65 | 3 (2025)

PRISPEVKI ZA NOVEJŠO ZGODOVINO

3

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="CVET-1887_7_2_43-46" xml:lang="sl">
 <teiHeader>
   <fileDesc>
     <titleStmt>
       <title>Vzroki in koristi terpljenja pobožnih duš. (1887) [CVET]</title>
       <author>Repič, Hijacint</author>
     </titleStmt>
     <editionStmt><edition>1.0</edition></editionStmt>
     <extent><measure unit="words">1203</measure></extent>
     <publicationStmt>
       <distributor>CLARIN.SI</distributor>
       <idno type="handle">http://hdl.handle.net/11356/1226</idno>
       <date when="2024-05-07"> 7. maj 2024</date>
     </publicationStmt>
     <sourceDesc>
       <bibl type="article">
         <author>Repič, Hijacint</author>
         <title level="a">Vzroki in koristi terpljenja pobožnih duš.</title>
         <bibl type="monogr">
           <title level="j">Cvetje z vertov sv. Frančiška</title>
           <biblScope unit="volume">7</biblScope>
           <biblScope unit="number">2</biblScope>
           <biblScope unit="page" from="11" to="14">11–14</biblScope>
           <date when="1887">1887</date>
           <idno type="URI" subtype="URN">https://www.dlib.si/.../PDF</idno>
         </bibl>
       </bibl>
     </sourceDesc>
```

INŠTITUT ZA NOVEJŠO ZGODOVINO

INŠTITUT ZA NOVEJŠO ZGODOVINO

PRISPEVKI ZA NOVEJŠO ZGODOVINO

Letnik LXV Ljubljana 2025 Številka 3

Prispevki za novejšo zgodovino
Contributions to the Contemporary History
Contributions a l'histoire contemporaine
Beiträge zur Zeitgeschichte

UDC/UDK 94(497.4) "18/19 "
ISSN 0353-0329 (tiskana izdaja)
2463-7807 (spletna izdaja); https://ojs.inz.si/pnz
DOI https://doi.org/10.51663/pnz.65.3

Izdajatelj/Published by: Inštitut za novejšo zgodovino/Institute of Contemporary History, Privoz 11, SI-1000 Ljubljana, tel. (386) 01 200 31 20, fax (386) 01 200 31 60, e-mail: jure.gasparic@inz.si

Sofinancer/Financially supported by: Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije/ Slovenian Research and Innovation Agency

Računalniški prelom/Typesetting: Studio Aleja d.o.o.

Tisk/Printed by: Medium d.o.o.

Cena/Price: 15,00 EUR

Zamenjave/Exchange: Inštitut za novejšo zgodovino/Institute of Contemporary History, Privoz 11, SI-1000 Ljubljana

Prispevki za novejšo zgodovino so indeksirani v/are indexed in: Scopus, ERIH Plus, Historical Abstract, ABC-CLIO, PubMed, CEEOL, Ulrich's Periodicals Directory, EBSCOhost

Številka vpisa v razvid medijev: 720

Fotografija na naslovnici:

Primer začetka zapisa posameznega besedila v formatu TEI.

Vir: Diana Košir in Tomaž Erjavec.

Vsebina

Razprave – Articles

Ocene in poročila – Reviews and Reports


SISTORY
ZGODOVINA SLOVENIJE

Uredniško obvestilo

Prispevki za novejšo zgodovino je ena osrednjih slovenskih znanstvenih zgodovinopisnih revij, ki objavlja teme s področja novejše zgodovine (19., 20. in 21. stoletje) srednje in jugovzhodne Evrope.

Od leta 1960 revijo redno izdaja Inštitut za novejšo zgodovino (do leta 1986 je izhajala pod imenom *Prispevki za zgodovino delavskega gibanja*).

Revija izide trikrat letno v slovenskem jeziku in v naslednjih tujih jezikih: angleščina, nemščina, srbščina, hrvaščina, bosanščina, italijanščina, slovaščina in češčina. Članki izhajajo z izvlečki v angleščini in slovenščini ter povzetki v angleščini.

Arhivski letniki so dostopni na **Zgodovina Slovenije - SIstory**.

Informacije za avtorje in navodila so dostopni na https://ojs.inz.si/pnz.

Editorial Notice

Contributions to Contemporary History is one of the central Slovenian scientific historiographic journals, dedicated to publishing articles from the field of contemporary history (the 19th, 20th and 21st century).

It has been published regularly since 1960 by the Institute of Contemporary History, and until 1986 it was entitled Contributions to the History of the Workers' Movement.

The journal is published three times per year in Slovenian and in the following foreign languages: English, German, Serbian, Croatian, Bosnian, Italian, Slovak and Czech. The articles are all published with abstracts in English and Slovenian as well as summaries in English.

The archive of past volumes is available at the **History of Slovenia - SIstory** web portal.

Further information and guidelines for the authors are available at https://ojs.inz.si/pnz.

Razprave – Articles

Jezikovne tehnologije in digitalna humanistika

Tematska številka, ki je pred vami, prinaša izbrane in nadgrajene prispevke z bienalne konference Jezikovne tehnologije in digitalna humanistika. Leta 2024 je konferenca potekala že štirinajstič, petič zapored pa je poleg jezikovnih tehnologij obravnavala tudi teme s področja digitalne humanistike. Dogodek je potekal ob podpori Slovenskega društva za jezikovne tehnologije (SDJT), Centra za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT UL) ter raziskovalnih infrastruktur CLARIN.SI in DARIAH-SI.

Konferenca je pritegnila številne predstavnice in predstavnike raziskovalne, razvojne, študentske in širše skupnosti, ki so v dveh dneh predstavljali in spoznavali najnovejše raziskave, izsledke in aktivnosti, pa tudi izzive in težave, s katerimi se področje trenutno sooča. Avtorje in avtorice najbolje ocenjenih prispevkov smo povabili k vsebinski nadgradnji in sodelovanju v pričujoči tematski številki. Nastalo je deset zanimivih in ažurnih razprav, ki odlično odražajo aktualne raziskovalne trende in novosti: v ospredju so predstavitve novih virov in orodij, zlasti za raziskave govorjenega jezika in zgodovinskih virov, popisane so novosti strojnega jezikoslovnega označevanja in procesiranja jezikovnih podatkov, kot povsem nova tema pa se izrisuje ocenjevanje uporabnosti velikih jezikovnih modelov za različne naloge uporabnega jezikoslovja.

Kaja Dobrovoljc predstavlja novo različico drevesnice govorjene slovenščine, razširjene z več kot 3000 na novo razčlenjenimi izjavami, in ob tem poudari razlike med govorjenim in pisnim jezikom ter njihov vpliv na delovanje razčlenjevalnikov. Ajda Pretnar Žagar se posveča analizi zgodovinskega korpusa sPeriodika in razkriva osrednjo vlogo slovenskih časopisov pri narodnem prebujanju, tematsko raznolikost objav ter omejitve, ki jih povzročajo napake OCR pri digitalizaciji besedil. Diana Košir in Tomaž Erjavec opisujeta korpus CVET, sestavljen iz besedil patra Hijacinta Repiča, in analizirata njegov pripovedni slog ter normativne posebnosti starejšega slovenskega jezika. Katja Meden, Ana Cvek, Vid Klopčič, Mihael Ojsteršek, Matevž Pesek, Mojca Šorn in Andrej Pančur analizirajo portal SIstory, ki po tehnični nadgradnji omogoča večjo preglednost, interoperabilnost in dostopnost zgodovinskih vsebin širši javnosti.

Luka Terčon, Kaja Dobrovoljc in Nikola Ljubešić opisujejo razvoj in zmogljivosti orodja CLASSLA-Stanza, ki uspešno nadgrajuje cevovod Stanza in omogoča kakovostno označevanje besedil v južnoslovanskih jezikih, vključno s spletnimi viri in transkripcijami govora. Jaka Čibej in Tina Munda predlagata polavtomatski pristop k popravljanju lematizacije in oblikoskladenjskih oznak, ki z uporabo oblikoslovnega leksikona Sloleks občutno zmanjša časovno potratnost ročnega pregledovanja označenih besedil in s tem pripravo jezikovnih učnih podatkov. Mojca Brglez, Veronika Bajt, Senja Pollak, Špela Rot in Matej Martinc predstavijo sistem za zaznavo pomenskih premikov v slovenščini in na primeru tematizacije migracij osvetlijo pomen konteksta in družbenih dejavnikov za jezikovno rabo v določenih obdobjih.

Špela Arhar Holdt, Magdalena Gapsa, Polona Gantar in Iztok Kosem preverjajo zmožnosti ChatGPT-ja pri prepoznavanju in razvrščanju sopomenk ter izdelavi slovarskih gesel v slovenščini, pri čemer pokažejo, da se model kljub določenim omejitvam izkaže kot obetaven pomočnik v digitalnem slovaropisju. Podobno temo naslavljata tudi Ivana Filipović Petrović in Slobodan Beliga, ki preučujeta, ali umetna inteligenca zmore prepoznati pomen in konceptualno povezanost hrvaških idiomov. Matej Klemen pa poroča o rezultatih dveh testov besedišča med govorci slovenščine kot drugega ali tujega jezika, ki se izkažeta kot učinkovita alternativa obstoječim metodam za razvrščanje govorcev glede na njihovo jezikovno raven.

Takšno pokonferenčno gostovanje v reviji Prispevki za novejšo zgodovino ni prvo – podobna tematska številka je bila uspešno pripravljena že leta 2019. Tako kot tedaj revija tudi tokrat s svojo široko, empirično utemeljeno in interdisciplinarno naravnanostjo ponuja izvrsten okvir za predstavitev novosti na področju jezikovnih virov in digitalne humanistike. Posebna vrednost sodelovanja je v prepletu področij, pristopov in raziskovalnih generacij, kar prispeva k utrjevanju odprte, strokovno utemeljene in dolgoročno vzdržne skupnosti. Ob tem velja poudariti vlogo nacionalnih in mednarodnih infrastrukturnih pobud, ki omogočajo razvoj orodij, virov in povezovalnih okolij, nujnih za konkurenčnost nacionalnih prostorov in jezikov v širšem jezikovnotehnološkem kontekstu.

Uredniki se iskreno zahvaljujemo avtoricam in avtorjem, ki so svoje prispevke nadgradili in prilagodili za objavo, recenzentkam in recenzentom za njihovo strokovno in konstruktivno delo ter vsem drugim sodelujočim. Želimo si, da tematska številka, ki je pred vami, ne bi bila le pregled stanja, temveč tudi pregleden nabor dobrih praks in izraz skupne ambicije po nadaljnjem povezovanju, krepitvi interdisciplinarnih pristopov ter ustvarjanju odprtega raziskovalnega prostora, v katerem bo tudi v prihodnje mogoče odgovarjati na kompleksna vprašanja sodobnega jezikovnega in kulturnega okolja.

Ljubljana, 12. avgust 2025

Špela Arhar Holdt, Tomaž Erjavec, Mojca Šorn

Language Technologies and Digital Humanities

This thematic issue gathers selected and expanded contributions from the Language Technologies and Digital Humanities biennial conference. In 2024, the conference was organised for the 14th time. For the fifth time it focused on topics related to digital humanities alongside those of language technologies. The event was supported by the Slovenian Language Technologies Society (SDJT), the Centre for Language Resources and Technologies of the University of Ljubljana (CJVT UL), and the research infrastructures CLARIN.SI and DARIAH-SI.

The conference attracted numerous representatives from the research, development, student, and broader community who spent two days sharing and exploring the latest research, discoveries, activities, and the current challenges in the field. We invited the authors of the top-ranked papers to contribute to the present thematic issue. They have created ten engaging and topical discussions that reflect current research trends and innovations. The contributions focus on new resources and tools, especially for spoken language research; historical sources; innovations in the machine tagging and processing of linguistic data; and introduce a completely new topic of assessing the applicability of large-scale language models to various applied linguistic tasks.

Kaja Dobrovoljc presents a new version of the spoken Slovenian treebank, enriched with over 3,000 newly segmented utterances, emphasising the differences between spoken and written language and their influence on parser performance. Ajda Pretnar Žagar analyses the historical corpus sPeriodika and highlights the central role of Slovenian newspapers in the national awakening, the thematic diversity of publications, and the limitations caused by OCR errors during text digitisation. Diana Košir and Tomaž Erjavec describe the CVET corpus, which comprises texts by Fr. Hijacint Repič, and analyse his narrative style and the normative features of the older Slovenian language. Katja Meden, Ana Cvek, Vid Klopčič, Mihael Ojsteršek, Matevž Pesek, Mojca Šorn, and Andrej Pančur analyse the SIstory portal, which, after a technical upgrade, enables greater transparency, interoperability, and accessibility of historical content to the general public.

Luka Terčon, Kaja Dobrovoljc, and Nikola Ljubešić outline the development and features of the CLASSLA-Stanza tool, which builds on the Stanza pipeline to enable accurate annotation of texts in South Slavic languages, including online resources and transcriptions of speech. Jaka Čibej and Tina Munda propose a semi-automatic method for correcting lemmatisation and morphosyntactic tags by utilising the Sloleks morphological lexicon, which significantly reduces the time-consuming manual review of annotated texts and, consequently, the creation of language learning datasets. Mojca Brglez, Veronika Bajt, Senja Pollak, Špela Rot, and Matej Martinc present a system for detecting semantic shifts in the Slovenian language, using the topic of migration as a case study to emphasise the significance of context and social influences in language use during specific periods.

Špela Arhar Holdt, Magdalena Gapsa, Polona Gantar, and Iztok Kosem evaluate ChatGPT's abilities in recognising synonyms, classifying them, and generating dictionary entries in Slovenian. They demonstrate that, despite certain limitations, it is a promising tool for digital lexicography. A similar topic is also addressed by Ivana Filipović Petrović and Slobodan Beliga, who investigate whether artificial intelligence can recognise the meaning and conceptual relations between Croatian idioms. Matej Klemen reports on the results of two vocabulary tests among speakers of Slovenian as a second or foreign language, which prove to be an effective alternative to traditional methods of classifying speakers based on their language proficiency.

This is not the first time the journal Prispevki za novejšo zgodovino has published a post-conference issue of this kind; a similar thematic issue was successfully published in 2019. The journal's broad, empirically based, and interdisciplinary orientation provides an excellent framework for presenting new developments in language resources and digital humanities. The true benefit of cooperation lies in combining disciplines, methods, and research generations, which encourages an open, professionally based, and sustainable community in the long term. In this context, it is essential to emphasise the role of national and international infrastructure initiatives, which support the development of tools, resources, and networking environments crucial for the competitiveness of national languages and spaces within the broader linguistic-technological framework.

The editors wish to thank the authors who have refined and adapted their contributions for publication, the reviewers for their professional and constructive efforts, and all other contributors. We hope the thematic issue before you will serve not only as an overview of the current situation, but also highlight good practices and reflect our shared ambition to foster stronger connections, enhance interdisciplinary approaches, and create an open research space, where the complex issues of the modern linguistic and cultural environment can continue to be explored.

Ljubljana, 12 August 2025

Špela Arhar Holdt, Tomaž Erjavec, Mojca Šorn

1.01 DOI: https://doi.org/10.51663/pnz.65.3.01

Kaja Dobrovoljc*

Treebanking Spoken Slovenian: New Data, Models, and Lessons Learned

IZVLEČEK

DREVESNICA GOVORJENE SLOVENŠČINE: NOVI PODATKI, MODELI IN KLJUČNI NAUKI

Prispevek predstavlja novo različico drevesnice govorjene slovenščine (SST), uravnotežene in reprezentativne zbirke transkribiranega spontanega govora z ročno označenimi lemami, besednimi vrstami, oblikoslovnimi značilnostmi in skladenjskimi odvisnostmi, ki je bila nedavno razširjena z več kot 3.000 na novo razčlenjenimi izjavami. Po kratkem pregledu postopkov vzorčenja, označevanja in poenotenja korpusnih podatkov – ki smo jih podrobneje predstavili že v predhodni razpravi – ponazorimo pomen tega jezikovnega vira za raziskave na področju jezikoslovja in strojne obdelave jezika. S primerjavo govorne in pisne drevesnice najprej izpostavimo leksikalne ter oblikoslovno-skladenjske posebnosti govora v primerjavi s pisnim jezikom, nato pa predstavimo njihov vpliv na delovanje orodij za samodejno slovnično razčlenjevanje govornih transkripcij. Na koncu predstavimo metodološke izkušnje, pridobljene pri razvoju drevesnice, razpravljamo o njenem potencialu za nadaljnje raziskave govorjenega jezika in poudarimo pomen tovrstnih virov z vidika naslavljanja jezikovne raznolikosti pri razvoju jezikovnih tehnologij.

Ključne besede: označevanje korpusov, odvisnostna drevesnica, spontani govor, Universal Dependencies, razčlenjevanje

* **PhD, Res. Assoc., University of Ljubljana, Faculty of Arts, Aškerčeva 2, SI-1000 Ljubljana; Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, kaja.dobrovoljc@ff.uni-lj.si; ORCID: 0000-0002-5909-7965**

ABSTRACT

This paper presents a new version of the Spoken Slovenian Treebank (SST), a balanced and representative collection of transcribed spontaneous speech with manually annotated lemmas, part-of-speech tags, morphological features, and syntactic dependencies, recently expanded with over 3,000 newly annotated utterances. After a brief overview of the data sampling, annotation, and consolidation processes—presented in detail in previous work—we evaluate the significance of this new language resource for both linguistic research and natural language processing by first highlighting its distinctive lexical and morphosyntactic features in comparison to writing, and then assessing their impact on the performance of tools for automatic grammatical annotation. Finally, we reflect on the methodological insights gained during treebank creation, discuss the potential of SST for advancing spoken language research, and argue for the necessity of such resources in supporting linguistic diversity in language technology.

Keywords: corpus annotation, dependency treebank, spontaneous speech, Universal Dependencies, parsing

Introduction

Spoken language treebanks, i.e. syntactically annotated collections of transcribed speech, represent one of the fundamental language resources for data-driven spoken language research in both linguistics[1] and natural language processing.[2] Consequently,

1 Erhard Hinrichs and Sandra Kübler, "Treebank Profiling of Spoken and Written German," in Montserrat Civit Torruella, Sandra Kübler, and María Antonia Martí Antonín, eds., *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005): 9–10 December 2005, Barcelona,* 65–76 (Barcelona: Universitat de Barcelona, 2005). Paola Pietrandrea and Aline Delsart, "Chapter 16. Macrosyntax at Work: Functions and Distribution of Macrosyntactic Patterns in the Rhapsodie Corpus," in Anne Lacheret-Dujour, Sylvain Kahane, and Paola Pietrandrea, eds., *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French* (Amsterdam: John Benjamins Publishing Company, 2019), 285–314, https://doi.org/10.1075/scl.89.17pie. Ineke Schuurman, Marijke Schouppe, and Henk Hoekstra, "Harvesting Dutch Trees: Syntactic Properties of Spoken Dutch," in Tanya Gaustad, ed., *Computational Linguistics in the Netherlands 2002: Selected Papers from the Thirteenth CLIN Meeting* (Amsterdam: Rodopi, 2003), 129–41.

2 Zoey Liu and Emily Prud'hommeaux, "Dependency Parsing Evaluation for Low-Resource Spontaneous Speech," in Eyal Ben-David, Shay Cohen, Ryan McDonald, Barbara Plank, Roi Reichart, Guy Rotman, and Yftah Ziser, eds., *Proceedings of the Second Workshop on Domain Adaptation for NLP* (Kyiv, Ukraine: Association for Computational Linguistics, April 2021), 156–65, https://aclanthology.org/2021.adaptnlp-1.16/. Anouck Braggaar and Rob van der Goot, "Challenges in Annotating and Parsing Spoken, Code-Switched, Frisian-Dutch Data," in Eyal Ben-David, Shay Cohen, Ryan McDonald, Barbara Plank, Roi Reichart, Guy Rotman, and Yftah Ziser, eds., *Proceedings of the Second Workshop on Domain Adaptation for NLP* (Kyiv, Ukraine: Association for Computational Linguistics, April 2021), 50–58, https://aclanthology.org/2021.adaptnlp-1.6/. Caines, Andrew, Michael McCarthy, and Paula Buttery, "Parsing Transcripts of Speech," in Nicholas Ruiz and Srinivas Bangalore, eds., *Proceedings of the Workshop on Speech-Centric Natural Language Processing (SCNLP@EMNLP 2017)* (Copenhagen, Denmark: Association for Computational Linguistics, September 7, 2017), 27–36, https://doi.org/10.18653/v1/w17-4604. Kaja Dobrovoljc and Matej Martinc, "Er ... Well, It Matters, Right? On the Role of Data Representations in Spoken Language Dependency Parsing," in Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster, eds., *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (Brussels, Belgium: Association for Computational Linguistics, November 2018), 37–46, https://doi.org/10.18653/v1/W18-6005.

many spoken language treebanks have been developed over the recent decades, such as the Switchboard corpus for English,[3] CGN for Dutch,[4] PDTSL for Czech,[5] NDC and LIA for Norwegian,[6] Rhapsodie for French,[7] as well as the multilingual Verbmobil[8] and CHILDES[9] collections. Recently, many such treebanks have emerged as part of the expanding multilingual Universal Dependencies (UD) dataset.[10]

For Slovenian, the Spoken Slovenian Treebank (SST)[11] has been the only language resource of this kind to date. To support computational and corpus linguistic research alike, the SST treebank was designed as a representative sample of the GOS reference

3 J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92),* 517–20, vol. 1 (San Francisco, CA, USA: IEEE, 1992), https://doi.org/10.1109/ICASSP.1992.225858. John J. Godfrey and Edward Holliman, *Switchboard-1 Release 2 LDC97S62,* Web Download (Philadelphia: Linguistic Data Consortium, 1993), https://doi.org/10.35111/sw3h-rw02.

4 Ton van der Wouden, Heleen Hoekstra, Michael Moortgat, Bram Renmans, and Ineke Schuurman, "Syntactic Analysis in the Spoken Dutch Corpus (CGN)," in Manuel González Rodríguez and Carmen Paz Suarez Araujo, eds., *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)* (Las Palmas, Canary Islands, Spain: European Language Resources Association (ELRA), May 2002), https://aclanthology.org/L02-1071/. Dutch Language Institute. *Corpus Gesproken Nederlands – CGN (Version 2.0.3)*, 2014, data set, http://hdl.handle.net/10032/tm-a2-k6.

5 Jan Hajič, Silvie Cinková, Marie Mikulová, Petr Pajas, Jan Ptáček, Josef Toman, and Zdeňka Urešová, "PDTSL: An Annotated Resource for Speech Reconstruction," in *2008 IEEE Spoken Language Technology Workshop* (IEEE, 2008), 93–96, https://doi.org/10.1109/SLT.2008.4777848. Jan Hajič, Petr Pajas, David Mareček, Marie Mikulová, Zdeňka Urešová, and Petr Podveský, *Prague Dependency Treebank of Spoken Language (PDTSL) 0.5* (LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2009), http://hdl.handle.net/11858/00-097C-0000-0001-4914-D.

6 Lilja Øvrelid, Anne Kåsen, Kristin Hagen, Anders Nøklestad, and Janne Bondi Johannessen, "The LIA Treebank of Spoken Norwegian Dialects," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 4482–88, 2018, https://www.nb.no/sprakbanken/en/resource-catalogue/oai-tekstlab-uio-no-lia-trebanken/. Andre Kåsen, Kristin Hagen, Anders Nøklestad, Joel Priestly, Per Erik Solberg, and Dag Trygve Truslew Haug, "The Norwegian Dialect Corpus Treebank," in Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille, France: European Language Resources Association, June 2022), 4827–32, https://aclanthology.org/2022.lrec-1.516/, https://www.nb.no/sprakbanken/en/resource-catalogue/oai-tektlab-uio-no-ndc-trebanken/.

7 Anne Lacheret-Dujour, Sylvain Kahane, and Paola Pietrandrea, eds., *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French.* Studies in Corpus Linguistics 89 (Amsterdam: John Benjamins Publishing Company, 2019), https://doi.org/10.1075/scl.89, https://github.com/UniversalDependencies/UD_French-Rhapsodie.

8 Erhard W. Hinrichs, Julia Bartels, Yasuhiro Kawata, Valia Kordoni, and Heike Telljohann, "The Tübingen Treebanks for Spoken German, English, and Japanese," in Wolfgang Wahlster, ed., *Verbmobil: Foundations of Speech-to-Speech Translation* (Berlin, Heidelberg: Springer Berlin Heidelberg, 2000), 550–74, https://doi.org/10.1007/978-3-662-04230-4_40. Bavarian Archive for Speech Signals (BAS). *VM2 – Speech Corpus,* 2016, http://hdl.handle.net/11022/1009-0000-0000-FC55-5.

9 Lisa Pearl and Jon Sprouse, "Syntactic Islands and Learning Biases: Combining Experimental Syntax and Computational Modeling to Investigate the Language Acquisition Problem," *Language Acquisition* 20, No. 1 (2013): 23–68, https://doi.org/10.1080/10489223.2012.738742.

10 Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman, "Universal Dependencies," *Computational Linguistics* 47, No. 2 (2021): 255–308, https://doi.org/10.1162/coli_a_00402. Kaja Dobrovoljc, "Spoken Language Treebanks in Universal Dependencies: An Overview," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille, France: ELRA, 2022), 1798–806, https://aclanthology.org/2022.lrec-1.191/.

11 Kaja Dobrovoljc and Joakim Nivre, "The Universal Dependencies Treebank of Spoken Slovenian," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (Portorož: ELRA, 2016), 1566–73, https://aclanthology.org/L16-1248/.

corpus of spoken Slovenian[12] and features manually annotated transcriptions on the levels of lemmatization, MULTEXT-East[13] morphological tags, and morphosyntactic annotations following the aforementioned UD annotation scheme, which includes cross-lingually comparable annotations of part-of-speech categories, morphological features and syntactic dependencies (Figure 1). As such, the treebank complements the SSJ reference treebank of written Slovenian, which features identical annotations,[14] and has already been used as the main data source for the development of specialized computational models for grammatical annotation of spoken Slovenian.[15]

To address the limitations of the original SST treebank—namely its relatively small size (approximately 3,100 parsed utterances, totalling 30,000 annotated tokens) and its diverse but fragmented data (short samples from numerous speech events)—the treebank has recently been expanded to more than three times its original size. This major extension, carried out as part of the ongoing SPOT project,[16] was first presented at the Language Technologies and Digital Humanities conference in September 2024.[17] In this paper, we build on that work—selected to appear in this special issue—by summarizing the entire process, describing the very latest release of the SST treebank, published as part of UD release v2.15,[18] and evaluating newly developed parsing models for state-of-the-art automatic grammatical annotation of spoken Slovenian. Finally, we conclude by summarizing key lessons learned from the dataset creation, annotation, and exploitation.

12 Anja Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej, and Tomaž Erjavec, "Spoken Corpus Gos 1.1," http://hdl.handle.net/11356/1438 (Slovenian language resource repository CLARIN.SI, 2021). Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek, and Marko Stabej, "Compilation, Transcription and Usage of a Reference Speech Corpus: The Case of the Slovene Corpus GOS," *Language Resources and Evaluation* 47, No. 4 (2013): 1031–48, https://doi.org/10.1007/s10579-013-9216-5.

13 Tomaž Erjavec, "MULTEXT-East," in Nancy Ide and James Pustejovsky, eds., *Handbook of Linguistic Annotation*, 441–62 (Dordrecht: Springer, 2017), https://doi.org/10.1007/978-94-024-0881-2_17.

14 Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek, "The Universal Dependencies Treebank for Slovenian," in *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, (Association for Computational Linguistics, 2017), 33–38, https://doi.org/10.18653/v1/W17-1406. Kaja Dobrovoljc, Luka Terčon, and Nikola Ljubešić, "Universal Dependencies za slovenščino: Nove smernice, ročno označeni podatki in razčlenjevalni model," *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research* 11, No. 1 (2023): 218–46, https://doi.org/10.4312/slo2.0.2023.1.218-246.

15 Dobrovoljc and Martinc, "Er ... Well, It Matters, Right?," 37–46, https://doi.org/10.18653/v1/W18-6005. Darinka Verdonik, Kaja Dobrovoljc, Tomaž Erjavec, and Nikola Ljubešić, "Gos 2: A New Reference Corpus of Spoken Slovenian," in Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (Torino, Italy: ELRA and ICCL, May 2024), 7825–30, https://aclanthology.org/2024.lrec-main.691/.

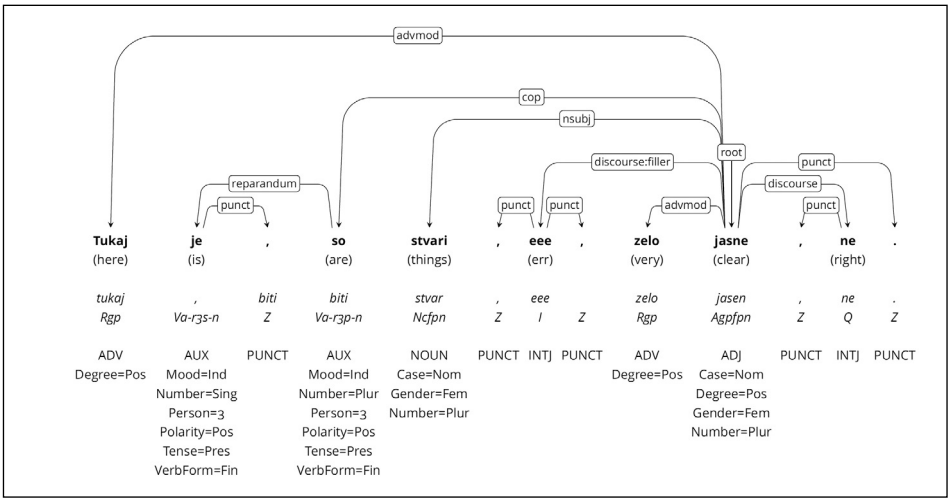
16 *Treebank-driven approach to the study of Spoken Slovenian*, ARIS grant No. Z6-4617, https://spot.ff.uni-lj.si/

17 Kaja Dobrovoljc, "Extending the Spoken Slovenian Treebank," in *Proceedings of the Conference on Language Technologies and Digital Humanities* (Ljubljana, Slovenia, 2024), 116–46, https://doi.org/10.5281/zenodo.13936393.

18 Zeman et al., "Universal Dependencies 2.15," http://hdl.handle.net/11234/1-5787 (LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2024).

We describe this major improvement of the SST treebank by summarizing the data sampling, annotation, and final dataset consolidation in Section SST Treebank Extension. Section New SST Treebank Overwiev provides an integrated overview of the resulting language resource, including its format and availability. To exemplify its value for further empirical investigations of lexical and grammatical characteristics of Slovenian speech, we compare the new SST treebank to the SSJ treebank of written Slovenian in Section Comparison with the SSJ Treebank of Written Slovenian and present a comparative analysis of the newly available Trankit and CLASSLA-Stanza NLP models for processing spoken Slovenian in Section 5. Finally, Section Discussion concludes with a discussion of key lessons learned and broader implications of this work.

Figure 1: Example of a grammatically annotated utterance in the SST treebank (roughly translated as *Things here are very clear, right.*) featuring UD syntactic annotations (top), part-of-speech tags and morphological features (bottom), as well as MULTEXT-East lemmas and morphosyntactic tags (italics)



Source: Own work

SST Treebank Extension

The extension of the Spoken Slovenian Treebank (SST) has been extensively documented in the aforementioned previous work,[19] which is why we only summarize the key steps below and refer readers to that paper for detailed descriptions.

19 Kaja Dobrovoljc, "Extending the Spoken Slovenian Treebank."

Data sampling

To address the limitations of the original SST corpus—namely its relatively small size and fragmented data—the treebank was extended by a minimum of 50,000 new tokens, while maintaining representativeness with respect to the updated GOS 2.1 reference corpus of spoken Slovenian.[20] The sampling procedure, designed in collaboration with the Mezzanine project,[21] involved two main steps. First, 22 samples from GOS 1 events in the original SST were expanded by approximately 450 additional words each, yielding about 10,000 new words. Second, 57 new speech events from the ARTUR subset were added, each contributing around 800 words, totaling approximately 40,000 new words. To ensure coherent syntactic structures, the ARTUR data—originally segmented at pause boundaries—was automatically re-segmented based on sentence-final punctuation, producing more syntactically and semantically meaningful units.[22] The exact counts, which also account for the post-festum modifications of the data described in the following sections, are reported in Section New SST Treebank Overwiew (Table 1).

Data annotation

The annotation process began with a semi-automated morphological annotation, documented by Čibej and Munda (2024),[23] which was then used for automatic parsing using the Trankit dependency parser. Each of the 79 document-level files was subsequently assigned to 2–3 independent annotators, who manually verified and corrected the annotations using the Q-CAT tool,[24] enhanced with audio support via embedded URLs.

20 Darinka Verdonik, Kaja Dobrovoljc, Tomaž Erjavec, and Nikola Ljubešić, "Gos 2: A New Reference Corpus of Spoken Slovenian," in *Proceedings of the 2024 Joint International Conference.* Darinka Verdonik, Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej, Tomaž Erjavec, Tomaž Potočnik, Mirjam Sepesy Maučec, Simona Majhenič, Andrej Žgank, Andreja Bizjak, Lucija Gril, Simon Dobrišek, Janez Križaj, Marko Bajec, Iztok Lebar Bajec, Tjaša Jelovšek, Mitja Trojar, Mitja Bernjak, Naum Dretnik, Gregor Strle, Kaja Dobrovoljc, Nikola Ljubešić, and Peter Rupnik, *Spoken Corpus Gos 2.1 (Transcriptions)* (Slovenian language resource repository CLARIN.SI, 2023), http://hdl.handle.net/11356/1863.

21 *Mezzanine; temeljne raziskave za razvoj govornih vIrov in tehnologij za slovenščino*, https://mezzanine.um.si/. Darinka Verdonik, Nikola Ljubešić, Peter Rupnik, Kaja Dobrovoljc, and Jaka Čibej, "Izbor in urejanje gradiv za učni korpus govorjene slovenščine ROG," paper presented at the *14th Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, Ljubljana, Slovenia, September 19–20, 2024 (Institute of Contemporary History, 2024), https://doi.org/10.5281/zenodo.13936425.

22 The re-segmentation was performed fully automatically, except for a few outliers where the absence of sentence-final punctuation in the original ARTUR transcriptions led to exceptionally long utterances. These cases were manually segmented for the UD release v2.16.

23 Jaka Čibej and Tina Munda, "Metoda polavtomatskega popravljanja lem in oblikoskladenjskih oznak na primeru učnega korpusa govorjene slovenščine ROG," paper presented at the *14th Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, Ljubljana, Slovenia, September 19–20, 2024 (Institute of Contemporary History, 2024), https://doi.org/10.5281/zenodo.13936390.

24 Janez Brank, Q-CAT Corpus Annotation Tool 1.5. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1844.

The curation process was completed in WebAnno[25] to reconcile multiple annotations and ensure consistency with updated guidelines.

In the process, we developed a new and improved version of the UD guidelines for Slovenian, which now account for numerous speech-specific phenomena such as self-repairs and discourse markers. These updates build upon our previous annotation experience[26] and incorporate recent practices and discussions within the community.[27] The guidelines are available as a standalone document in Slovenian[28] and as an abbreviated version of the Slovenian UD guidelines online in English.[29]

Data consolidation

Finally, both manually revised datasets—the original SST and the new GOS 2 data—were merged and consolidated. This process involved harmonizing metadata formatting, punctuation, and letter-case principles across the subsets. Sentence-medial and sentence-final punctuation was semi-automatically added to GOS 1 transcriptions using the Slovene Punctuator tool,[30] followed by manual corrections to align with the conventions of the ARTUR dataset.[31] Non-lexical tokens, such as [audience:laughter] and [pause], were removed to ensure consistency with ARTUR and UD treebank trends in general.[32] These and other non-lexical phenomena can still be accessed from the transcriptions of the reference GOS 2.1 corpus if necessary.

The final data consolidation also included correcting transcription errors, such as erroneous capitalization caused by automatic letter case unification in GOS 2.1, and resolving tokenization issues flagged during UD validation. Morphological annotation inconsistencies were also corrected, including lemmatization errors and refinements to specific categories like colloquial expressions and anonymized names.

25 Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann, "WebAnno: A Flexible, Web-Based and Visually Supported System for Distributed Annotations," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 1–6, 2013, https://aclanthology.org/P13-4001. *WebAnno - Log in*, https://www.clarin.si/webanno/login.html.

26 Kaja Dobrovoljc and Joakim Nivre, "The Universal Dependencies Treebank of Spoken Slovenian," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (Portorož: ELRA, 2016), 1566–73, https://aclanthology.org/L16-1248/.

27 Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes, "Annotation Guidelines of UD and SUD Treebanks for Spoken Corpora: A Proposal," in Daniel Dakota, Kilian Evang, and Sandra Kübler, eds., *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, 35–47 (Sofia, Bulgaria: Association for Computational Linguistics, December 2021), https://aclanthology.org/2021.tlt-1.4/. Dobrovoljc, "Spoken Language Treebanks in Universal Dependencies," 1798-806.

28 Kaja Dobrovoljc and Luka Terčon, *Universal Dependencies: Smernice za označevanje besedil v slovenščini. Različica 1.7*. (Center za jezikovne vire in tehnologije Univerze v Ljubljani, 2024), https://wiki.cjvt.si/attachments/71.

29 Example of the online Slovenian UD guidelines for speech repairs: https://universaldependencies.org/sl/dep/reparandum.html.

30 *GitHub - clarinsi/Slovene_punctator,* https://github.com/clarinsi/Slovene_punctuator.

31 Darinka, Verdonik and Andreja Bizjak, *Pogovorni zapis in označevanje govora v govorni bazi Artur projekta RSDO.* Elaborat, predštudija, študija (Maribor: Univerza, 2023), https://dk.um.si/IzpisGradiva.php?lang=slv&id=85198.

32 Dobrovoljc, "Spoken Language Treebanks in Universal Dependencies," 1798–806.

New SST Treebank Overview

This section presents the contents of the new SST treebank with respect to its size, diversity of spoken data included, and availability.

Treebank size

As shown in Table 1, the resulting new, extended and revised, SST treebank based on approximately 10 hours of transcribed speech includes 344 unique speech events (documents) with a total of 6,108 utterances and 98,393 tokens. In comparison to the previous edition of the treebank (prior to the revisions presented in this paper),[33] the new SST treebank includes more than triple the number of transcribed tokens (+334%) and almost double the number of utterances (+196%), as well as a more varied set of events (+ 11%) and speakers (+ 11%). The average length of a (sampled) document has been extended from an average of 103 tokens per document to 286 tokens per document. As such, the SST treebank is one of the largest spoken language treebanks annotated in Universal Dependencies, surpassed only by the Naija UD treebank.

Table 1: Overview of the new SST treebank and its subsets

Subset	Events	Speakers	Utterances	Tokens
SST-2016 (revised)	287	594	2.903	36.960
New from GOS 1	22	61	1.236	13.112
New from ARTUR	57	72	1.969	48.321
SST-2024 (UD 2.15)	**344**	**676**	**6.108**	**98.393**

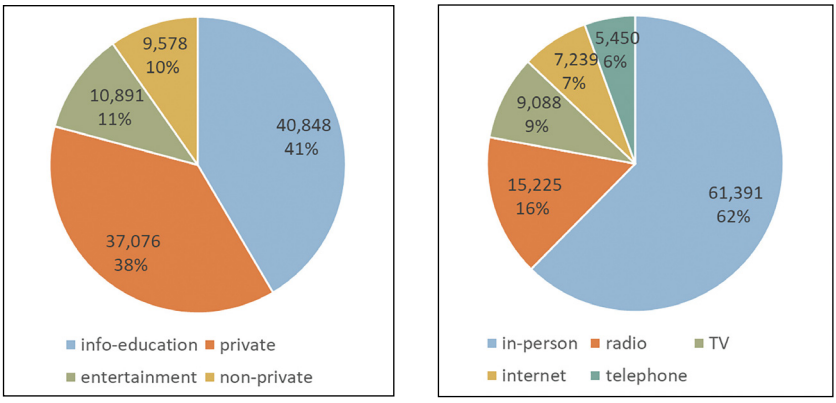
Source: Own work

33 The original version of the SST treebank (Dobrovoljc and Nivre, 2016) featured 287 events, 594 speakers, 3,188 utterances and 29,488 tokens.

Data diversity

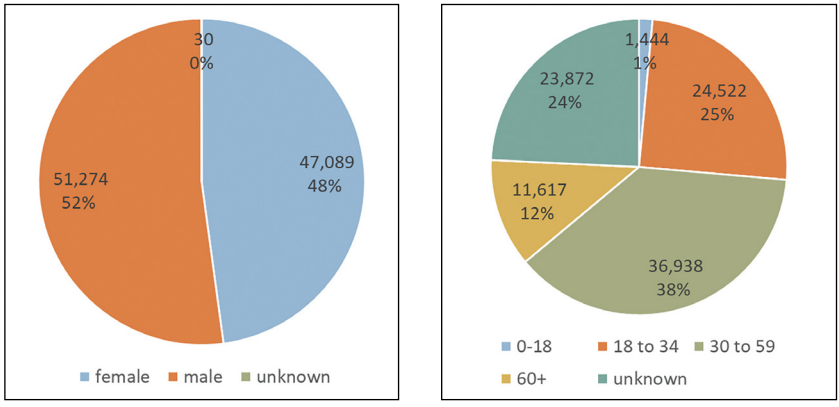
At the same time, the new SST treebank remains representative with respect to the reference GOS 2.1 and, indirectly, to Slovenian speech in general, as shown in Figures 2 to 5, which report the number of tokens per different types of speech events,[34] communication channels and speaker demographics.

Figure 2: Number of tokens in SST with respect to the event type (left) and channel (right)



Source: Own work

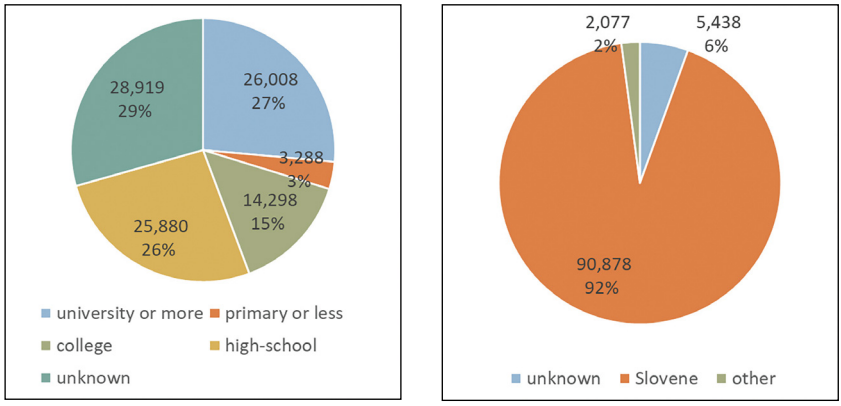
Figure 3: Number of tokens in SST with respect to speaker gender (left) and age (right)



Source: Own work

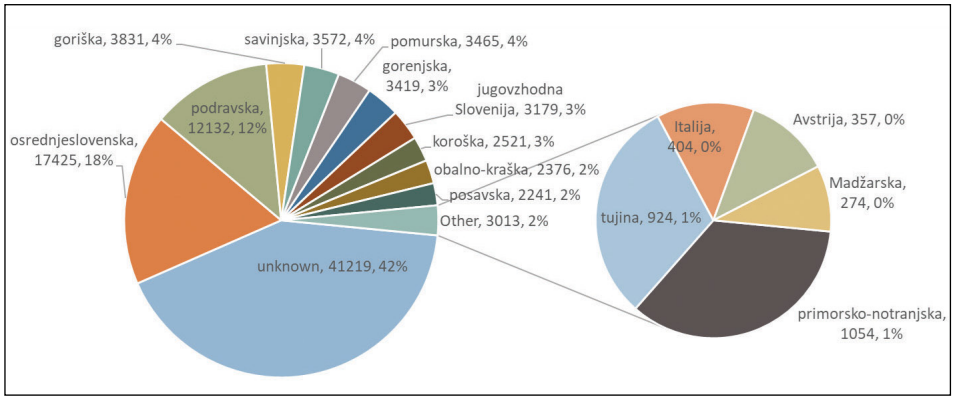
34 Generally, all events feature spontaneous speech, i.e. unscripted verbal communication that occurs naturally in real-time, albeit with varying amounts of planning in public and non-public situations. A more detailed characterisation of speech events can be retrieved from the metadata available in the reference GOS 2 corpus.

Figure 4: Number of tokens in SST with respect to speaker education (left) and first language (right)



Source: Own work

Figure 5: Number of tokens in SST with respect to the region of speaker residence



Source: Own work

Treebank availability

The latest version of the SST treebank was released as part of UD v2.15,[35] and is freely available under the CC-BY license, replacing the previous CC-BY-NC license that prohibited commercial use. The dataset follows the standard UD data split protocol, dividing the data into training, development, and test sets with approximate token distributions of 80%, 10%, and 10%, respectively. This revision aligns the SST data split with the ROG corpus (see below), ensuring an even distribution of

35 Zeman et al., "Universal Dependencies 2.15."

ROG-ARTUR data across subsets, while maintaining the original principles of randomized, document-level segmentation for representativeness with respect to different event and speaker types (Section Data diversity).

The treebank is encoded in the standard CONLL-U format,[36] illustrated in Figure 6, with detailed token-level annotations and metadata in comment lines (e.g., speaker ID, document ID, audio URLs, pronunciation-based spelling).[37] This ensures that all additional metadata—such as non-lexical tokens and detailed event and speaker data—can be traced and retrieved from the reference GOS corpus using persistent IDs.[38]

In addition to the official CONLL-U release and its availability on GitHub,[39] the SST treebank is also accessible via online tools for UD querying and visualization, including Grew-match,[40] INESS,[41] and the locally developed Drevesnik service,[42] based on the open-source dep_search tool.[43] These services support comprehensive exploration of the dataset, with some offering advanced query functions or even enabling audio playback.

Finally, the new SST treebank also serves as the backbone of the recently released ROG training corpus of spoken Slovenian,[44] which includes additional annotation layers for disfluencies, dialogue acts, and prosody boundaries in the ROG-ARTUR subset and is available in formats that support visualization and browsing in the EXMARaLDA tool.[45]

36 *CoNLL-U Format,* https://universaldependencies.org/format.html.

37 Due to space limitations, the CONLL-U example in Figure 6 only shows the first feature in the FEATS column (but see the example in Figure 1) and omits the contents of the MISC column altogether (e.g., pronunciation=tuki|GOS2.1_token_id=GOS119.tok1104).

38 This includes the retrieval audio recordings of the events, which are freely available under CC-BY for the ARTUR subset (Verdonik et al., 2023), and for research purposes for the GOS 1 subset (Verdonik et al., 2024).

39 *GitHub - UniversalDependencies/UD_Slovenian-SST*, https://github.com/UniversalDependencies/UD_Slovenian-SST/.

40 *Grew-match*, https://universal.grew.fr/. Bruno Guillaume, "Graph Matching and Graph Rewriting: GREW Tools for Corpus Exploration, Maintenance and Conversion," in Dimitra Gkatzia and Djamé Seddah, eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (Online: Association for Computational Linguistics, April 2021), 168–75, https://doi.org/10.18653/v1/2021.eacl-demos.21.

41 *INESS : Home*, https://clarino.uib.no/iness. Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik, "An Open Infrastructure for Advanced Treebanking," in Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco, eds., *META-RESEARCH Workshop on Advanced Treebanking at LREC2012* (Istanbul, Turkey, May 2012), 22–29.

42 *Drevesnik*, https://orodja.cjvt.si/drevesnik/. Miha Štravs, Kaja Dobrovoljc, and Luka Bezgovšek, *Service for Querying Dependency Treebanks Drevesnik 1.2* (Slovenian language resource repository CLARIN.SI, 2025), http://hdl.handle.net/11356/2034.

43 Juhani Luotolahti, Jenna Kanerva, and Filip Ginter, "Dep_search: Efficient Search Tool for Large Dependency Parsebanks," in Jörg Tiedemann and Nina Tahmasebi, eds., *Proceedings of the 21st Nordic Conference on Computational Linguistics* (Gothenburg, Sweden: Association for Computational Linguistics, May 2017), 255–58, https://aclanthology.org/W17-0233/.

44 Darinka Verdonik, Kaja Dobrovoljc, Peter Rupnik, Nikola Ljubešić, Simona Majhenič, Jaka Čibej, and Thomas Schmidt, *Training Corpus of Spoken Slovenian ROG 1.0* (Slovenian language resource repository CLARIN.SI, 2024), http://hdl.handle.net/11356/1992.

45 *EXMARaLDA*, https://www.exmaralda.org/.

Figure 6: Example of an annotated utterance (shown in Figure 1) in the CONLL-U format

```
# newdoc_id = GOS119
# sent_id = GOS119.s72
# speaker_id = Bm-gost-07155
# sound_url = https://nl.ijs.si/project/gos20/GOS119/GOS119.s72.mp3
# text = tukaj je , so stvari , eee , zelo jasne , ne .
1    tukaj   tukaj  ADV    Rgp       Degree=Pos     10    advmod           _    _
2    je      biti   VERB   Va-r3s-n  Mood=Ind...    4     reparandum       _    _
3    ,       ,      PUNCT  Z         _              2     punct            _    _
4    so      biti   AUX    Va-r3p-n  Mood=Ind...    10    cop              _    _
5    stvari  stvar  NOUN   Ncfpn     Case=Nom...    10    nsubj            _    _
6    ,       ,      PUNCT  Z         _              7     punct            _    _
7    eee     eee    INTJ   I         _              10    discourse:filler _    _
8    ,       ,      PUNCT  Z         _              7     punct            _    _
9    zelo    zelo   ADV    Rgp       Degree=Pos     10    advmod           _    _
10   jasne   jasen  ADJ    Agpfpn    Case=Nom...    0     root             _    _
11   ,       ,      PUNCT  Z         _              12    punct            _    _
12   ne      ne     PART   Q         Polarity=Neg   10    discourse        _    _
13   .       .      PUNCT  Z         _              10    punct            _    _
```

Source: Own work

Comparison with the SSJ Treebank of Written Slovenian

To illustrate the relevance of this newly created resource for further research on spoken Slovenian, we compare the new SST treebank with its written counterpart, the SSJ UD treebank of written Slovenian,[46] which has been annotated using the same annotation scheme and thus enables direct comparison of annotations on various levels. To neutralize the effect of punctuation—an artefact in spoken language—the comparison is based on versions of the treebanks with punctuation removed. The results thus reflect the analysis of all uttered phenomena rather than all transcribed phenomena.

Vocabulary

The comparison of the vocabulary in Table 2 shows that, despite the spoken SST treebank being much smaller than its written counterpart, there are as many as 5,242 unique words (39.5% of all word types in SST) and 2,293 (30.1%) unique lemmas

46 Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek, "The Universal Dependencies Treebank for Slovenian," in *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing* (Association for Computational Linguistics, 2017), 33–38, https://doi.org/10.18653/v1/W17-1406. Kaja Dobrovoljc, Luka Terčon, and Nikola Ljubešić, "Universal Dependencies za slovenščino: Nove smernice, ročno označeni podatki in razčlenjevalni model," *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research* 11, No. 1 (2023): 218–46, https://doi.org/10.4312/slo2.0.2023.1.218-246.

featured in the SST treebank that do not occur in the written SSJ treebank, confirming previous findings[47] on the unique lexical characteristics of spoken Slovenian.[48]

Table 2: Comparison of vocabulary diversity in the spoken (SST) and written (SSJ) treebank

	SST (spoken)	SSJ (written)
Words	76.341	227.619
Word types	13.268	48.570
Unique word types	5.242	40.544
Lemma types	7.617	25.352
Unique lemma types	2.293	20.028

Source: Own work

Part-of-speech Categories

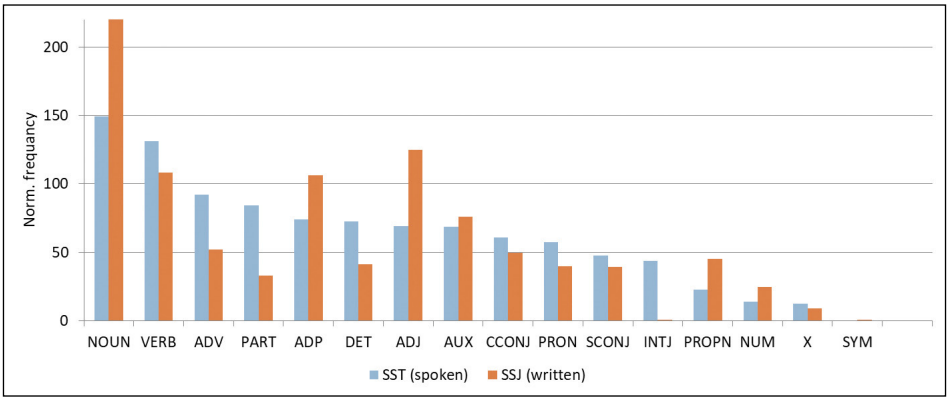
The comparison of part-of-speech tag frequencies per thousand words shown in Figure 7 reveals that the two modalities also differ with respect to the type of vocabulary used. For instance, spoken language exhibits a much higher frequency of word classes pertaining to interaction, subjectivity, deixis and modification, such as particles (PART), adverbs (ADV), interjections (INTJ), determiners (DET) and pronouns (PRON). The higher frequency of verbs (VERB) in spoken language also suggests a more dynamic narrative style, while a higher frequency of nouns (NOUN, PROPN), adjectives (ADJ) and prepositions (ADP) in written communication suggests a denser information structure and more descriptive content. Our findings confirm that spoken and written communication exhibit distinct tendencies towards nominal and verbal styles, aligning with Douglas Biber's seminal work on register variation.[49]

47 Darinka Verdonik and Mirjam Sepesy Maučec, "A Speech Corpus as a Source of Lexical Information," *International Journal of Lexicography* 30, No. 2 (June 2017): 143–66, https://doi.org/10.1093/ijl/ecw004. Kaja Dobrovoljc, "Formulaičnost v slovenskem jeziku," *Slovenščina 2.0: Empirične, aplikativne in interdisciplinarne raziskave* 6, No. 2 (2018): 67–95, https://doi.org/10.4312/slo2.0.2018.2.67-95.

48 Examples of most frequent unique lemmas in SST include filled pauses (e.g. *eee*), response tokens (e.g. *aja*), anonymized names (e.g. *[name:personal]*), and colloquial expressions (e.g. *ke*), while most frequent unique lemmas in SSJ include roman numbers (e.g. *2*), abbreviations (e.g. *dr.*), acronyms (e.g. *EU*) and culturally obsolete vocabulary (e.g. *tolar*).

49 Douglas Biber, *Variation across Speech and Writing* (Cambridge: Cambridge University Press, 1988), https://doi.org/10.1017/CBO9780511621024. Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan, *Longman Grammar of Spoken and Written English* (Berlin: De Gruyter Mouton, 2010).

Figure 7: Comparison of the distribution of POS categories in the spoken (SST) and written (SSJ) treebank



Source: Own work

Dependency relations

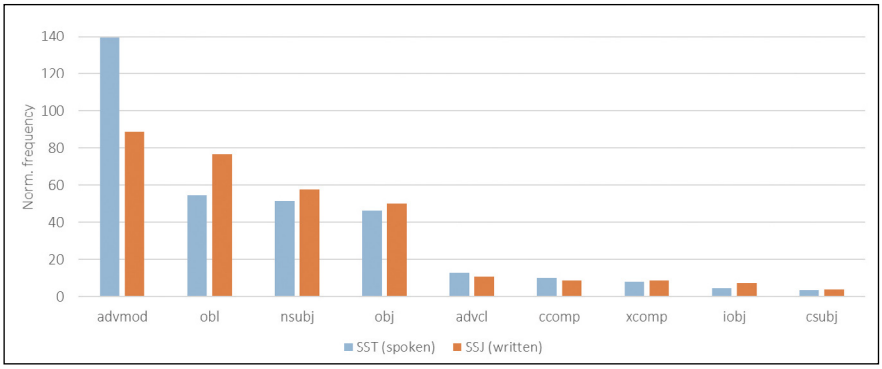
Finally, we compare the distribution of the dependency relations (syntactic functions of words) across the two datasets.

Core dependants of predicates

Figure 8 shows the comparison of the distribution of the predicate arguments, namely the nominal or clausal subjects (*nsubj, csubj*), objects (*obj, iobj, ccomp*) and adjuncts (*advmod, obl, advcl*). Interestingly, there are no major differences observed in the distribution of core arguments within each treebank, confirming that similar clause pattern strategies are used in both modalities. However, the notable differences in the frequency of some relations in both treebanks confirm the aforementioned nominal-heavy nature of written communication, i.e. more nominal subject (*nsubj*), objects (*obj, iobj*) and adjuncts (*obl*) in the written SSJ treebank. At the same time, the clauses in spoken language contain a much higher percentage of adverbial modification (*advmod*),[50] which could be explained by the abundance of modal adverbials, which speakers use to express stance, convey attitude, and balance the interaction.

50 The advmod relation is used both for modification of predicates (e.g. Pride jutri.) but also for modification of other modifier words, such as adjectives (e.g. zelo umazana posoda), so the number reflects both.

Figure 8: Comparison of core predicate arguments in the spoken (SST) and written (SSJ) treebank

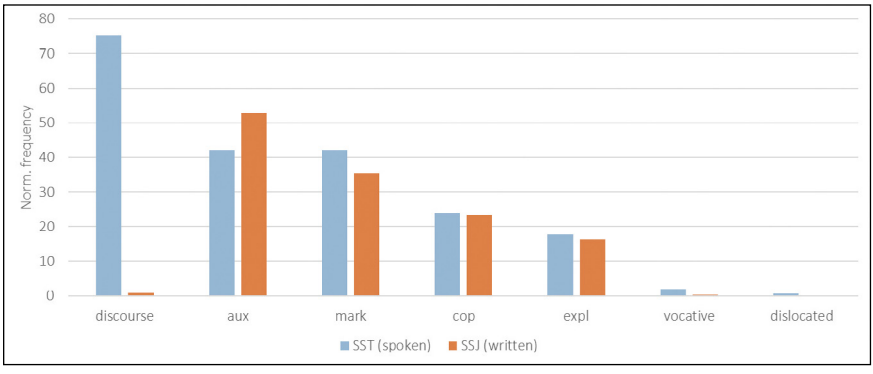


Source: Own work

Other dependants of predicates

In contrast to the much higher number of discourse elements (*discourse*), vocatives (*vocative*), and fronted or postponed elements (*dislocated*) in SST, which only rarely occur in written data, the differences in the distribution of other dependants of predicates (reported in Figure 9) are less pronounced, with two exceptions. First, spoken communication seems to show a preference for simple verbs phrases in the present tense (i.e. less auxiliary verbs marked with *aux*). Second, despite the very similar frequency of subordinate clauses in both modalities (*csubj, ccomp* and *advcl* in Figure 8 and *acl* in Figure 10), spoken data exhibits a higher number of subordinate conjunctions (*mark*). This finding requires further investigation, but may be related to the more frequent use of subordinate clauses as standalone utterances in spoken interaction—for example, as responses or elaborations on prior turns in conversation (e.g. replying *Ker dežuje* 'Because it is raining' to a question about why an event was cancelled).

Figure 9: Comparison of the non-core predicate arguments in the spoken (SST) and written (SSJ) treebank

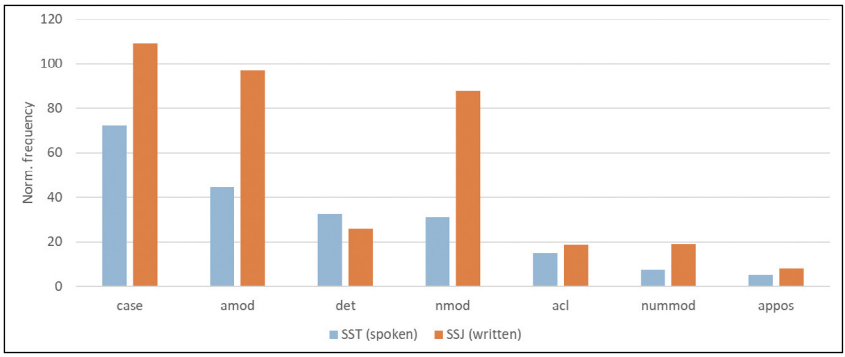


Source: Own work

Dependants of nominals

The comparison of the distribution of the relations pertaining to the dependents of nominals (e.g. noun phrase constituents) in Figure 10 shows that spoken communication exhibits a lower frequency of modifiers of nouns, such as adjectival (*amod*), nominal and prepositional (*nmod, case*), numerical (*nummod*), clausal (*acl*) and appositional (*appos*) modifiers. This is in line with the aforementioned lower number of nominal phrases in speech (Figure 7), but also suggests an overall simpler structure of such phrases (i.e. less pre- and post-modification of nouns). The only exception to this rule is the higher frequency of determiners (*det*) in SST, which can be explained by the frequent use of demonstrative pronouns and other context-grounding deictical premodifiers in speech.

Figure 10: Comparison of the dependents of nominals in the spoken (SST) and written (SSJ) treebank



Source: Own work

Other relations

Last, Figure 11 shows the comparison of the distribution for all other types of dependency relations that do not fall into any of the main syntactic categories mentioned above. Naturally, the biggest differences between both modalities can be observed for the *reparandum* relation pertaining to speech repairs, which only occur in the spoken treebank.

The second important observation is that sentences in speech are generally much shorter than in writing. This is not only reflected by the difference in the average number of words per utterance/sentence (i.e. the frequency of root elements in a treebank),[51] but also by the higher frequency of *parataxis* relation, which is used for run-on clauses with no linking conjunction.

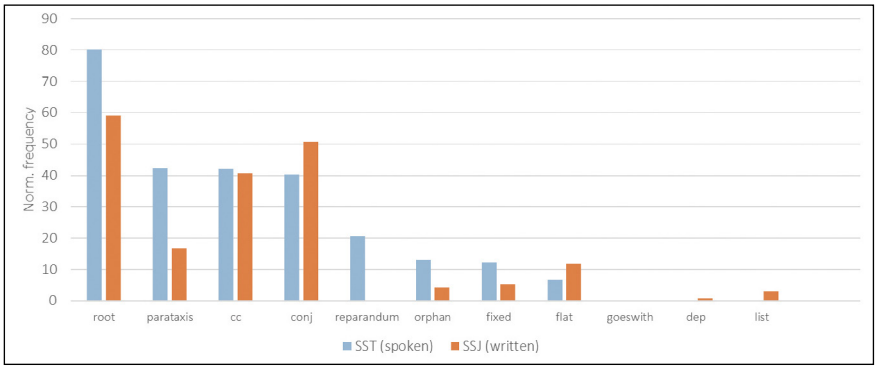
Our results also confirm the elliptical nature of spoken communication, with SST exhibiting a higher frequency of *orphan* relations, which are used to mark core

51 Average sentence length without punctuation is 12.5 tokens per utterance in SST and 17 tokens per sentence in SSJ.

arguments in cases of predicate ellipsis. We can also observe that speech features a higher number of coordinating conjunctions (*cc*) in relation to the number of coordinating conjuncts (*conj*); however, the cause might be attributed to various reasons, such as a higher number of discourse-structuring devices in speech in general (see the higher frequency of subordinating conjunctions labeled as *mark* in Figure 10) or longer coordination phrases in writing (i.e. multiple conjuncts).

Last, SST treebank also features a larger number of fixed multi-word expressions, which is in line with previous findings on the formulaic nature of this type of communication.[52] On the other hand, flat multi-word expressions (mainly encompassing personal names and foreign named entities) occur less often in speech.

Figure 11: Comparison of all other relations in the spoken (SST) and written (SSJ) treebank



Source: Own work

New Models for Grammatical Annotation of Spoken Slovenian

Finally, the new SST treebank was also used to train speech-specific models of two state-of-the-art tools for automatic grammatical annotation: Trankit[53] and CLASSLA-Stanza,[54] to complement their standard models trained solely on written data. While various speech-specific models incorporating SST data have been developed for

52 Kaja Dobrovoljc, "Formulaičnost v slovenskem jeziku," *Slovenščina 2.0: Empirične, aplikativne in interdisciplinarne raziskave* 6, No. 2 (2018): 67–95, https://doi.org/10.4312/slo2.0.2018.2.67-95.

53 Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen, "Trankit: A Light-Weight Transformer-Based Toolkit for Multilingual Natural Language Processing," in Dimitra Gkatzia and Djamé Seddah, eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations,* 80–90, Online (Association for Computational Linguistics, April 2021), https://doi.org/10.18653/v1/2021.eacl-demos.10.

54 Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc, "CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages," paper presented at the *14th Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, Ljubljana, Slovenia, September 19–20, 2024, (Institute of Contemporary History, 2024), https://doi.org/10.5281/zenodo.13936405.

experimental purposes, we present the best-performing speech-specific model(s)[55] of each tool and compare it to its standard counterparts trained only on written data. For both tools, the best-performing speech models were trained on a combination of spoken (SST) and written (SSJ/SUK) data, aligning with previous findings on the advantages of joint modelling for spoken data annotation.[56]

Thus, the following models have been featured in the evaluation:

1. The standard Trankit model[57] trained on SSJ
2. The spoken Trankit model[58] trained on SSJ and SST
3. The standard CLASSLA-Stanza models[59] trained on SSJ/SUK
4. The spoken CLASSLA-Stanza models[60] trained on SSJ/SUK and SST

We report the evaluation of all models on both written (SSJ) and spoken test sets (SST) in Table 3, using the standard F1 evaluation metric for lemmatization (LEMMA), part-of-speech tagging (UPOS), full morphology prediction (XPOS) and labelled attachment score (LAS), which measures the correct assignment of dependency heads and relations.[61] Due to space restrictions, we highlight here only the four most relevant findings among the many interesting results.

55 While Trankit follows a single-model architecture, CLASSLA-Stanza employs a modular approach with separate models for each processing layer, such as lemmatization, tagging, and parsing.

56 Dobrovoljc and Martinc, "Er ... Well, It Matters, Right?," 37–46, https://doi.org/10.18653/v1/W18-6005. Darinka Verdonik, Kaja Dobrovoljc, Tomaž Erjavec, and Nikola Ljubešić, "Gos 2: A New Reference Corpus of Spoken Slovenian," in *Proceedings of the 2024 Joint International Conference*.

57 Luka Krsnik, Kaja Dobrovoljc, and Luka Terčon, *The Trankit Model for Linguistic Processing of Standard Written Slovenian 1.1* (Slovenian Language Resource Repository CLARIN.SI, 2024), http://hdl.handle.net/11356/1963.

58 Luka Krsnik, Kaja Dobrovoljc, and Luka Terčon, *The Trankit Model for Linguistic Processing of Written and Spoken Slovenian 1.2* (Slovenian Language Resource Repository CLARIN.SI, 2024), http://hdl.handle.net/11356/1997.

59 Luka Terčon, Jaka Čibej, and Nikola Ljubešić, *The CLASSLA-Stanza Model for Lemmatization of Standard Slovenian 2.0* (Slovenian Language Resource Repository CLARIN.SI, 2023), http://hdl.handle.net/11356/1768. Nikola Ljubešić, Luka Terčon, and Jaka Čibej, *The CLASSLA-Stanza Model for Morphosyntactic Annotation of Standard Slovenian 2.0* (Slovenian Language Resource Repository CLARIN.SI, 2023), http://hdl.handle.net/11356/1767. Luka Terčon, Kaja Dobrovoljc, and Nikola Ljubešić, *The CLASSLA-Stanza Model for UD Dependency Parsing of Standard Slovenian 2.2* (Slovenian Language Resource Repository CLARIN.SI, 2025).

60 Luka Terčon, Kaja Dobrovoljc, and Nikola Ljubešić, *The CLASSLA-Stanza Model for Lemmatization of Spoken Slovenian 2.2* (Slovenian Language Resource Repository CLARIN.SI, 2025), http://hdl.handle.net/11356/2017. Luka Terčon, Kaja Dobrovoljc, and Nikola Ljubešić, *The CLASSLA-Stanza Model for Morphosyntactic Annotation of Spoken Slovenian 2.2* (Slovenian Language Resource Repository CLARIN.SI, 2025), http://hdl.handle.net/11356/2016. Luka Terčon, Kaja Dobrovoljc, and Nikola Ljubešić, *The CLASSLA-Stanza Model for UD Dependency Parsing of Spoken Slovenian 2.2* (Slovenian language resource repository CLARIN.SI, 2025), http://hdl.handle.net/11356/2018.

61 To neutralize the impact of the non-trivial task of speech segmentation, the evaluation of all models is performed on pre-segmented and pre-tokenized test sets.

Table 3: F1 performance of best-performing Trankit and CLASSLA-Stanza models for written and spoken Slovenian, evaluated on the SSJ and SST test sets. Best-performing models for each modality are marked in bold.

	SSJ-test (written)				**SST-test (spoken)**			
	Lemmas	UPOS	XPOS	LAS	Lemmas	UPOS	XPOS	LAS
Standard Trankit	98,07	99,12	98,24	95,48	98,16	95,33	93,93	79,14
Spoken Trankit	98,1	99,17	98,27	95,36	98,85	98,97	98,02	87,93
Standard CLASSLA-Stanza	98,87	98,52	96,89	90,42	98,68	92,86	91,39	69,81
Spoken CLASSLA-Stanza	98,8	98,66	96,65	90,09	99,23	98,15	96,76	81,91

Source: Own work

Performance of standard models on spoken data

Our results confirm previous findings that the performance of standard models trained on written data drops significantly when applied to transcribed speech. This is especially evident in syntactic parsing, where we observe an LAS decrease of 16.3pp for the standard Trankit model and 20.6pp for the standard CLASSLA-Stanza model, when applied to the spoken SST test set. These results reinforce that spoken data presents a significant challenge for standard off-the-shelf NLP models.

Performance of spoken models on spoken data

The performance of both tools on spoken data improves substantially when spoken (SST) data is included in training, as seen in the newly released speech-adapted models. For part-of-speech tagging and morphological feature prediction, both tools show a gain of approximately 3–5pp when using spoken models. Notably, their performance on spoken data now matches their performance on written data, achieving F1 scores of 98–99 for lemmatization, part-of-speech tagging, and full morphology prediction in both modalities.

For syntactic parsing, the improvements are even more pronounced. Compared to their written counterparts, the spoken Trankit model achieves an LAS improvement of 8.8pp, while the spoken CLASSLA-Stanza model sees a 12.1pp gain. As expected, spoken data parsing remains challenging, with an approximately 8pp gap between the best-performing parsing scores on speech and writing for both tools.[62] Nevertheless,

62 For a detailed evaluation of spoken models' accuracy with respect to specific part-of-speech tags or dependency relations—particularly relevant for targeted research applications—readers are referred to Terčon et al., 2025, published in this same journal.

these significant improvements highlight the value of SST data in enhancing spoken language processing and underscore the need for continued development of speech-aware models and tools.

Performance of spoken models on written data

Perhaps the most surprising finding is that for both tools, the newly available *spoken* models—trained on both written and spoken data—perform just as well on written data as the standard models, trained on written data alone. In other words, incorporating spoken data into training improves performance on speech without compromising performance on standard written text. This challenges the traditional *written* vs. *spoken* or *standard* vs. *domain-adapted* divide and suggests that these models should not be seen as speech-specific but rather as new state-of-the-art *universal* models, capable of delivering top-tier performance across both language modalities.

Comparing Trankit and CLASSLA-Stanza models

The timely inclusion of SST in two state-of-the-art tools for processing Slovenian is a significant step forward, as each tool has its own strengths. However, in terms of overall performance, the transformer-based Trankit generally outperforms CLASSLA-Stanza across both modalities and all metrics, except for lemmatization, where CLASSLA-Stanza has a slight advantage due to larger training set for written data (1M words in the SUK corpus compared to 270k in SSJ) and lexicon control via the Sloleks morphological dictionary. Trankit, in particular, demonstrates a notable advantage in dependency parsing, with LAS scores of the spoken-universal model reaching 95.36 on written data and 87.93 on spoken data. In comparison, CLASSLA-Stanza exhibits approximately 5–6pp lower parsing performance across both modalities.

Discussion

We have presented a new version of the reference morphosyntactically parsed corpus of spoken Slovenian, the Spoken Slovenian Treebank (SST), which has recently been extended to include more than triple the number of transcribed words and almost double the number of utterances compared to the original SST. As such, this newly available resource represents a significant addition to the Slovenian language resource landscape—ready to be exploited and further extended—and provides a valuable model for similar efforts in other languages. To support such initiatives, we share several key lessons learned during the development of SST.

First, we offer several recommendations for developing a spoken treebank resource. Anchoring the treebank in a well-established reference corpus, such as GOS, reduces annotation workload, increases the visibility and uptake of the resource, and allows direct mapping to richly transcribed speech phenomena—such as original audio recordings, layered orthographic and phonetic transcriptions, and detailed speaker or event metadata—that go well beyond what can be represented in minimalist tabular formats like CoNLL-U. Sampling is equally critical: longer and more representative speech events expand the range of linguistic phenomena that can be studied, including discourse-level structures, rather than limiting analyses to isolated sentences or speaker turns. Some limitations of the current SST still reflect overly short or fragmented sampling.

When it comes to annotation, pre-annotation with high-accuracy parsers—whether off-the-shelf or custom-trained—can significantly speed up the process by allowing annotators to focus on the more challenging structures, which are particularly frequent in the under-researched spoken language communication. However, these structures also require consistent, well-documented treatment; it is therefore crucial to clearly record annotation principles, from segmentation to specific grammatical annotation decisions. In the context of the UD annotation initiative, this means not only maintaining internal consistency, but also aligning with cross-linguistic practices and established guidelines. Current efforts within the UniDive COST action[63] are especially valuable in this regard, as they aim to harmonize treebank guidelines for spoken language across multiple languages and projects.

Second, the SST provides essential infrastructure for advancing linguistic research on spoken language. Our comparison with the written SSJ treebank shows that the differences between speech and writing are not limited to lexis, but extend to the distribution of parts of speech, syntactic relations, and overall structural organization. These distinctions highlight the need for dedicated spoken treebanks, which can reveal patterns that remain obscured in written data alone. With its rich metadata and representative sampling, the SST enables targeted investigations into how syntactic choices vary across social and contextual dimensions—such as gender, age, dialect, or communicative setting. Some of these possibilities have already been illustrated, for example in studies of self-repair strategies across private and public speech, and speaker gender.[64] Moreover, having the treebank aligned with a language-independent scheme, such as UD, opens up unprecedented possibilities for cross-linguistic investigations of spoken language grammar, enabling systematic identification of truly universal and

63 Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska, "UniDive: A COST Action on Universality, Diversity and Idiosyncrasy in Language Technology," in Maite Melero, Sakriani Sakti and Claudia Soria, eds., *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024* (Torino, Italia: ELRA and ICCL, 2024), 372–82, https://aclanthology.org/2024.sigul-1.45/.

64 Kaja Dobrovoljc, "Uporaba drevesnice SST v raziskavah govorjene slovenščine: prednosti in omejitve," *Jezik in slovstvo* 69, No. 4 (2024): 187–209, https://doi.org/10.4312/jis.69.4.187-209.

language-specific grammatical features in speech. The SST has already proven useful in such efforts, including recent comparative studies on syntactic diversity[65] and word order variation[66] across speech and writing.

Third, the distinct nature of spoken language has important implications for NLP. Our evaluation clearly shows that parsing models trained on the new SST substantially outperform standard models trained exclusively on written data when applied to transcribed speech. This highlights the practical importance of developing treebanks that reflect the characteristics of spoken communication. With the SST now available, we now have state-of-the-art models capable of accurately parsing spoken Slovenian, opening up the possibility of extending grammatical annotation to larger corpora, such as the full GOS reference corpus. Looking ahead, further improvements will depend on moving beyond transcripts and incorporating the full communicative context—including prosody, audio, and multimodal cues—to better approximate how humans process language. Since SST includes aligned recordings, much of this groundwork is already in place. More broadly, our findings contribute to the growing recognition within the NLP community that diversity in training data—including non-standard and spoken varieties—not only enhances the processing of underrepresented domains but also strengthens the performance of general-purpose models. Our findings thus suggest that investing in broadly trained parsing models on diverse data can support both accurate processing of underrepresented varieties and more inclusive data analysis.

Conclusion

In this paper, we presented the recent extension of the Spoken Slovenian Treebank with more than 3,000 new manually parsed utterances, resulting in a new, balanced and representative, version of the corpus to be used in linguistic, computational and other empirical investigations of spoken Slovenian. We made a first step in this direction by showcasing the key lexical and morphosyntactic characteristics that distinguish speech from writing, and presenting their significance for developing speech-aware NLP tools. Our findings suggest that training parsers on richly varied data—rather than restricting them to narrow domains—may be a worthwhile direction for building more inclusive and robust language processing tools.

65 Kaja Dobrovoljc, *Counting Trees: A Treebank-Driven Exploration of Syntactic Variation in Speech and Writing across Languages*, arXiv preprint (arXiv:2505.22774, 2025), https://arxiv.org/abs/2505.22774.

66 Nives Hüll and Kaja Dobrovoljc, "Word Order Variation in Spoken and Written Corpora: A Cross-Linguistic Study of SVO and Alternative Orders," in *Proceedings of SyntaxFest 2025,* 2025.

Acknowledgements

This work was financially supported by the Slovenian Research and Innovation Agency through the research projects *Treebank-Driven Approach to the Study of Spoken Slovenian* (Z6-4617), *Large Language Models for Digital Humanities* (GC-0002), and the research program *Language Resources and Technologies for Slovene* (P6-0411). In addition to the collaborators from the Mezzanine project (J74642) who have been involved with the data sampling and morphological annotation (Jaka Čibej, Tina Munda, Nikola Ljubešić, Peter Rupnik, Darinka Verdonik), we also wish to thank the data annotators (Nives Hüll, Karolina Zgaga, Luka Terčon, Matija Škofljanec) and the technical collaborators who have contributed to data pre-annotation (Luka Krsnik), audio re-segmentation (Janez Križaj, Simon Dobrišek, Tomaž Erjavec), and model evaluation (Luka Krsnik, Luka Terčon). Generative AI tools were used to support language editing during the preparation of this manuscript; full responsibility for the content remains with the authors.

Sources and Literature

Literature

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *Longman Grammar of Spoken and Written English*. Berlin: De Gruyter Mouton, 2010.

Biber, Douglas. *Variation across Speech and Writing*. Cambridge: Cambridge University Press, 1988. https://doi.org/10.1017/CBO9780511621024.

Braggaar, Anouck, and Rob van der Goot. "Challenges in Annotating and Parsing Spoken, Code-Switched, Frisian-Dutch Data." In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, edited by Eyal Ben-David, Shay Cohen, Ryan McDonald, Barbara Plank, Roi Reichart, Guy Rotman, and Yftah Ziser, 50–58. Kyiv, Ukraine: Association for Computational Linguistics, April 2021. https://aclanthology.org/2021.adaptnlp-1.6/.

Caines, Andrew, Michael McCarthy, and Paula Buttery. "Parsing Transcripts of Speech." In *Proceedings of the Workshop on Speech-Centric Natural Language Processing (SCNLP@EMNLP 2017)*, edited by Nicholas Ruiz and Srinivas Bangalore, 27–36. Copenhagen, Denmark: Association for Computational Linguistics, September 7, 2017. https://doi.org/10.18653/v1/w17-4604.

Čibej, Jaka, and Tina Munda. "Metoda polavtomatskega popravljanja lem in oblikoskladenjskih oznak na primeru učnega korpusa govorjene slovenščine ROG." Paper presented at the *14th Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, Ljubljana, Slovenia, September 19–20, 2024. Institute of Contemporary History, 2024. https://doi.org/10.5281/zenodo.13936390.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. "Universal Dependencies." *Computational Linguistics* 47, No. 2 (2021): 255–308. https://doi.org/10.1162/coli_a_00402.

Dobrovoljc, Kaja, and Joakim Nivre. "The Universal Dependencies Treebank of Spoken Slovenian." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1566–73. Portorož: European Language Resources Association, 2016. https://aclanthology.org/L16-1248/.

Dobrovoljc, Kaja, and Luka Terčon. *Universal Dependencies: Smernice za označevanje besedil v slovenščini. Različica 1.7*. Center za jezikovne vire in tehnologije Univerze v Ljubljani, 2024. https://wiki.cjvt.si/attachments/71.

Dobrovoljc, Kaja, and Matej Martinc. "Er ... Well, It Matters, Right? On the Role of Data Representations in Spoken Language Dependency Parsing." In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, edited by Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster, 37–46. Brussels, Belgium: Association for Computational Linguistics, November 2018. https://doi.org/10.18653/v1/W18-6005.

Dobrovoljc, Kaja, Luka Terčon, and Nikola Ljubešić. "Universal Dependencies za slovenščino: Nove smernice, ročno označeni podatki in razčlenjevalni model." *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research* 11, No. 1 (2023): 218–46. https://doi.org/10.4312/slo2.0.2023.1.218-246.

Dobrovoljc, Kaja, Tomaž Erjavec, and Simon Krek. "The Universal Dependencies Treebank for Slovenian." In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, 33–38. Association for Computational Linguistics, 2017. https://doi.org/10.18653/v1/W17-1406.

Dobrovoljc, Kaja. "Extending the Spoken Slovenian Treebank." In *Proceedings of the Conference on Language Technologies and Digital Humanities*, 116–46. Ljubljana, 2024. https://doi.org/10.5281/zenodo.13936393.

Dobrovoljc, Kaja. "Formulaičnost v slovenskem jeziku." *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research* 6, No. 2 (2018): 67–95. https://doi.org/10.4312/slo2.0.2018.2.67-95.

Dobrovoljc, Kaja. "Spoken Language Treebanks in Universal Dependencies: An Overview." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1798–806. Marseille: European Language Resources Association, 2022. https://aclanthology.org/2022.lrec-1.191.

Dobrovoljc, Kaja. "Uporaba drevesnice SST v raziskavah govorjene slovenščine: prednosti in omejitve." *Jezik in slovstvo* 69, No. 4 (2024): 187–209. https://doi.org/10.4312/jis.69.4.187-209.

Erjavec, Tomaž. "MULTEXT-East." In *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, 441–62. Dordrecht: Springer, 2017. https://doi.org/10.1007/978-94-024-0881-2_17.

Godfrey, John J., Edward C. Holliman, and Jane McDaniel. "SWITCHBOARD: Telephone Speech Corpus for Research and Development." In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, 517–20. IEEE, 1992. https://doi.org/10.1109/ICASSP.1992.225858.

Guillaume, Bruno. "Graph Matching and Graph Rewriting: GREW Tools for Corpus Exploration, Maintenance and Conversion." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, edited by Dimitra Gkatzia and Djamé Seddah, 168–75. Online. Association for Computational Linguistics, April 2021. https://doi.org/10.18653/v1/2021.eacl-demos.21.

Hajič, Jan, Silvie Cinková, Marie Mikulová, Petr Pajas, Jan Ptáček, Josef Toman, and Zdeňka Urešová. "PDTSL: An Annotated Resource for Speech Reconstruction." In *2008 IEEE Spoken Language Technology Workshop*, 93–96. IEEE, 2008. https://doi.org/10.1109/SLT.2008.4777848.

Hinrichs, Erhard W., Julia Bartels, Yasuhiro Kawata, Valia Kordoni, and Heike Telljohann. "The Tübingen Treebanks for Spoken German, English, and Japanese." In *Verbmobil: Foundations of Speech-to-Speech Translation*, edited by Wolfgang Wahlster, 550–74. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. https://doi.org/10.1007/978-3-662-04230-4_40.

Hinrichs, Erhard, and Sandra Kübler. "Treebank Profiling of Spoken and Written German." In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005): 9–10 December 2005, Barcelona*, edited by Montserrat Civit Torruella, Sandra Kübler, and María Antonia Martí Antonín, 65–76. Barcelona: Universitat de Barcelona, 2005.

Kahane, Sylvain, Bernard Caron, Emmett Strickland, and Kim Gerdes. "Annotation Guidelines of UD and SUD Treebanks for Spoken Corpora: A Proposal." In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, edited by Daniel Dakota,

Kilian Evang, and Sandra Kübler, 35–47. Sofia, Bulgaria: Association for Computational Linguistics, December 2021. https://aclanthology.org/2021.tlt-1.4/.

Kåsen, Andre, Kristin Hagen, Anders Nøklestad, Joel Priestly, Per Erik Solberg, and Dag Trygve Truslew Haug. "The Norwegian Dialect Corpus Treebank." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, 4827–32. Marseille, France: European Language Resources Association, June 2022. https://aclanthology.org/2022.lrec-1.516/.

Liu, Zoey, and Emily Prud'hommeaux. "Dependency Parsing Evaluation for Low-Resource Spontaneous Speech." In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, edited by Eyal Ben-David, Shay Cohen, Ryan McDonald, Barbara Plank, Roi Reichart, Guy Rotman, and Yftah Ziser, 156–65. Kyiv, Ukraine: Association for Computational Linguistics, April 2021. https://aclanthology.org/2021.adaptnlp-1.16/.

Ljubešić, Nikola, Luka Terčon, and Kaja Dobrovoljc. "CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages." Paper presented at the *14th Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, Ljubljana, Slovenia, September 19–20, 2024. Institute of Contemporary History, 2024. https://doi.org/10.5281/zenodo.13936405.

Luotolahti, Juhani, Jenna Kanerva, and Filip Ginter. "Dep_search: Efficient Search Tool for Large Dependency Parsebanks." In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, edited by Jörg Tiedemann and Nina Tahmasebi, 255–58. Gothenburg, Sweden: Association for Computational Linguistics, May 2017. https://aclanthology.org/W17-0233/.

Nguyen, Minh Van, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. "Trankit: A Light-Weight Transformer-Based Toolkit for Multilingual Natural Language Processing." In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, edited by Dimitra Gkatzia and Djamé Seddah, 80–90. Online: Association for Computational Linguistics, April 2021. https://doi.org/10.18653/v1/2021.eacl-demos.10.

Pietrandrea, Paola, and Aline Delsart. "Chapter 16. Macrosyntax at Work: Functions and Distribution of Macrosyntactic Patterns in the Rhapsodie Corpus." In *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, edited by Anne Lacheret-Dujour, Sylvain Kahane, and Paola Pietrandrea, 285–314. Amsterdam: John Benjamins Publishing Company, 2019. https://doi.org/10.1075/scl.89.17pie.

Rhapsodie: *A Prosodic and Syntactic Treebank for Spoken French*, edited by Anne Lacheret-Dujour, Sylvain Kahane, and Paola Pietrandrea. Studies in Corpus Linguistics 89. Amsterdam: John Benjamins Publishing Company, 2019. https://doi.org/10.1075/scl.89.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. "An Open Infrastructure for Advanced Treebanking." In *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, edited by Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco, 22–29. Istanbul, Turkey, May 2012.

Schuurman, Ineke, Marijke Schouppe, and Henk Hoekstra. "Harvesting Dutch Trees: Syntactic Properties of Spoken Dutch." In *Computational Linguistics in the Netherlands 2002: Selected Papers from the Thirteenth CLIN Meeting*, edited by Tanya Gaustad, 129–41. Amsterdam: Rodopi, 2003.

van der Wouden, Ton, Heleen Hoekstra, Michael Moortgat, Bram Renmans, and Ineke Schuurman. "Syntactic Analysis in the Spoken Dutch Corpus (CGN)." In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, edited by Manuel González Rodríguez and Carmen Paz Suarez Araujo. Las Palmas, Canary Islands, Spain: European Language Resources Association (ELRA), May 2002. https://aclanthology.org/L02-1071/.

Verdonik, Darinka, and Andreja Bizjak. *Pogovorni zapis in označevanje govora v govorni bazi Artur projekta RSDO*. Developement research. Maribor: University of Maribor, 2023. https://dk.um.si/IzpisGradiva.php?lang=slv&id=85198.

Verdonik, Darinka, Iztok Kosem, Ana Zwitter Vitez, Simon Krek, and Marko Stabej. "Compilation, Transcription and Usage of a Reference Speech Corpus: The Case of the Slovene Corpus GOS." *Language Resources and Evaluation* 47, No. 4 (2013): 1031–48. https://doi.org/10.1007/s10579-013-9216-5.

Verdonik, Darinka, Kaja Dobrovoljc, Tomaž Erjavec, and Nikola Ljubešić. "Gos 2: A New Reference Corpus of Spoken Slovenian." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, 7825–30. Torino, Italy: ELRA and ICCL, May 2024. https://aclanthology.org/2024.lrec-main.691/.

Verdonik, Darinka, Nikola Ljubešić, Peter Rupnik, Kaja Dobrovoljc, and Jaka Čibej. "Izbor in urejanje gradiv za učni korpus govorjene slovenščine ROG." Paper presented at the *14th Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, Ljubljana, Slovenia, September 19–20, 2024. Institute of Contemporary History, 2024. https://doi.org/10.5281/zenodo.13936425.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. "WebAnno: A Flexible, Web-Based and Visually Supported System for Distributed Annotations." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 1–6. 2013. https://aclanthology.org/P13-4001.

Other sources

Bavarian Archive for Speech Signals (BAS). *VM2 – Speech Corpus*. 2016. http://hdl.handle.net/11022/1009-0000-0000-FC55-5.

Brank, Janez. *Q-CAT Corpus Annotation Tool 1.5*. Slovenian language resource repository CLARIN.SI, 2023. http://hdl.handle.net/11356/1844.

Dutch Language Institute. *Corpus Gesproken Nederlands – CGN (Version 2.0.3)*. 2014. Data set. http://hdl.handle.net/10032/tm-a2-k6.

Godfrey, John J., and Edward Holliman. *Switchboard-1 Release 2 LDC97S62.* Web Download. Philadelphia: Linguistic Data Consortium, 1993. https://doi.org/10.35111/sw3h-rw02.

Hajič, Jan, Petr Pajas, David Mareček, Marie Mikulová, Zdeňka Urešová, and Petr Podveský. *Prague Dependency Treebank of Spoken Language (PDTSL) 0.5*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2009. http://hdl.handle.net/11858/00-097C-0000-0001-4914-D.

Krsnik, Luka, Kaja Dobrovoljc, and Luka Terčon. *The Trankit Model for Linguistic Processing of Standard Written Slovenian 1.1*. Slovenian Language Resource Repository CLARIN.SI, 2024. http://hdl.handle.net/11356/1963.

Krsnik, Luka, Kaja Dobrovoljc, and Luka Terčon. *The Trankit Model for Linguistic Processing of Written and Spoken Slovenian 1.2*. Slovenian Language Resource Repository CLARIN.SI, 2024. http://hdl.handle.net/11356/1997.

Ljubešić, Nikola, Luka Terčon, and Jaka Čibej. *The CLASSLA-Stanza Model for Morphosyntactic Annotation of Standard Slovenian 2.0*. Slovenian Language Resource Repository CLARIN.SI, 2023. http://hdl.handle.net/11356/1767.

Štravs, Miha, and Kaja Dobrovoljc. *Service for Querying Dependency Treebanks Drevesnik 1.1*. Slovenian Language Resource Repository CLARIN.SI, 2024. http://hdl.handle.net/11356/1923.

Štravs, Miha, Kaja Dobrovoljc, and Luka Bezgovšek. *Service for Querying Dependency Treebanks Drevesnik 1.2*. Slovenian language resource repository CLARIN.SI, 2025. http://hdl.handle.net/11356/2034.

Terčon, Luka, Jaka Čibej, and Nikola Ljubešić. *The CLASSLA-Stanza Model for Lemmatization of Standard Slovenian 2.0*. Slovenian Language Resource Repository CLARIN.SI, 2023. http://hdl.handle.net/11356/1768.

Terčon, Luka, Kaja Dobrovoljc, and Nikola Ljubešić. *The CLASSLA-Stanza Model for Lemmatization of Spoken Slovenian 2.2*. Slovenian Language Resource Repository CLARIN.SI, 2025. http://hdl.handle.net/11356/2017.

Terčon, Luka, Kaja Dobrovoljc, and Nikola Ljubešić. *The CLASSLA-Stanza Model for Morphosyntactic Annotation of Spoken Slovenian 2.2*. Slovenian Language Resource Repository CLARIN.SI, 2025. http://hdl.handle.net/11356/2016.

Terčon, Luka, Kaja Dobrovoljc, and Nikola Ljubešić. *The CLASSLA-Stanza Model for UD Dependency Parsing of Standard Slovenian 2.2*. Slovenian Language Resource Repository CLARIN.SI, 2025. http://hdl.handle.net/11356/2015.

Terčon, Luka, Kaja Dobrovoljc, and Nikola Ljubešić. *The CLASSLA-Stanza Model for UD Dependency Parsing of Spoken Slovenian 2.2*. Slovenian language resource repository CLARIN.SI, 2025. http://hdl.handle.net/11356/2018.

Verdonik, Darinka, Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej, Tomaž Erjavec, Tomaž Potočnik, Mirjam Sepesy Maučec, Simona Majhenič, Andrej Žgank, Andreja Bizjak, Lucija Gril, Simon Dobrišek, Janez Križaj, Marko Bajec, Iztok Lebar Bajec, Tjaša Jelovšek, Mitja Trojar, Mitja Bernjak, Naum Dretnik, Gregor Strle, Kaja Dobrovoljc, Nikola Ljubešić, and Peter Rupnik. *Spoken Corpus Gos 2.1 (Transcriptions)*. Slovenian language resource repository CLARIN.SI, 2023. http://hdl.handle.net/11356/1863.

Verdonik, Darinka, Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Tomaž Erjavec, Tomaž Potočnik, Andreja Bizjak, Andrej Žgank, Mitja Bernjak, Špela Antloga, Simona Majhenič, Peter Čakš, Matevž Pucer, Mitja Cvetko, Jani Pavlič, Simon Dobrišek, Janez Križaj, Marko Bajec, Iztok Lebar Bajec, Tjaša Jelovšek, Mitja Trojar, Naum Dretnik, David Bordon, VideoLectures.NET, and Janez Križaj. *Spoken Corpus Gos 2.1 (Audio, Video)*. Slovenian language resource repository CLARIN.SI, 2024. http://hdl.handle.net/11356/1973.

Verdonik, Darinka, and Mirjam Sepesy Maučec. "A Speech Corpus as a Source of Lexical Information." *International Journal of Lexicography* 30, No. 2 (June 2017): 143–66. https://doi.org/10.1093/ijl/ecw004.

Verdonik, Darinka, Andreja Bizjak, Andrej Žgank, Mitja Bernjak, Špela Antloga, Simona Majhenič, Peter Čakš, Matevž Pucer, Mitja Cvetko, Marijana Zelenik, Jani Pavlič, Simon Dobrišek, Janez Križaj, Gregor Strle, Marija Ivanovska, Klemen Grm, Marko Bajec, Iztok Lebar Bajec, Tjaša Jelovšek, Jure Lokovšek, Jure Longyka, Mitja Trojar, Jerneja Žganec Gros, Aleš Mihelić, Boštjan Vesnicer, Naum Dretnik, and David Bordon. *ASR Database ARTUR 1.0 (Audio)*. Slovenian language resource repository CLARIN.SI, 2023. http://hdl.handle.net/11356/1776.

Verdonik, Darinka, Anja Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej, and Tomaž Erjavec. *Spoken Corpus GOS 1.1*. Slovenian Language Resource Repository CLARIN.SI, 2021. http://hdl.handle.net/11356/1438.

Verdonik, Darinka, Kaja Dobrovoljc, Peter Rupnik, Nikola Ljubešić, Simona Majhenič, Jaka Čibej, and Thomas Schmidt. *Training Corpus of Spoken Slovenian ROG 1.0*. Slovenian language resource repository CLARIN.SI, 2024. http://hdl.handle.net/11356/1992.

Zeman, Daniel, et al. *Universal Dependencies 2.15*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2024. http://hdl.handle.net/11234/1-5787.

Kaja Dobrovoljc

DREVESNICA GOVORJENE SLOVENŠČINE: NOVI PODATKI, MODELI IN KLJUČNI NAUKI

POVZETEK

Prispevek predstavlja novo različico drevesnice govorjene slovenščine (SST), skladenjsko razčlenjenega korpusa transkribiranega spontanega govora, ki je bil nedavno razširjen z več kot 3.000 novimi izjavami oz. več kot 60.000 pojavnicami. Nova različica SST temelji na referenčnem korpusu GOS 2.1 in vključuje 344 govornih dogodkov z raznoliko zastopanostjo govorcev in sporazumevalnih okoliščin. Po kratkem pregledu postopkov vzorčenja, označevanja in končnega usklajevanja podatkov predstavimo primerjalno analizo med govorno drevesnico SST in pisno drevesnico SSJ, ki potrjuje številne leksikalne, oblikoslovne in skladenjske posebnosti govorjenega jezika v primerjavi s pisnim. V govorjeni slovenščini tako najdemo več struktur, povezanih z interakcijo, sprotnim načrtovanjem govora, subjektivnostjo, deiktičnostjo, modifikacijo, elipso in strukturiranjem diskurza, po drugi strani pa manj samostalniških zvez, ki so tudi bolj preprosto sestavljene.

Glede na to, da je bila drevesnica SST kmalu po objavi že uporabljena za razvoj novih modelov za slovnično označevanje (transkripcij) govorjene slovenščine v orodjih CLASSLA-Stanza in Trankit, v nadaljevanju predstavimo sistematično primerjavo modelov, naučenih zgolj na pisnih besedilih, in modelov, naučenih tako na pisnih kot govorjenih besedilih. Rezultati kažejo, da so pri slovničnem označevanju govora modeli, naučeni na kombinaciji govorjenih in pisanih podatkov, bistveno boljši od modelov, naučenih zgolj na pisnih besedilih, zlasti pri nalogi skladenjskega razčlenjevanja. Obenem so ti 'mešani' modeli tudi pri označevanju pisnih besedil ohranjajo enako stopnjo natančnosti kot standardni pisni modeli, na podlagi česar lahko sklenemo, da gre za robustne univerzalne modele, ki dosegajo najboljše možne rezultate tako na pisnih kot govorjenih besedilih.

V diskusiji rezultate povzamemo in jih nadgradimo z razpravo o najpomembnejših izkušnjah, pridobljenih med razvojem drevesnice SST, ki letos obeležuje že deset let od prve objave. Izpostavimo prednosti uporabe referenčnega korpusa kot izhodišča, priporočimo vzorčenje daljših, zaključenih govornih besedil in sistematično dokumentiranje označevalnih smernic ter poudarimo pomen usklajenosti z mednarodnimi označevalnimi pobudami, kot je shema Universal Dependencies. V sklepnem delu neprecenljiv metodološki potencial tega jezikovnega vira ponazorimo z omembo več aktualnih raziskav in številnimi možnostmi nadaljnjih raziskav tako na področju jezikoslovja kot jezikovnih tehnologij.

1.01 DOI: https://doi.org/10.51663/pnz.65.3.02

Ajda Pretnar Žagar*

Računalniška analiza slovenskih zgodovinskih časopisov (1771–1914): jezikovni, tematski in državotvorni uvidi

IZVLEČEK

Prispevek predstavlja računalniško-jezikoslovno analizo sPeriodike, zgodovinskega korpusa slovenskih periodičnih publikacij, izdanih med letoma 1771 in 1914. Z analizo ključnih besed ter diahrono analizo smo raziskali jezikovne, tematske in zgodovinske razsežnosti desetih najvidnejših časopisov v korpusu. Ugotovitve razkrivajo osrednjo vlogo teh časopisov pri oblikovanju slovenskega narodnega prebujanja v obdobju po letu 1848, hkrati pa poudarjajo raznolike tematske usmeritve posameznih periodičnih publikacij, kot so kmetijstvo, pedagogika, književnost in oglaševanje. Poleg tega raziskava obravnava izzive, ki jih prinaša slaba kakovost optičnega prepoznavanja znakov (OCR) pri digitalizaciji zgodovinskih besedil, ter njihove posledice za jezikovno in vsebinsko analizo. Združevanje računalniških metod z zgodovinskim raziskovanjem v tej študiji ponuja vpogled v razvoj slovenskega jezika, vlogo medijev pri oblikovanju narodne identitete in možnosti za izboljšanje besedilnih virov, temelječih na OCR.

Ključne besede: zgodovinski časopisi, analiza ključnih besed, napake OCR, korpusno jezikoslovje

* **Dr., asistent z doktoratom, Inštitut za novejšo zgodovino, Privoz 11, SI-1000 Ljubljana, ajda.pretnar@inz.si; ORCID: 0000-0002-5927-4538**

ABSTRACT

COMPUTATIONAL ANALYSIS OF SLOVENIAN HISTORICAL NEWSPAPERS (1771–1914): LINGUISTIC, THEMATIC, AND NATION-BUILDING INSIGHTS

This paper presents a computational linguistic analysis of sPeriodika, a historical corpus of Slovenian periodicals published between 1771 and 1914. Using keyword analysis and diachronic analysis, we explore the linguistic, thematic, and historical dimensions of ten prominent newspapers in the corpus. Our findings reveal the centrality of these newspapers in shaping Slovenian nation-building during the post-1848 period, while also highlighting the diverse thematic orientations of individual periodicals, including agriculture, pedagogy, literature, and advertising. Moreover, the study examines the challenges posed by low-quality Optical Character Recognition (OCR) in historical text digitisation and its implications for linguistic and content analysis. By combining computational methods with historical inquiry, this research provides insights into the evolution of the Slovenian language, the media's role in nation-building, and the potential for improving OCR-based textual resources.

Keywords: historical periodicals, keyword analysis, OCR errors, corpus linguistics

Uvod

V zadnjem desetletju smo priča porastu raziskav zgodovinskih časopisov.[1] Rast je posledica vse večjega priznanja zgodovinskih časopisov kot dragocenih primarnih virov, ki ponujajo vpogled v pretekle družbe, kulture in dogodke. Raziskave pokrivajo širok spekter aplikacij, od digitalizacije zgodovinskih časopisov in ustvarjanja obsežnih visokokakovostnih digitalnih korpusov do naprednih računalniških pristopov za analizo jezikovnih sprememb, sentimenta in diskurza v zgodovinskih kontekstih.

Hkrati se sodobne metodologije vse bolj prilagajajo specifičnim izzivom zgodovinskih časopisov, kot so degradirana besedila, nekonsistenten zapis besed in večjezične zbirke. Ti pristopi preoblikujejo obdelavo zgodovinskih časopisov v interdisciplinarno področje, ki povezuje digitalno humanistiko, računalništvo in arhivske študije.

sPeriodika[2] je nedavno objavljen korpus zgodovinskih slovenskih periodičnih publikacij iz obdobja 1771–1914. Korpus je obsežen in temelji na digitaliziranih časopisih iz digitalne knjižnice dLib, ki jo upravlja Narodna in univerzitetna knjižnica

1 Maud Ehrmann et al., »Computational Approaches to Digitised Historical Newspapers,« *Dagstuhl Reports* 12, št. 7 (2023): 112–79, pridobljeno 5. 2. 2025, https://doi.org/10.4230/DagRep.12.7.112.

2 Filip Dobranić et al., »Corpus of Slovenian periodicals (1771–1914) sPeriodika 1.0«(2023), *Slovenian language resource repository CLARIN.SI,* http://hdl.handle.net/11356/1881.

Slovenije. Vsebuje nekatere najpomembnejše časopise tistega časa, ki so prispevali k večji pismenosti in narodnemu prebujanju v Sloveniji.[3, 4]

Prispevek je korpusno-jezikoslovna študija korpusa sPeriodika in predstavlja dopolnitev ter prevod prispevka na konferenci JTDH.[5] Razširitev zajema dodatno poglavje o zgodovinskem razvoju jezika z analizo arhaičnih besed (razdelek 3.3), analiza napak OCR pa je razširjena v samostojno poglavje. Izbrali smo deset časopisov z največjim številom izdaj in izvedli osnoven kvantitativni pregled vsebine. Kakovost optične prepoznave znakov (OCR) v korpusu je nizka, a primerljiva s podobnimi zgodovinskimi digitaliziranimi časopisi,[6] zato nas je zanimalo, ali lahko kljub temu izluščimo značilnosti časopisov s pomočjo analize ključnih besed, pogostosti besed in konkordanc. V rezultatih podamo splošen kvantitativni opis časopisov, vpogled v zgodovinski razvoj slovenskega jezika in pregled napak OCR. Z raziskavo poudarimo pomen označenih zgodovinskih izdaj za slovensko raziskovalno skupnost, saj bi brez digitalno dostopnega in označenega korpusa tak pregled težko izvedli.

Sorodna dela

Zgodovinski časopisi se pogosto uporabljajo v digitalni humanistiki, predvsem zaradi sodobnih prizadevanj za digitalizacijo, dostopnih vmesnikov za raziskovanje vsebine[7] in odprtih repozitorijev. Raziskave zajemajo širok spekter, od diahronih in primerjalnih analiz do diskurzivnih študij, pri čemer je analiza premika konceptov ena izmed najvidnejših metod. Primerjalne študije se osredotočajo na primerjave med državami[8] ali raziskovanje regionalnih razlik.[9] Diahrone študije pogosto raziskujejo

3 Marijan Dović, »Literatura in mediji v Jurčičevem času,« *Slavistična revija* 54, št. 4 (2006): 543–57.

4 Smilja Amon, »Vloga slovenskega časopisja v združevanju in ločevanju slovenske javnosti od 1797–1945,« *Javnost* 15 (2008): S9–S24.

5 Ajda Pretnar Žagar, »A corpus linguistic characterization of speriodika,« v: *Proceedings of the Conference on Language Technologies and Digital Humanities* (Ljubljana: Inštitut za novejšo zgodovino 2024), 384–406.

6 Kimmo Kettunen in Tuula Pääkkönen, »Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means,« v: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (Portorož: ELRA, 2016), 956–61.

7 Maud Ehrmann et al., »Historical Newspaper User Interfaces: A Review,« v: *85th IFLA General Conference and Assembly (IFLA)* (Zenodo, 2019).

8 Adán Mayer et al., »Underlying sentiments in 1867: A study of news flows on the execution of Emperor Maximilian I of Mexico in digitized newspaper corpora,« *Digital Humanities Quarterly* 16, št. 4 (2022).

9 Jaihyun Park in Ryan Cordell, »A quantitative discourse analysis of Asian workers in the US historical newspapers,« v: *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages* (Tokio: Association for Computational Linguistics, 2023), 7–15.

premike konceptov,[10, 11, 12] semantične spremembe[13] ali spremembe tematik skozi čas.[14] Druga veja raziskav vključuje vsebinsko usmerjen pristop, ki se osredotoča na nastanek javnih diskurzov[15] ali državotvorno besedišče.[16, 17] Nekatere raziskave se osredotočajo tudi na večjezičnost,[18, 19] ki je značilna za zgodovinske časopise in otežuje primerjalno analizo.

Izven digitalne humanistike so slovenski zgodovinski časopisi priljubljena tema raziskav. Večina teh se osredotoča na procese narodnega prebujanja, zlasti po marčni revoluciji leta 1848.[20, 21] Najobsežnejšo študijo je izvedla Smilja Amon,[22] ki predstavlja pregled slovenskega novinarstva. *Ljubljanski zvon* iz leta 1885 ponuja podroben pregled časopisov tistega časa,[23] pri čemer navaja 34 časopisov v slovenščini, skupaj z opisi, uredniki, izdajatelji in cenami. Druge raziskave se večinoma osredotočajo na *Kmetijske in rokodelske novice*,[24] ki so postavile temelje slovenskemu novinarstvu.[25] Jezikovne analize so prav tako pogoste, le malo raziskav pa se posveča vsebinski analizi in primerjavam. Ena takih je analiza Štepca[26] (1987), ki obravnava poročanje o zločinih v *Slovencu* in *Slovenskem narodu*. Štepec ugotavlja, da konservativni *Slovenec*

10 Japp Verheul et al., »Using word vector models to trace conceptual change over time and space in historical newspapers 1840–1914,« *Digital Humanities Quarterly* 16, št. 2 (2022).

11 Jani Marjanen et al., »The Expansion of Isms, 1820–1917: Data-Driven Analysis of Political Language in Digitized Newspaper Collections,« Journal of Data Mining & Digital Humanities 2020, https://doi.org/10.46298/jdmdh.6159.

12 Lidia Pivovarova et al., »Word Clustering for Historical Newspapers Analysis,« v: *Proceedings of the Workshop on Language Technology for Digital Historical Archives* (Varna: INCOMA Ltd., 2019), 3–10.

13 Nilo Pedrazzini in Barbara McGillivray, »Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers,« v: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities* (Tajpej: Association for Computational Linguistics, 2022), 85–95.

14 Jani Marjanen et al., »Topic Modelling Discourse Dynamics in Historical Newspapers,« v: *Digital Humanities in the Nordic Countries 2020* (CEUR-WS.org, 2021), 63–77.

15 Jani Marjanen et al., »A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917,« *Journal of European Periodical Studies* 4, št. 1 (2019).

16 Jonathan Schoots, »Analyzing political formation through historical isiXhosa text analysis: Using frequency analysis to examine emerging African nationalism in South Africa,« v: *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023)* (Dubrovnik: Association for Computational Linguistics, 2023), 65–75, https://doi.org/10.18653/v1/2023.rail-1.8.

17 Simon Hengchen et al., »A Data-Driven Approach to Studying Changing Vocabularies in Historical Newspaper Collections,« *Digital Scholarship in the Humanities* 36, dodatek 2 (2021): ii109-ii126, https://doi.org/10.1093/llc/fqab032.

18 Marjanen, »A National Public Sphere?«

19 Mayer, »Underlying sentiments in 1867.«

20 Obdobje pred marčno revolucijo leta 1848 se običajno imenuje predmarčno obdobje. V članku obdobje po revoluciji imenujemo pomarčno.

21 Nataša Stergar, »Narodnostno vprašanje v predmarčnih letnikih Bleiweisovih Novic,« *Kronika* 25, št. 3 (1977).

22 Amon, »Vloga slovenskega časopisja v združevanju in ločevanju slovenske javnosti.«

23 Anonymous, »Slovenski časopisi leta 1885,« *Ljubljanski zvon* 5, 1885, 631–35.

24 Stane Mihelič, »Kmetijska družba in ustanovitev 'Novic',« *Slavistična revija* 1, št. 1/2 (1948).

25 Prva slovenska periodična publikacija so bile *Lublanske novize* Valentina Vodnika leta 1797, a niso izhajale dolgo.

26 Marko Štepec, »Zločin v slovenskem časopisju v 80. letih 19. stoletja,« *Kronika* 35, št. 1/2 (1987): 30–38.

prepušča poročanje o zločinih liberalnemu *Slovenskemu narodu,* saj to vidi kot nekatoliško in nepotrebno. Druge raziskave obravnavajo jezikovno vprašanje v *Slovenskem pravniku,*[27] novice o Istri,[28] modo v ženskih časopisih[29] in socialnodemokratsko periodiko.[30]

sPeriodika

sPeriodika[31] je korpus slovenskih zgodovinskih časopisov, izdanih med letoma 1771 in 1914. Korpus je ustvaril Dobranić s sodelavci,[32] temelji pa na optično prepoznanih zapisih, ki so jih v različnih obdobjih z različnimi tehnologijami ustvarili v Narodni in univerzitetni knjižnici Slovenije, pri čemer so avtorji izvedli dodatno čiščenje in predobdelavo. Korpus je na voljo v repozitoriju CLARIN.SI[33] in v konkordančniku NoSketch Engine.[34]

Opis

Korpus sPeriodika vsebuje 216 časopisov z različnim številom izdaj (največ 28.406, najmanj 1). Skupno število izdaj je 148.457. Kot prikazuje Slika 1, se je aktivnost izdajanja postopoma povečevala do prve svetovne vojne, ko je večina časopisov prenehala izhajati. Zadnje desetletje vključuje podatke samo do leta 1914, kar pojasnjuje upad frekvence.

27 Tone Zorn, »Odmevnost jezikovnega vprašanja v listu Slovenski pravnik v letih 1871–1918,« *Kronika* 35, št. 3 (1987): 146–55.

28 Branko Marušič, »Izbor vesti o Istri v slovenskem časopisju do leta 1880,« *Annales* 17, št. 1 (2007): 65–82.

29 Maja Ilich, »Nekaj o modi v slovenskem časopisju na prelomu stoletja (1895–1915),« *Zgodovina za vse* 6, št. 2 (1999): 98–108.

30 Dušan Kermavner, »Drugi slovenski socialnodemokratski listi,« *Kronika* 10 (1962): 80–89.

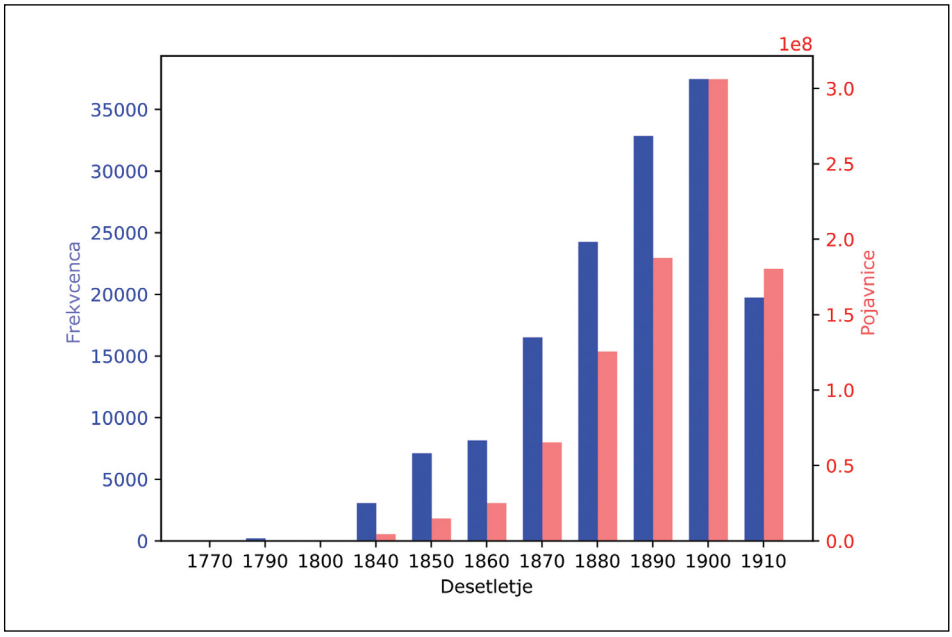
31 Dobranić et al., »*Corpus of Slovenian Periodicals (1771–1914) sPeriodika 1.0*«.

32 Filip Dobranić et al., »A Lightweight Approach to a Giga-Corpus of Historical Periodicals: The Story of a Slovenian Historical Newspaper Collection,« v: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (Italija: ELRA in ICCL, 2024).

33 CLARIN.SI, http://hdl.handle.net/11356/1881.

34 NoSketch Engine, https://www.clarin.si/ske/#dashboard?corpname=speriodika.

Slika 1: Frekvenca izdaj (modri stolpci) in pojavnic (rdeči stolpci) po desetletjih v sPeriodiki



Vir: avtorica iz podatkov NoSketchEngine

Zaradi dolgega repa v distribuciji izdaj po časopisih smo se odločili analizirati deset časopisov z največ izdajami, kar predstavlja 78 odstotkov korpusa. Takšno merilo smo izbrali, da zajamemo časopise z največjim nacionalnim dosegom in dovolj dolgim časovnim razponom. Tabela 1 prikazuje deset izbranih časopisov s številom in deležem izdaj (zaokroženo na dve decimalni mesti).

Naslovi časopisov nosijo pomenske poudarke, ki na splošno določajo njihovo vsebino: *Kmetijske in rokodelske novice*, *Slovenski gospodar*, *Učiteljski tovariš*, *Slovenski narod*, *Dom in svet*, *Slovenec*, *Edinost*, *Ljubljanski zvon*, *Vertec* in *Soča*.

Primerjava ključnih besed

Za obravnavane časopise smo s pomočjo orodja NoSketch Engine izluščili ključne besede. Te smo primerjali s sPeriodiko, kar pomeni, da smo izluščili leme, ki so v določenem časopisu močno zastopane in zato statistično značilne. Lematizacija je bila izvedena s postopkom CLASSLA-Stanza, kot je navedeno v izvirnem članku o sPeriodiki.[35] Ključnost (angl. *keyness*) je v NoSketch določena na osnovi *enostavne matematične metode*[36] s parametrom glajenja N = 1 (privzeta nastavitev).

35 Dobranić et al., »A Lightweight Approach to a Giga-Corpus of Historical Periodicals.«

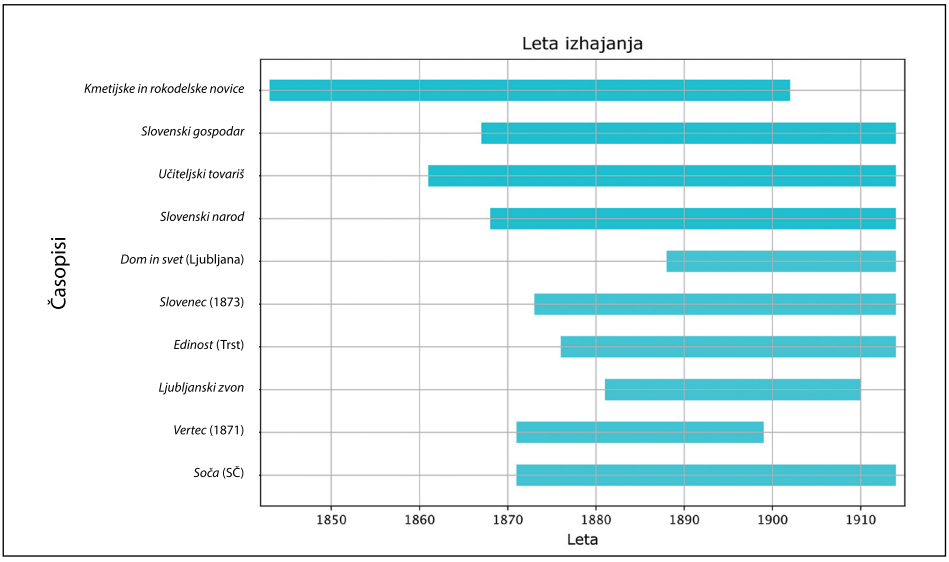
36 Adam Kilgarriff, »Simple Maths for Keywords,« v: *Proceedings of Corpus Linguistics* 6 (Liverpool, VB: University of Liverpool, 2009).

Tabela 1: Časopisi z največjim številom izdaj v korpusu sPeriodika

Časopis	št. objav	% objav	št. pojavnic
Kmetijske in rokodelske novice (KRN)	28406	19	29.834.568
Slovenski gospodar (SG)	16009	11	22.602.374
Učiteljski tovariš (UT)	15674	11	24.337.225
Slovenski narod (SN)	14039	9	183.294.799
Dom in svet (DS)	11073	7	32.326.449
Slovenec (SVN)	10897	7	137.506.802
Edinost (ED)	8371	6	98.274.429
Ljubljanski zvon (LZ)	3923	3	15.590.800
Vertec (VT)	3515	2	3.170.465
Soča (SČ)	3367	2	38.879.707

Vir: NoSketchEngine

Slika 2: Leta izdajanja za deset izbranih časopisov



Vir: NoSketchEngine

Analizirali smo prvih sto ključnih besed in jih predstavili v Tabeli 2. Očitne napake OCR smo izključili, saj želimo prikazati osrednjo vsebino časopisa, ne naključnih napak. Poročamo tudi o odstotku napak OCR (delež napak med 100 zadetki).

Kmetijske in rokodelske novice

Kmetijske in rokodelske novice so zveste svojemu imenu, saj obravnavajo kmetijske teme (*kmetovavec, žlahen, žebec*[37]) ter lokalne novice (*Kranjska*). Časopis je bil prvi polnopravni časnik v slovenščini, zato vsebuje več arhaičnih besed (*onidan, en malo*) kot drugi časopisi. Preostale ključne besede sodijo v raznolike kategorije, od rubrik v časopisu (*novičar*) in financ (*dnar*) do novic o Rusiji (*rusovski*) ter narodno-prosvetnih tem (*čitavnica*[38]). Analiza ključnih besed kaže širok spekter tem, ki jih je časopis pokrival, ter njegovo dolgoletno osrednjo vlogo v kulturnem življenju Slovencev.[39]

Slovenski gospodar

Slovenski gospodar je prvi časopis na seznamu, ki ga močno zaznamujejo napake OCR (94 odstotkov[40]). Pregled v konkordančniku pokaže, da je črka »n« pogosto prepisana kot »a« (*sloveaski* –> slovenski, *aaš* –> naš, *aemški* –> nemški), črka »v« pa kot »7« (*pra7*). Druge ključne besede razkrivajo, da je pogosto napačna tudi zamenjava »č« za »6«. Omenja se tudi izraz Stajerc, ki je napačna oblika besede Štajerc. Pojem lahko pomeni prebivalca Štajerske, vendar se najpogosteje nanaša na časopis *Štajerc*, ki je izhajal med letoma 1900 in 1918. Ton je precej žaljiv, saj je bil *Slovenski gospodar* katoliški in konservativen časopis, medtem ko je bil *Štajerc* napreden pronemški časnik (podrobneje opisan v Jezernik[41]). Pomenske ključne besede se nanašajo na sejme (*sermon*), dogajanje (*izgoditi*), zlatnike (*fl*), šolsko zvezo (*šulverein*), ljudi (*poslanec dr. Franc Radaj*; *Franc Kosar*), spoštovane (*vlč, velečastiti*) in *posilinemce* (posmehljiv izraz za pronemške Slovence).

Učiteljski tovariš

Učiteljski tovariš je zvest svojemu imenu. Večina ključnih besed se nanaša na pedagogiko (*zavezin*[42] *konvikt,*[43]*učiteljstvo, učiteljski, lehrerbund, pedagoški, koleginja, ljudski*). V razpravah je opaziti politični vidik, saj se pogosto omenja »*Slomškar*«, kar se nanaša na konkurenčno »Slomškovo zvezo«, zvezo katoliških učiteljev. Pri besedi »*tovarišica*« niti iz kolokacij ni jasno, ali ima političen prizvok. Vendar pa sta obe sklicevanji na ženske kolegice (*tovarišica* in *koleginja*) v *Učiteljskem tovarišu* močno zastopani, kar morda kaže na to, da je časopis ženskam prisojal večjo stopnjo enakopravnosti. Pogostost omenjenih besed je namreč v tem časopisu bistveno večja v primerjavi s

37 Arhaično za žrebec.

38 Čitavnica je pogostejša v zgodnjih izdajah KRN, kasneje pa jo nadomešča izraz čitalnica.

39 Stergar, »Narodnostno vprašanje v predmarčnih letnikih Bleiweisovih Novic.«

40 Stopnja napake 94 odstotkov se nanaša na rezultate analize ključnih besed in ne na celotno vsebino časopisa.

41 Božidar Jezernik. »Katoliška duhovščina na prelomu devetnajstega in dvajsetega stoletja in proces modernizacije na Slovenskem,« *Traditiones* 51, št. 1 (2022): 103–45.

42 Zaveza se nanaša na Zvezo avstrijskih jugoslovanskih učiteljskih društev.

43 Konvikt je izobraževalni zavod s celodnevno oskrbo, predvsem za duhovnike.

splošnim korpusom, vendar kolokacije ne razkrivajo posebnih razlik v kontekstu. Učiteljski tovariš prav tako vsebuje veliko nemških izposojenk (*Lehrerbund, Lehrer, Volkschule, Lehrerschaft, Gesuche, Vorgeschriebenen*) in omembe oseb (*Črnagoj, Jelenc, Maier, Strmšek, Režek, Požegar, Gangl*).

Slovenski narod

Analiza ključnih besed časnika *Slovenski narod* razkriva številne specifične rubrike. Časopis je redno objavljal železniške vozne rede za avstrijske železnice (*amstetten, pontabel, selzthal*), poročila z dunajske borze (*prior oblig.*), meteorološka poročila (smeri vetrov) in specifične oglase (*Moll Seidlitz prašek, Revaliescere du Barry, Berger Kotran milo*). Nekatere besede se nanašajo na uvodni odstavek časopisa, ki je vseboval navodila za pošiljanje prispevkov (*izvoti*,[44] *četiristopne*). Opazili smo tudi nekatere za *Slovenski narod* značilne napake OCR, ki so morda posledica izbire pisave (*tuđi, tuđ*,[45] *ćel*[46]). Nekateri rezultati so morda posledica prekomernega popravljanja, saj Dobranić in sodelavci[47] omenjajo statistično osnovano združevanje razdeljenih besed (*Trammwaydrušt, Stražatoplice*).

Dom in svet

Dom in svet (Ljubljana) je močno literarno in umetniško usmerjen. Za časopis so značilna imena literarnih junakov (*bodriški nadknez Gotšalk, Viljenica, Virida, Maruška, Ančka*) in avtorjev zgodb (*Podgoričan*), ki jih je časopis stalno objavljal. Velik del njihovih novic omenja umetniška dela (*spominiki, bilina, pasionski*) in publikacije (besedilo o klinopisnih spomenikih, ki ga je napisal F. Sedej in je bilo objavljeno v istem časopisu). Najpresenetljivejši je močan vpliv slovanskega umetniškega sveta na časopis. *Dom in svet* redno objavlja biografije srednje-, vzhodno- in južnoslovanskih avtorjev ter seznam slovanskih publikacij (zlasti ruskih, srbskih in hrvaških).

Slovenec

Podobno kot *Slovenski narod* tudi analiza ključnih besed časnika *Slovenec* razkriva specifične rubrike, na primer poročila z dunajske borze (*vravnaven, salmov, dunavski, napoleondor, napoleond*,[48] *waldsteinov*), meteorološka poročila in podlistek Pismo Boltatovega Pepeta,[49] napisan v narečju (*gespud, tku, kokr*). Med ključnimi besedami so tudi oglasi, na primer za Merkur Exchange Limited Company (*kurzen*), steklarske

44 Napačen leksem besede »izvoliti«.
45 V pomenu »tudi«.
46 V pomeni »celo« ali »čelo«.
47 Dobranić et al., »A Lightweight Approach to a Giga-Corpus of Historical Periodicals.«
48 Obe pojavnici predstavljata dobesedni prepis izraza za francoski zlatnik »napoléon d'or«.
49 Psevdonim za Srečka Magoliča.

delavnice in trgovino z oljnimi barvami. Nekaj ključnih besed se nanaša na jugovzhodno Evropo (*Hrvaška, Madžarska, Bolgarija*), kar delno nakazuje politično usmeritev časopisa. Vendar smo glede na politično pomembnost časnika v slovenskem prostoru pričakovali večji delež političnih besed. Mnogo ključnih besed izhaja iz glave časopisa, kjer so bile podane praktične informacije o naročilu in distribuciji časopisa. Vendar so tudi drugi časopisi, kot so *Slovenski narod*, *Slovenski gospodar*, *Edinost* in *Soča*, imeli obsežne glave. Visoka pogostost ključnih besed iz glave je morda posledica jezikovnih značilnosti glave časopisa *Slovenec*.

Edinost (Trst)

Edinost (Trst), vodilni časopis tržaških Slovencev, vsebuje veliko besed, povezanih z oglasi. 68 odstotkov ključnih besed se nanaša na ulice ali kraje poslovanja (*barriera, nuova, vecchia, piazza, galatti*). Večinoma gre za italijanska imena ulic, a so omenjeni tudi istrski kraji (*Pula, Rovinj*). *Edinost* je pokrivala istrsko regijo do leta 1902, ko je bilo ustanovljeno Politično društvo Hrvatov in Slovencev v Istri.[50] Pri omembi Primorske se večina pojavnic nanaša na vremensko napoved in podnaslov časopisa (Glasilo političnega društva »Edinost« za Primorsko). Omenjeni so tudi denarni izrazi (*nvč* je okrajšava za »novčič«, kovanec v vrednosti 1/100 zlatnika) in prostor za oglase (*inseratni* označuje oddelek časopisa za oglase). Oglasi vključujejo ponavljajoče se reklame za kavo (*kava Santos good average*), zdravstvene storitve (*izdiranje, plombiranje, ambulatorij*) in živila (*pekarna, butejka*). Podobno kot drugi časopisi tistega časa je *Edinost* redno objavljala železniške vozne rede. Besedi »*Medpostaja*« in »*Pula*« sta največkrat uporabljeni v kontekstu železniških voznih redov, podobno kot v *Slovenskem narodu*, vendar osredotočeno na italijanske železnice. Novice o železniških voznih redih kažejo, da je bil časopis zelo praktičen; ponujal je oglaševalski prostor za lokalna podjetja in podajal informacije o prevozu. Mnogi časopisi tistega časa so imeli podobne vsebine.

Ljubljanski zvon

Ljubljanski zvon je bil vodilna literarna revija pomarčne dobe. Večina desetih najpogostejših ključnih besed se nanaša na literarne like (*Gojko, Samorad, Trenk, Abadon, Zdenka*). 29 odstotkov ključnih besed predstavljajo imena literarnih likov, kar poudarja literarno naravo revije. Vendar vsebine niso bile zgolj leposlovne. Omenjeni so denimo Slovniški razgovori, kjer je revija objavljala nasvete o pravilnem slovenskem črkovanju in slovnici (*sedanjik, sgl, Miklošič, dovršnik*), in Štrekljeve jezikoslovne mrvice, kjer je avtor razlagal slovnično sestavo, pomen in izvor določenih besed (*subst*). Veliko ključnih besed je posledica napak OCR, natančneje 36 odstotkov. Težava s ključnimi besedami pri *Ljubljanskem zvonu* je nekoliko posebna. Podobno kot pri *Slovenskem gospodarju* so najpogostejše napačno transkribirane besede.

50 Darko Darovec, *Pregled zgodovine Istre* (Koper: Zgodovinsko društvo za južno Primorsko, Založba Annales; Čentur: Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja, 2023), 66.

Te napake so tesno povezane z literarno naravo revije. *Ljubljanski zvon* je namreč edini analizirani časopis, ki dosledno uporablja naglase na samoglasnikih. Naglasi v slovenščini niso pogosti, vendar so bili v tej reviji verjetno uporabljeni za poudarjanje ritma in pravilne izgovarjave besed, ta slogovna izbira pa povzroča težave modelu OCR.

Vertec (1871)

Vertec (1871) vsebuje veliko zgodb in je tako podoben *Domu in svetu* ter *Ljubljanskemu zvonu*, saj ga zaznamujejo literarni liki (*Marijca, Marijec,*[51] *Katarinka, Ivanek*). Delež omemb literarnih likov med ključnimi besedami je 38-odstoten. V primerjavi z drugimi periodičnimi publikacijami so imena pretežno pomanjševalnice, kar odraža usmeritev časopisa na mlajše bralce. Vendar pa ime včasih ne označuje literarnih likov, temveč resnične osebe. Časopis je namreč poimensko navajal avtorje pravilnih rešitev ugank, skupaj z lokacijo. Druge ključne besede so idilične, povezane z družino ali naravo (*dedek, sestrica, ptičica, čmrlj, lisica*). Stopnja napak OCR pri tem časopisu je precej visoka – 36-odstotna.

Soča

Soča je objavila več prevodov, vključno z deli *Trije muškетirji* Alexandra Dumasa (*Athos, Porthos, Artagnan, Aramis*), *Grof Monte Cristo* (*Villefort*), *Quo Vadis?* (*Vinicij*) in *Križarski vitezi* (*Zbišek*) Henryka Sienkiewicza ter *Foma Gordejev* Maksima Gorkega. Ključne besede v skupnem obsegu vključujejo 23 odstotkov imen likov. Časopis ima nekaj regionalnih posebnosti, na primer besedo »*nunc*«, ki v goriškem narečju označuje starejšega znanca. Regionalni značaj se odraža tudi v omembah lokalnih političnih osebnosti, kot sta Alojzij Pajer-Monriva, proitalijanski odvetnik in politik, ter Ivan Berbuč, politik in sourednik *Soče*. Zanimiva najdba je ključna beseda »*prismojenec*«. »Prismojenec« je bil vzdevek za *Primorski list*, konservativni časopis, ki je nasprotoval *Soči*, podobno kot je *Slovenski gospodar* nasprotoval *Štajercu.* Kljub temu je bila *Soča* vsebinsko bolj podobna *Slovencu.*[52] Časopis vsebuje 53 odstotkov napak OCR, zaradi česar je eden najtežjih za analizo. Tipična napaka OCR za ta časopis je opuščanje strešice (*uze,*[53] *dezelni, drzaven, goriski*). Poleg tega ima časopis nizko kakovost slik dokumentov, kar še povečuje verjetnost napak OCR.

51 Marijec je napačna oblika leme za besedo Marijca.

52 Branko Marušič, *Pregled politične zgodovine Slovencev na Goriškem: 1848–1899* (Nova Gorica: Goriški muzej, 2005), 326

53 Izvirno uže.

Zgodovinski razvoj jezika

Za dodatno analizo ključnih besed in preučitev razvoja jezika v slovenskih časopisnih publikacijah v poznem 19. in zgodnjem 20. stoletju smo izluščili frekvenčne podatke za izbrane besede iz prejšnjega poglavja. S tem smo želeli točneje opredeliti specifike časopisov z ozirom na razvoj slovenščine. Za primerjavo smo uporabili korpus sPeriodika (ne le deset glavnih časopisov), da bi celovito identificirali trende rabe besed. Izbrane besede so bile prepoznane arhaične besede iz analize ključnih besed: *berž* (brž), *denes* (danes), *sklenica* (steklenica), *menenje* (mnenje), *rekši* (rekoč), *smijati* (smejati), *zanimljiv* (zanimiv), *žnjo/žnjim/žnjimi* (z njo, z njim, z njimi) in *zvršetek* (konec). Čeprav je bilo kandidatov več, smo izbrali tiste pojavnice, ki so imele najvišje število pojavitev v svoji arhaični obliki. Večina frekvenčnih podatkov je bila pridobljena z iskanjem po lemi, razen za *žnjo/žnjim/žnjimi*, kjer je bil uporabljen poizvedbeni jezik CQL.

Frekvence smo pridobili neposredno iz okolja NoSketch Engine. Frekvenčne podatke za časopise, katerih leta izhajanja vključujejo obseg (npr. 1901–1914), smo enakomerno porazdelili med leta, medtem ko smo za tiste, ki vključujejo sezono (npr. 1888/1889), podatke dodelili prvemu navedenemu letu (v tem primeru 1888). Poenostavljena porazdelitev po letih povzroči nekaj netočnosti, vendar je zaradi osredotočenosti na trende in ne na natančne številke taka poenostavitev zadostna.

Rezultati so prikazani na stolpičnem diagramu (Slika 3). Grafe smo začeli z letom 1850, pri čemer smo podatke združili po desetletjih za lažjo primerjavo med grafikoni.

Menjava arhaičnih besed s sodobnimi

Presenetljivo je, da je edina beseda, ki kaže visoko frekvenco na začetku obdobja z nenadnim prevzemom sodobne oblike, *berž* (Slika 3A). *Berž* so najpogosteje uporabljali v časopisu *Kmetijske in rokodelske novice* (2215 pojavitev, relativna gostota[54] 959,6), vendar pri relativni gostoti vodi *Slovenska č(e)bela* (3573,2). Raba besede *berž* je po koncu Bachovega absolutizma, ko so bile dovoljene tudi druge periodične publikacije, postopoma upadala. To je razvidno tudi iz rabe besede v *Kmetijskih in rokodelskih novicah* (Slika 4), kjer raba pada na podoben način.

Regionalne arhaične besede

Beseda *denes* se je uporabljala pretežno v manjših časopisih (*Slovenski tednik*, *Naprej*), medtem ko je bila oblika *danes* v rabi bistveno pogosteje. *Denes* se je uporabljal približno do osemdesetih let 19. stoletja, ko je začel prevladovati sodobni zapis *danes*. Arhaična oblika verjetno izhaja iz kajkavskega jezika, ki je močno vplival na

54 Relativna gostota (*relative density*) primerja pogostost izbranega besedilnega tipa s pogostostjo v celotnem korpusu.

severovzhodni del današnje Slovenije,[55] kjer so izhajale prve izdaje časopisa *Slovenski narod*. *Menenje* je bilo pogosto v regionalnih (južno)zahodnih časopisih (*Gospodarski list, Novičar, Edinost, Slovenka*). Prav tako izrazit regionalni značaj kaže *zvršetek*, z visoko relativno frekvenco v podobnih časopisih. Vendar je splošna frekvenca arhaičnih besed zelo nizka. Te besede so lahko posledica vpliva lokalnega narečja ali italijanskega jezika.

Literarni jezik

Na podlagi primerjave rabe besed sta dva časopisa najbolj odstopala od jezikovne norme tistega časa. To sta literarna časopisa *Ljubljanski zvon* in *Vertec*. V *Ljubljanskem zvonu* so objavljali mnogi znani slovenski avtorji, kot so Anton Aškerc, Simon Gregorčič in Oton Župančič. Podobno so v *Vertcu* objavljali Fran Levstik, Dragotin Kette in Fran Saleški Finžgar. Glede na to, da so časopisa oblikovali pisatelji in pesniki, lahko jezikovna odstopanja pripišemo svobodi literarnega izražanja in eksperimentiranju.

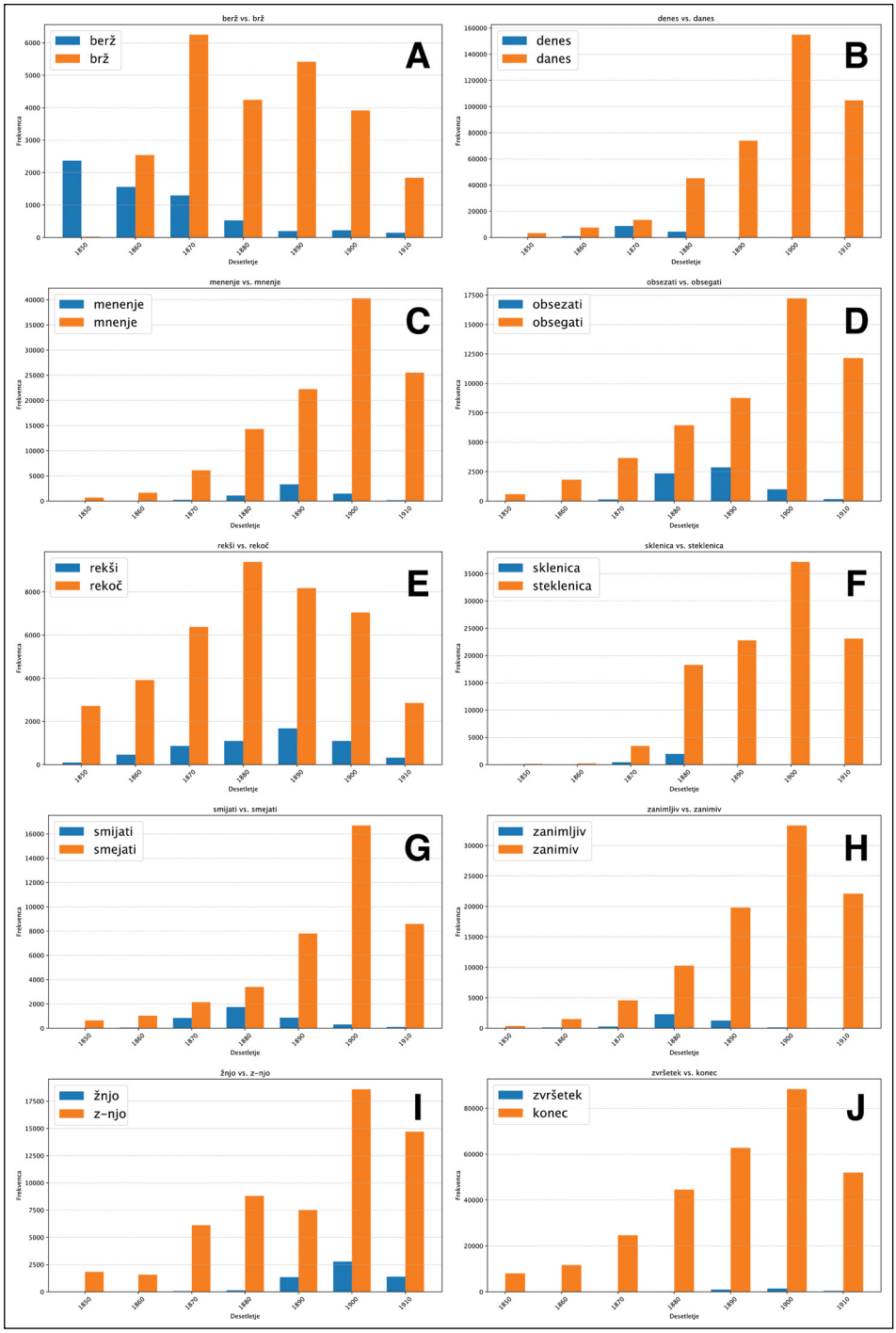
Tabela 2: Prvih 10 ključnih besed (lem) v izbranih časopisih. Celice vsebujejo lemo in njeno frekvenco v določenem časopisu. Zadnja vrstica prikazuje odstotek napak OCR med 100 najpogostejšimi ključnimi besedami.

Razvr - stitev	KRN	SG	UT	SN	DS	SVN	ED	LZ	VT	SČ
1	unidan	sejmov	zavezin	amstetten	nadknez	vravnaven	nvč	gojko	marijca	athos
	-1.552	-843	-2.265	-11.058	-738	-3.299	-12.057	-889	-269	-2040
2	novičar	izgoditi	konvikt	izvoti	virida	gespud	galatti	samorad	otiti	porthos
	-3421	-481	-5.486	-7.416	-798	-3.447	-5.504	-679	-475	-1.411
3	čitavnica	fl	učiteljstvo	pontabel	spominik	tku	barriera	trenk	štir	artagnan
	-2.044	-12.467	-54.905	-6.225	-1.029	-4.680	-7.162	-713	-368	-1.369
4	rusovski	šulverein	učiteljski	selzthal	bodriški	salmov	inseraten	abadon	vrtčev	aramis
	-1.714	-677	-58.083	-8.551	-631	-2.996	-7.641	-549	-220	-1.253
5	kmetovavec	radaj	slomškar	oblig	viljenica	kokr	nuova	zdenka	katarinka	nunec
	-2.481	-541	-1.244	-6.752	-638	-3.535	-7.977	-826	-172	-1.946
6	dnar	vlč	tovarišica	franzensfeste	juriš	napoleondor	konsorcija	groga	ivanek	zbišek
	-2.238	-903	-4.632	-7.256	-912	-3.206	-5.091	-1.046	-181	-1.004
7	žlahen	kosar	koleginja	četiristopen	gotšalk	kursen	pula	cetinovič	pesenca	meljavec
	-1.433	-673	-1.031	-3.690	-610	-2.771	-7.343	-334	-203	-928
8	krajnski	posilinemec	lehrerbund	steyr	maruška	dunavski	vecchia	dramatiški	marijec	villefort
	-3.076	-463	-902	-5.488	-670	-4.189	-6.292	-642	-155	-846
9	žebec	-	pedagoški	osoben	podgoričan	waldsteinov	medpostaja	obsezati	vzpomlad	vinicij
	-632		-2.796	-28.671	-996	-2.349	-3.331	-1.943	-176	-821
10	enmalo	-	črnagoj	vara	ančka	napoleond	piazza	premec	ivanko	foma
	-823		-779	-13.567	-1.407	-2.234	-14.364	-381	-170	-916
napake	**5%**	**92%**	**12%**	**19%**	**1%**	**15%**	**0%**	**36%**	**36%**	**53%**

Vir: NoSketchEngine

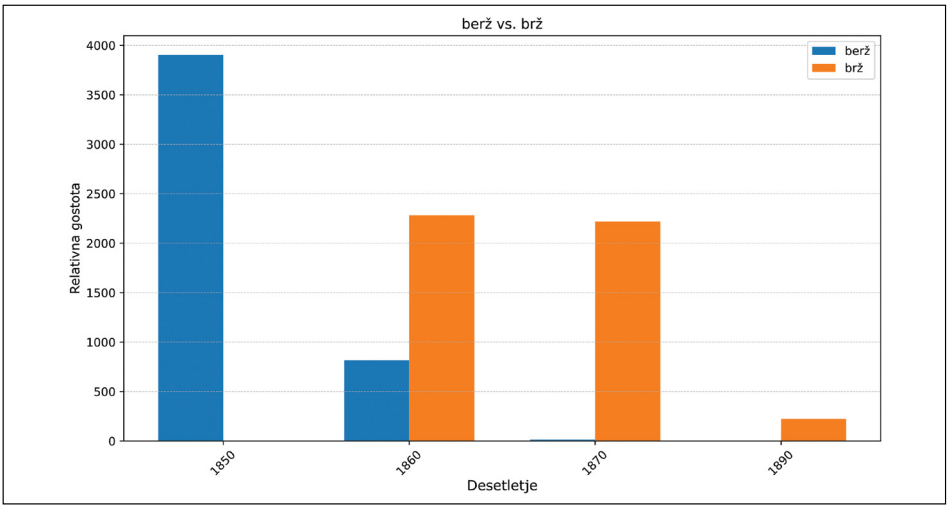
55 Breda Pogorelec, *Zgodovina slovenskega knjižnega jezika* (Ljubljana: Založba ZRC, 2011), 28.

Slika 3: Frekvence arhaičnih besed v primerjavi z njihovimi sodobnimi različicami v korpusu sPeriodika. Modri stolpci predstavljajo arhaične, oranžni stolpci pa sodobne oblike.



Vir: avtorica iz podatkov NoSketchEngine

Slika 4: Relativna gostota besed »berž« in »brž« v *Kmetijskih in rokodelskih novicah* po desetletjih



Vir: avtorica iz podatkov NoSketchEngine

Pojavnico *obsezati* so uporabljali pretežno v časopisu *Ljubljanski zvon*. Zabeležen je relativno velik porast besede v osemdesetih in devetdesetih letih 19. stoletja. Podobno je tudi z besedo *zanimljiv*, ki je imela višjo relativno frekvenco v *Ljubljanskem zvonu* kot v drugih časopisih, ter besedo *smijati*, ki se je pretežno uporabljala v literarnih časopisih v sedemdesetih in osemdesetih letih 19. stoletja (*Vertec* z relativno frekvenco 65,9 in *Ljubljanski zvon* s frekvenco 55,4). V devetdesetih letih 19. stoletja se je trend rabe besede *smijati* začel zmanjševati, prevladovati pa je začela sodobna oblika *smejati*. *Rekši*, arhaična oblika besede *rekoč*, je deležniška oblika glagola reči. Obe obliki sta bili v rabi v opazovanem obdobju, pri čemer je bila *rekši* precej manj priljubljena kot *rekoč*. Raba besede *rekoč* v sodobnem času prav tako upada (vir: metaFida v1.0). Časopisi so uporabljali obe obliki in niso pokazali večjih pristranskosti do *rekši*, razen *Vertca* (40,05) in *Ljubljanskega zvona* (27,64), kjer se oblika *rekši* uporablja nekoliko pogosteje. Nekateri manjši časopisi pa imajo še višjo relativno frekvenco. *Sklenica* je bila uporabljena zgolj občasno v sedemdesetih in osemdesetih letih 19. stoletja, brez kakšnega večjega časopisa, ki bi jo uporabljal v veliki meri. *Žnjim/žnjo/žnjimi* kaže porast na prelomu 20. stoletja, vendar interpretacija za to besedo ni povsem jasna.

Oblikovanje slovenskega jezika je tesno povezano s strokovno razpravo o jezikovnih pravilih. Prva slovenska slovnica je bila Bohoričeva *Arcticae horulae succisivae*, izdana leta 1584. Skoraj dve stoletji je trajalo, da je bila izdana druga slovnica. Leta 1768 je izšla Pohlinova slovnica *Kranjska gramatika*, napisana v nemščini in osredotočena na kranjsko narečje. V začetku 19. stoletja je bilo veliko poskusov modernizacije slovenščine, kar je pripeljalo do izdaj slovnic pomembnih avtorjev, med njimi Kopitarja (1809), Vodnika (1811), Dajnka (1824), Metelka (1825), Murka (1832/43/50), Majarja (1850) in Miklošiča (1852). Kljub številnim konkurenčnim slovnicam je bil

prvi slovenski pravopis objavljen šele leta 1899, avtor pa je bil Fran Levec. Plodovita dejavnost izdajanja slovnic priča o obdobju oblikovanja jezika, v katerem so se soočala nasprotujoča si stališča do pravopisa, izgovarjave, pisanja in skladnje. Ta soočenja so verjetno prispevala k vzporedni rabi določenih arhaičnih in/ali narečnih besed, kar je razvidno iz analize zgodovinskih časopisov.

Analiza napak OCR

Napake OCR so predstavljale pomemben izziv pri analizi določenih časopisov (*Slovenski gospodar*, *Soča*). S pomočjo analize ključnih besed smo ročno identificirali napake OCR iz nabora 100 pojavnic. Pomanjkanje strešic smo obravnavali kot napako OCR, saj se beseda brez strešic šteje kot drugačna od besede s strešicami (*drzaven/ državen*) ali pa lahko pomeni povsem drugo besedo (*čelo/celo*). V širšem pregledu 1000 ključnih besed smo odkrili 266 napak, vključno z manjkajočimi diakritičnimi znaki, zamenjavo znakov in napačno interpretacijo naglasnih znamenj kot številk (npr. *dom6v* namesto *domov*). Pomembno je omeniti, da so bili časopisi digitalizirani z različnimi modeli OCR, kar je povzročilo specifične napake v posameznih publikacijah.

Splošne napake OCR

Nekatere napake OCR se ponavljajo in kažejo na temeljne slabosti modelov OCR za arhaične zapise in slovenščino. Najpogostejša napaka (24 odstotkov) je prepis črk n, s ali š kot a. Te napake so najpogostejše v časopisu *Slovenski gospodar*, ki ima tudi sicer največ napak OCR. Druga najpogostejša napaka (21 odstotkov) je pomanjkanje strešic (*stajerski, drzaven*), tretja (9 odstotkov) pa napačna transkripcija naglasnih znamenj kot številk – zlasti 6, 7 ali 2 (*dom6v, rek6, už6, u2e, pra7*). Naglasna znamenja so pogosto zapisana tudi kot d (*takdj*). Črka n na začetku pogosto pomeni, da se beseda začne z narekovaji (*nkaj, nne, njaz*). Zamenjava črk je zelo pogosta, zlasti med č in e (*oee, užč*), i in l (*ijubi, nefranklran*), c in e (*Marijea, evetice*) ter u in n (*nčenki*).

Naglasna znamenja

Naglasna znamenja in strešice predstavljajo poseben problem pri transkripciji sPeriodike. Tukaj je primer iz *Ljubljanskega zvona*, edinega časopisa, ki redno uporablja naglasna znamenja na samoglasnikih (medtem ko *Vertec* to počne občasno):

Takó kričálo vse je gôri náme. (izvirnik)
Takd kričdlo vse je g6ri ndme. (prepis OCR)

Napake vizualno delujejo smiselno. Ó in á sta prepisana kot d (ali občasno 6), ô kot 6, á tudi kot ä, é pa kot č. Kljub temu težave s transkripcijo omejujejo semantično analizo ključnih besed.

Napake v specifičnih časopisih

Vertec ima specifične napake OCR. Čeprav te niso ekskluzivne za ta časopis, so v njem še posebej izrazite. Znakovne zamenjave pogosto prizadenejo črke in sklope črk m, u in ru. Zaradi podobnosti oblik se m pogosto prepiše kot ra, ni ali in. U se prepiše kot ii, ru pa kot ni ali m. Črka v je pogosto prepisana kot r, ó pa kot d ali 6.

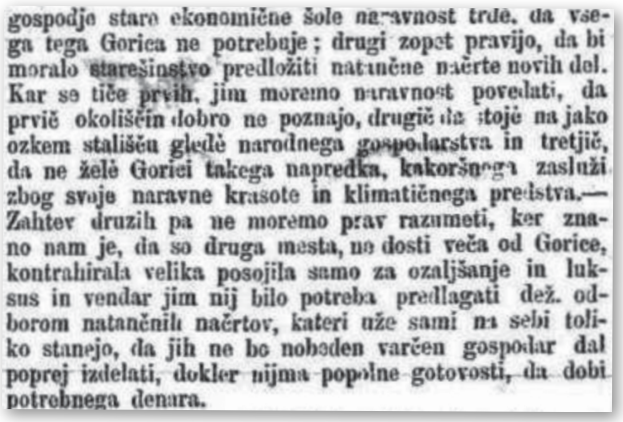
Pri časopisih, kjer se napake pogosto pojavljajo v ključnih besedah, smo primerjali pogostost napačno zapisanih besed s pravilnimi oblikami. Napačna oblika *sloveaski* se v korpusu pojavi 1855-krat, pravilna oblika *slovenski* pa 45.759-krat. Ključnih besed ni mogoče analizirati semantično, saj bi bilo treba vse napačne oblike najprej pretvoriti v pravilne. Vendar se napaka znatno pogosteje pojavlja v *Slovenskem gospodarju* kot v katerem koli drugem časopisu. Razlika v pogostosti pomeni, da ta napaka značilno označuje časopis in bi jo bilo mogoče uporabiti pri postopkih naknadne obdelave. Z drugimi besedami, takšne napačne oblike bi lahko naknadno popravili v izbrani publikaciji.

Kandidati za ponovno optično branje

S stopnjo napak lahko določimo tudi kandidate za ponovno optično branje. Nekateri optično prebrani dokumenti so že zdaj slabe kakovosti ali pa so bili med prvimi digitaliziranimi časopisi. Sodobne rešitve OCR bi lahko dale precej boljši rezultat od obstoječih različic, vendar je ponovno optično branje celotne sPeriodike zamudno in nepotrebno. Smiselna rešitev bi bilo oblikovanje seznama kandidatov za ponovno optično branje. Na podlagi naših rezultatov bi *Slovenski gospodar* in *Soča* pridobila tako s ponovnim optičnim branjem kot tudi s sodobno OCR-transkripcijo, medtem ko bi *Ljubljanski zvon* potreboval le izboljšano transkripcijo (saj so optično prebrani dokumenti že ustrezni).

Sodobne tehnologije OCR, skupaj z velikimi jezikovnimi modeli (VJM) in velikimi multimodalnimi modeli (VMM), odpirajo nove možnosti za izboljšanje natančnosti transkripcije. Na primer, GPT-4o je uspešno transkribiral slabše optično prebrani del časopisa *Soča* (Slika 5):

Slika 5: Del časopisa *Soča* s slabo kakovostjo optičnega branja



gospodje staro ekonomične šole naravnost trde, da vsega tega Gorica ne potrebuje; drugi zopet pravijo, da bi moralo starešinstvo predložiti natančne načrte novih del. Kar se tiče prvih, jim moremo naravnost povedati, da prvič okoliščin dobro ne poznajo, drugič da stojé na jako ozkem stališču glede narodnega gospodarstva in tretjič, da ne želé Gorici takega napredka, kakoršnega zasluži zbog svoje naravne krasote in klimatičnega prečistva.— Zahtev druzih pa ne moremo prav razumeti, ker znano nam je, da so druga mesta, no dosti veča od Gorice, kontrahirala velika posojila samo za ozaljšanje in luksus in vendar jim nij bilo potreba predlagati dež. odborom natančnih načrtov, kateri uže sami na sebi toliko stanejo, da jih ne bo nobeden varčen gospodar dal poprej izdelati, dokler nijma popolne gotovosti, da dobi potrebnega denara.

Vir: *Soča*, 17. 9. 1874, https://dlib.si/details/URN:NBN:SI:DOC-5JQMY60Z/

> gospodo staro ekonomične šole nezavnost trde, da vsega tega Gorica ne potrebuje; drugi zopet pravijo, da bi moralo starešinstvo predložiti natčene načrte novih del. Kar se tiče prvih, jim moramo naravnost povedati, da prvič okolišin dobro ne poznajo, drugič da stojé na jako ozkem stališču glede narodnega gospodarstva in tretjič, da ne želé Gorici takega napredka, kakoršnega zasluži zaradi svoje naravne krasote in klimatičnega prečistva. Zahtev drugih pa ne moremo prav razumeti, kar znano nam je, da so druga mesta, no dosti veča od Gorice, kontrahirala velika posojila samo za ozaljšanje in luksus in vendar jim ni bilo potrebno predlagati dež. odboru natancnih načrtov, kateri že sami na sebi toliko stanjo, da jih ne bo nobeden varčen gospodar dal poprej izdelati, dokler njim popolne gotovosti, da dobi potrebnega denarja.

Zmožnosti VJM in VMM omogočajo prepoznavo slabše optično prebranih dokumentov skoraj brez dodatnega prilagajanja. VMM presegajo sodobne rešitve OCR pri neposredni prepoznavi besedila,[56] tudi pri kompleksnih postavitvah, kot so izrezki iz starih kitajskih časopisov[57] in rokopisna besedila.[58] Medtem ko trenutne raziskave kažejo mešane rezultate za popravke po optični prepoznavi znakov,[59, 60]

56 Yuliang Liu et al., »On the Hidden Mystery of OCR in Large Multimodal Models,« *Sci. China Inf. Sci.* 67, 220102 (2024), https://doi.org/10.1007/s11432-024-4235-6.

57 Eric H. C. Chow, »An Experiment with Gemini Pro LLM for Chinese OCR and Metadata Extraction,« pridobljeno 5. 4. 2024, https://digitalorientalist.com/2024/04/05/an-experiment-with-gemini-pro-llm-for-chinese-ocr-and-metadata-extraction/.

58 Mark Humphries et al., »Unlocking the Archives: Large Language Models Achieve State-of-the-Art Performance on the Transcription of Handwritten Historical Documents,« pridobljeno 24. 10. 2024, http://dx.doi.org/10.2139/ssrn.5006071.

59 Alan Thomas, Robert Gaizauskas in Haiping Lu, »Leveraging LLMs for post-OCR correction of historical newspapers,« v: *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages* (Torino: ELRA in ICCL, 2024), 116–21.

60 Emanuela Boros et al., »Post-correction of historical text transcripts with large language models: An exploratory study,« v: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (LaTeCH-CLfL 2024) (St. Julians: Association for Computational Linguistics, 2024).

bi prilagoditev VMM za zgodovinske podatke lahko izboljšala rezultate. Ti napredki odpirajo pot za globlje analize zgodovinskih korpusov,[61] vključno s povzetki vsebine, analizo trendov in semantičnim iskanjem. Poleg tega nastajajo novi VJM, posebej prilagojeni zgodovinskim podatkom (ZVJM), ki omogočajo še podrobnejši vpogled v zgodovinske družbe.[62]

Razprava

Časopise smo opredelili z analizo ključnih besed na podlagi lem. Periodike so običajno opredeljene bodisi s svojo deklarirano usmeritvijo (npr. *KRN*, *Učiteljski tovariš*) bodisi s podlistki in oglasi (npr. *Dom in svet*, *Slovenski narod*) ali pa – žal – z napakami OCR (*Slovenski gospodar*).

Poudarek na podlistkih in oglasih se ujema s predhodnimi raziskavami o zgodovinskih slovenskih periodikah. Podlistki, tj. časopisni odseki, namenjeni leposlovju, so igrali ključno vlogo pri razvoju slovenske proze, saj so avtorjem omogočali zgodnji dostop do širšega občinstva.[63] Analiza ključnih besed je sicer identificirala zgolj značilne izraze, ki sovpadajo z določenimi literarnimi liki, vendar so ti neločljivo povezani s podlistki, v katerih se pojavljajo.

Nasprotno pa je bila vloga oglasov bolje poudarjena. V poznem 19. stoletju so oglasi zavzemali pomemben del periodik, pri čemer je bilo razmerje med uredniškimi in oglasnimi vsebinami 4 : 1.[64] Ključne besede so izpostavile specifične oglaševalce in tudi splošni oglaševalski jezik (npr. *inseraten, nvč*).

Analiza ključnih besed je razkrila prehodno stanje slovenskega jezika v tem obdobju. Vsaka periodika je imela svojevrstne pravopisne konvencije za knjižne besede. Na primer, *Kmetijske in rokodelske novice* uporabljajo *nograd* namesto *vinograd* in *berž* namesto *brž*, medtem ko *Slovenski narod* uporablja *denes* namesto *danes* in *sklenica* namesto *steklenica*. Celo časopisi, ki so bili v ospredju jezikovne standardizacije, denimo *Ljubljanski zvon*, vsebujejo besede, ki so danes arhaične (npr. *obsezati* namesto *obsegati* in *smijati* namesto *smejati*). Diahrona analiza je pokazala, da so bile nekatere besede specifične za določene regije, druge pa so odražale eksperimentalno ali umetniško rabo v vodilnih literarnih periodikah.

Napake OCR so predstavljale pomemben izziv pri analizi določenih periodik. Pri periodikah s pogostimi napakami OCR lahko napačne besede izkrivljajo analizo, zato bi popravljanje besedila po optični prepoznavi znakov izboljšalo natančnost

61 Giselle G. Garcia in Christian Weilbach, »If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History,« v: *Proceedings of the Computational Humanities Research Conference 2023* (Pariz: CHR, 2023), 616–38

62 Michael E. W. Varnum, Nicolas Baumard, Mohammad Atari in Kurt Gray, »Large language models based on historical text could offer informative tools for behavioral science,« *Proceedings of the National Academy of Sciences* 121, št. 42 (Washington, DC: National Academy of Sciences of the United Stated of America 2024): e2407639121, https://doi.org/10.1073/pnas.2407639121.

63 Dović, »Literatura in mediji v Jurčičevem času.«

64 Ibidem.

semantične analize. Kot poudarjajo Strange in sodelavci,[65] je popravljanje po OCR ključno za tehnike, kot je analiza ključnih besed. Nekatere periodike so bile zaradi ponavljajoče se vsebine podvržene pristranskostim v analizi ključnih besed. V *Slovencu*, na primer, je 29 odstotkov ključnih besed pripadalo glavi časopisa, medtem ko se je 68 odstotkov ključnih besed v *Edinosti* nanašalo na italijanska ulična imena. Takšne pristranskosti omejujejo uporabnost analize ključnih besed za vsebinsko karakterizacijo v teh primerih.

Zaključek

Analiza ključnih besed razkriva različne vidike periodik. Nekateri časopisi so opredeljeni s svojo splošno vsebino, kot je kmetijstvo (*Kmetijske in rokodelske novice*) ali pedagogika (*Učiteljski tovariš*); drugi so opredeljeni s ponavljajočimi se podlistki, ki jih objavljajo (*Dom in svet*, *Slovenec*, *Vertec*, *Soča*); nekateri pa so prepoznavni po oglasnem prostoru (*Slovenski narod*, *Edinost*). *Slovenski gospodar* žal vsebuje preveč napak OCR, da bi analiza ključnih besed razkrila smiselne vpoglede. Ponavljajoče se napake OCR v periodikah bi lahko bile obravnavane v postopku obdelave po optični prepoznavi znakov.

Računalniški pregled ponuja številne možnosti za nadaljnje analize. Mogoče bi bilo, denimo, primerjalno analizirati prva dva slovenska dnevna časopisa, liberalni *Slovenski narod* in konservativnega *Slovenca*. Podobna primerjalna analiza bi se lahko uporabila za *Edinost* in *Sočo*, dva časopisa Slovencev v Italiji, ter razčlenitev njunih skupnih in različnih elementov (še posebej ob upoštevanju namenov za združitev teh časopisov). Kandidatne arhaične besede bi lahko izbrali s frekvenčnega seznama celotne sPeriodike in tako točneje opredelili razvoj slovenščine na prelomu 19. v 20. stoletje. Veliko zahtevnejša raziskava bi lahko preučila razlike v oglasih, saj so ti izstopali že pri analizi ključnih besed. Naloga je kompleksna, ker je izredno težko določiti meje posameznih oglasov, vendar bi bilo problem mogoče obravnavati tako, da bi periodike obravnavali kot slike[66] in uporabili iskanje sosedov za določanje podobnih oglasov. Velike multimodalne modele lahko uporabljamo za mnoge zgoraj omenjene naloge, ta tehnologija pa bo prihodnosti korenito spremenila zgodovinske raziskave, še posebej pri obravnavi korpusov nižje kakovosti.

65 Carolyn Strange, Daniel McNamara, Josh Wodak in Ian Wood, »Mining for the meanings of a murder: The impact of OCR quality on the use of digitized historical newspapers,« *Digital Humanities Quarterly* 8, št. 1 (2014).

66 Quintus van Galen, »The page is an image again: Bleedmapping as an analysis technique for historical newspapers,« *Digital Humanities Quarterly* 17, št. 1 (2023).

Zahvale

Iskreno se zahvaljujem dr. Nikoli Ljubešiću in Filipu Dobraniću za njun neprecenljiv prispevek k pričujočemu delu. Delo, opisano v tem članku, sta financirali Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije v okviru raziskovalnega programa P6-0436 *Digitalna humanistika: viri, orodja in metode* (2022–2027) ter raziskovalne infrastrukture DARIAH-SI in Evropska unija v okviru programa Horizon Europe (101186647 – AI4DH).

Viri in literatura

Literatura

Amon, Smilja. »Vloga slovenskega časopisja v združevanju in ločevanju slovenske javnosti od 1797–1945.« *Javnost* 15 (2008): S9–S24.

Anonymous, L.. »Slovenski časopisi leta 1885.« *Ljubljanski zvon* 5, 1885, 631–35.

Boros, Emanuela, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer in Frédéric Kaplan. »Post-correction of historical text transcripts with large language models: An exploratory study.« V: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, uredili Yuri Bizzoni, Stefania Degaetano-Ortlieb, Anna Kazantseva in Stan Szpakowicz. St. Julians: Association for Computational Linguistics, 2024.

Darovec, Darko. *Pregled zgodovine Istre.* Koper: Zgodovinsko društvo za južno Primorsko, Založba Annales; Čentur: Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja, 2023.

Dobranić, Filip, Bojan Evkoski in Nikola Ljubešić. »Corpus of Slovenian periodicals (1771–1914) sPeriodika 1.0« (2023). *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1881.

Dobranić, Filip, Bojan Evkoski in Nikola Ljubešić. »A Lightweight Approach to a Giga-Corpus of Historical Periodicals: The Story of a Slovenian Historical Newspaper Collection.« V: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, uredili Nicoletta Calzolari, Kan Min-Yen, Veronique Hoste et al. Torino: ELRA in ICCL, 2024.

Dović, Marijan. »Literatura in mediji v Jurčičevem času.« *Slavistična revija* 54, št. 4 (2006): 543–57.

Dović, Marijan. »Anatomy of the 'Deathly Silence': Slovenian Newspapers in Carniola and the Pre-March Censorship.« *Neohelicon* 50, št. 2 (2023): 543–60. https://doi.org/10.1007/s11059-023-00707-8.

Ehrmann, Maud, Estelle Bunout in Marten Düring. »Historical Newspaper User Interfaces: A Review.« V: *85th IFLA General Conference and Assembly (IFLA)*. Zenodo, 2019.

Ehrmann, Maud, Marten Düring, M., Clemens Neudecker in Antoine Doucet. »Computational Approaches to Digitised Historical Newspapers.« *Dagstuhl Reports* 12, št. 7 (2023): 112–79. Pridobljeno 5. 2. 2025. https://doi.org/10.4230/DagRep.12.7.112.

Garcia, Giselle G. in Christian Weilbach. »If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History.« V: *Proceedings of the Computational Humanities Research Conference 2023*, uredili Artjoms Šeļa, Fotis Jannidis in Iza Romanowska, 616–38. Pariz: CHR, 2023.

Hengchen, Simon, Ruben Ros, Jani Marjanen in Mikko Tolonen. »A Data-Driven Approach to Studying Changing Vocabularies in Historical Newspaper Collections.« *Digital Scholarship in the Humanities* 36, dodatek 2 (2021): ii109-ii126. https://doi.org/10.1093/llc/fqab032.

Humphries, Mark, Lianne C. Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray in Spence, Elizabeth. »Unlocking the Archives: Large Language Models Achieve State-of-the-Art Performance on the Transcription of Handwritten Historical Documents.« Pridobljeno 24. 10. 2024. http://dx.doi.org/10.2139/ssrn.5006071.

Ilich, Maja. »Nekaj o modi v slovenskem časopisju na prelomu stoletja (1895-1915).« *Zgodovina za vse* 6, št. 2 (1999): 98–108.

Jezernik, Božidar. »Katoliška duhovščina na prelomu devetnajstega in dvajsetega stoletja in proces modernizacije na Slovenskem.« *Traditiones* 51, št. 1 (2022): 103–45.

Kermavner, Dušan. »Drugi slovenski socialnodemokratski listi.« *Kronika* 10 (1962): 80–89.

Kettunen, Kimmo in Tuula Pääkkönen. »Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means.« V: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, uredili Khalid Choukri, Thierry Declerck, Sara Goggi et al., 956–61. Portorož: ELRA, 2016.

Kilgarriff, Adam. »Simple Maths for Keywords.« V: *Proceedings of Corpus Linguistics* 6. Liverpool, VB: University of Liverpool, 2009.

Liu, Yuliang, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Yin, Xucheng, Cheng-lin Liu, Lianwen Jin in Xiang Bai. »On the Hidden Mystery of OCR in Large Multimodal Models.« *Sci. China Inf. Sci.* 67, 220102 (2024). https://doi.org/10.1007/s11432-024-4235-6.

Marjanen, Jani, Jussi Kurunmäki, Lidia Pivovarova in Elaine Zosa. »The Expansion of Isms, 1820–1917: Data-Driven Analysis of Political Language in Digitized Newspaper Collections.« *Journal of Data Mining & Digital Humanities* 2020. https://doi.org/10.46298/jdmdh.6159.

Marjanen, Jani, Ville Vaara, Antti Kanner, Hege Roivainen, Eetu Mäkelä, Leo Lahti in Mikko Tolonen. »A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917.« *Journal of European Periodical Studies* 4, št. 1 (2019). https://doi.org/10.21825/jeps.v4i1.10483.

Marjanen, Jani, Elaine Zosa, Simon Hengchen, Lidia Pivovarova in Mikko Tolonen. »Topic Modelling Discourse Dynamics in Historical Newspapers.« V: *Digital Humanities in the Nordic Countries 2020*, uredili Sanita Reinsone, Inguna Skadiņa, Andra Baklāne in Jānis Daugavietis, 63–77. CEUR-WS.org, 2021.

Marušič, Branko. *Pregled politične zgodovine Slovencev na Goriškem: 1848-1899*. Nova Gorica: Goriški muzej, 2005.

Marušič, Branko. »Izbor vesti o Istri v slovenskem časopisju do leta 1880.« *Annales* 17, št. 1 (2007): 65-82.

Mayer, Adán. I. L., Ximena Gutierrez-Vasques, Ernesto P. Saiso in Hannu Salmi. »Underlying sentiments in 1867: A study of news flows on the execution of Emperor Maximilian I of Mexico in digitized newspaper corpora.« *Digital Humanities Quarterly* 16, št. 4 (2022).

Mihelič, Stane. »Kmetijska družba in ustanovitev 'Novic'.« *Slavistična revija* 1, št. 1/2 (1948).

Park, Jaihyun in Ryan Cordell. »A quantitative discourse analysis of Asian workers in the US historical newspapers.« V: *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, uredili Mika Hämäläinen, Emily Öhman, Flammie Pirinenm et al., 7–15. Tokio: Association for Computational Linguistics, 2023.

Pedrazzini, Nilo in Barbara McGillivray. »Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers.« V: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, uredili Mika Hämäläinen, Khalid Alnajjar, Niko Partanen in Jack Rueter, 85–95. Tajpej: Association for Computational Linguistics, 2022.

Pivovarova, Lidia, Elaine Zosa in Jani Marjanen. »Word Clustering for Historical Newspapers Analysis.« V: *Proceedings of the Workshop on Language Technology for Digital Historical Archives*, uredili Cristina Vertan, Petya Osenova in Dimitar Iliev, 3–10. Varna, Bulgarija: INCOMA Ltd., 2019.

Pogorelec, Breda. *Zgodovina slovenskega knjižnega jezika*, uredil Kozma Ahačič. Založba ZRC, 2011.

Pretnar Žagar, Ajda. »A corpus linguistic characterization of speriodika.« V: *Proceedings of the Conference on Language Technologies and Digital Humanities,* uredila Špela Arhar Holdt in Tomaž Erjavec, 384–406. Ljubljana: Inštitut za novejšo zgodovino, 2024.

Schoots, Jonathan. »Analyzing political formation through historical isiXhosa text analysis: Using frequency analysis to examine emerging African nationalism in South Africa.« V: *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023)*, uredili Rooweither Mabuya, Don Mthobela, Mmasibidi Setaka in Menno Van Zaanen, 65–75. Dubrovnik, Hrvaška: Association for Computational Linguistics, 2023. https://doi.org/10.18653/v1/2023.rail-1.8.

Stergar, Nataša. »Narodnostno vprašanje v predmarčnih letnikih Bleiweisovih Novic.« *Kronika* 25, št. 3 (1977).

Strange, Carolyn, Daniel McNamara, Josh Wodak in Ian Wood. »Mining for the meanings of a murder: The impact of OCR quality on the use of digitized historical newspapers.« *Digital Humanities Quarterly* 8, št. 1 (2014).

Štepec, Marko. »Zločin v slovenskem časopisju v 80. letih 19. stoletja.« *Kronika* 35, št. 1/2 (1987): 30–38.

Thomas, Alan, Robert Gaizauskas in Haiping Lu. »Leveraging LLMs for post-OCR correction of historical newspapers.« V: *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages*, uredila Rachele Sprugnoli in Marco Passarotti, 116–21. Torino: ELRA in ICCL, 2024.

van Galen, Quintus. »The page is an image again: Bleedmapping as an analysis technique for historical newspapers.« *Digital Humanities Quarterly* 17, št. 1 (2023).

Varnum, Michael E. W., Nicolas Baumard, Mohammad Atari in Kurt Gray. »Large language models based on historical text could offer informative tools for behavioral science.« *Proceedings of the National Academy of Sciences* 121, št. 42. Washington, DC: National Academy of Sciences of the United Stated of America 2024: e2407639121. https://doi.org/10.1073/pnas.2407639121.

Verheul, Japp, Hannu Salmi, Martin Riedl, Asko Nivala, Lorella Viola, Jana Keck in Emily Bell. »Using word vector models to trace conceptual change over time and space in historical newspapers 1840–1914.« *Digital Humanities Quarterly* 16, št. 2, (2022).

Zorn, Tone. »Odmevnost jezikovnega vprašanja v listu Slovenski pravnik v letih 1871-1918.« *Kronika* 35, št. 3 (1987): 146–55.

Spletni viri

Chow, Eric H. C. »An Experiment with Gemini Pro LLM for Chinese OCR and Metadata Extraction.« Pridobljeno 5. 4. 2024. https://digitalorientalist.com/2024/04/05/an-experiment-with-gemini-pro-llm-for-chinese-ocr-and-metadata-extraction/.

Ajda Pretnar Žagar

COMPUTATIONAL ANALYSIS OF SLOVENIAN HISTORICAL NEWSPAPERS (1771–1914): LINGUISTIC, THEMATIC, AND NATION-BUILDING INSIGHTS

SUMMARY

This paper presents a corpus linguistic study of *sPeriodika*, a recently published corpus of Slovenian historical periodicals (1771–1914), compiled from digitised newspapers in the digital repository of the Slovenian National and University Library (dLib.si). The corpus includes key periodicals that contributed to literacy and nation-building in Slovenia. The study focuses on the ten newspapers with the highest number of publications. The author uses keyword analysis, word frequency analysis, concordances, and diachronic analysis to characterise their content and the historical development of the Slovenian language. The study identifies specific thematic orientations of selected periodicals, such as agriculture, pedagogy, feuilletons and advertising, by extracting and analysing keywords. It links the findings to the intense nation-building that followed the March Revolution of 1848.

To characterise the development of the Slovenian language, the author uses diachronic analysis, comparing archaic and modern word forms identified by keyword analysis. Our results indicate that Slovenian literary and regional periodicals exhibited distinct linguistic conventions.

The author uses diachronic analysis to characterise the development of the Slovenian language, comparing archaic and modern word forms identified through keyword analysis. The results show that Slovenian literary and regional periodicals had a distinct set of linguistic conventions, which reflected broader trends in language standardisation.

The study also addresses the challenges posed by the poor quality of optical character recognition (OCR) in historical documents. OCR errors are a significant challenge in historical newspaper analysis. Our research identifies recurring OCR problems, including the misrecognition of characters and the omission of diacritics. Some newspapers, such as *Slovenski gospodar* and *Soča*, have exceptionally high OCR error rates, affecting the keyword analysis results. We discuss possible solutions, including post-OCR correction and using modern Large Multimodal Models (LMMs) and Large Language Models (LLMs) to improve OCR accuracy. Preliminary experiments with GPT-4o, a well-known LLM, show promising results in transcribing degraded historical texts. Future research could focus on refining OCR correction techniques and extending comparative analyses across historical newspapers.

In conclusion, this study highlights the value of computational methods in historical newspaper research despite the challenges of OCR. Keyword analysis effectively

differentiates newspapers based on content, thematic focus, and editorial stance. However, OCR errors need to be taken into account in future studies. Our findings suggest the potential of machine learning and AI-based OCR improvements for processing historical newspapers, paving the way for more refined analyses of historical corpora in Slovenian and other languages.

1.01 DOI: https://doi.org/10.51663/pnz.65.3.03

Diana Košir,* Tomaž Erjavec**

Korpusna analiza pripovednega sloga in jezikovne norme v starejši verski periodiki

IZVLEČEK

V prispevku je predstavljen postopek izdelave korpusa CVET, ki vsebuje besedila patra Hijacinta Repiča, objavljena v verski reviji Cvetje z vertov sv. Frančiška v obdobju 1881–1916. Korpus je uporabljen kot podlaga za jezikovno in stilistično analizo, opravljeno z orodjem noSketch Engine. Z analizo frekvenčnosti izbranih spremenljivk sta opisana besedišče patra Repiča in njegov pripovedni slog. Nazadnje je na primeru besed tipa bralec/bravec opazovan sinhrono-diahroni in normativni vidik starejšega slovenskega jezika besedil v korpusu.

Ključne besede: starejši slovenski jezik, verski tisk, korpusno jezikoslovje, korpusna stilistika, jezikovna norma

ABSTRACT

CORPUS ANALYSIS OF NARRATIVE STYLE AND LINGUISTIC NORM IN AN OLDER RELIGIOUS PERIODICAL

The article presents the process of creating the CVET corpus, which contains the texts of Father Hijacint Repič published in the religious journal Cvetje z vertov sv. Frančiška between 1881 and 1916. The corpus then serves as the basis for a linguistic and stylistic analysis, carried out with the noSketch Engine tool. Frequency analysis of selected variables is used to describe

* **Asistentka, Inštitut za jezikoslovne študije, Znanstveno-raziskovalno središče Koper, Garibaldijeva 1, SI-6000 Koper, diana.kosir@zrs-kp.si; ORCID: 0009-0009-4428-9698**

** **Dr., strokovno-raziskovalni svetnik, Institut »Jožef Stefan«, Odsek za tehnologije znanja, Jamova 39, SI-1000 Ljubljana, tomaz.erjavec@ijs.si; ORCID: 0000-0002-1560-4099**

Father Repič's vocabulary and narrative style. Finally, the synchronic-diachronic and normative aspects of the older Slovenian language in the texts of the corpus are considered using the example of words of the type bralec/bravec.

Keywords: older Slovenian language, religious press, corpus linguistics, corpus stylistics, linguistic norm

Uvod

Razvoj slovenskega knjižnega jezika do konca 19. stoletja je mogoče opazovati skozi tri pomembna obdobja, začenši z dobo utemeljevanja kranjske knjižne norme od Bohoriča do Kopitarja (1584–1808), ki sta ji sledila »puristično« obdobje s Kopitarjem, Miklošičem in Janežičem (1808–1854–1863) s pretenzijami po skupnem knjižnem jeziku in ukinjanju drugih deželnih knjižnih različic ter obdobje razvoja poklicnega slovanskega in slovenskega jezikoslovja (začetek Orožen označi z Miklošičevim nastopom profesure na dunajski univerzi v letih 1850–1880).[1]

Od sredine 19. stoletja je slovenski prostor zaznamovalo obdobje narodne prebuje s čitalniškim in taborskim gibanjem, vzponom prosvetno-kulturnih in gospodarskih društvenih organizacij ter porastom slovenskega knjižnega in periodičnega tiska.[2] Zlasti na obrobju slovenskega kulturnega prostora je imel tisk zaradi jezikovnega elementa tudi narodno identifikacijsko in narodnopovezovalno vlogo.[3] Slovenski jezik je tako za Slovence pomenil enega od ključnih gradnikov osebne in nacionalne identitete ter narodne zavesti.[4]

V obdobju do prve svetovne vojne se je ob slovenskem gospodarskopolitičnem in drugem periodičnem časopisju, denimo *Kmetijskih in rokodelskih novicah* (Ljubljana), *Slovenskem gospodarju* (Maribor), *Slovenskem narodu* (Maribor), *Slovencu* (Ljubljana), *Soči* (Gorica), *Gorici* (Gorica), *Primorskem listu* (Gorica), *Edinosti* (Trst), vzpostavila mreža literarnih revij, na primer *Slovenska bčela* (Celovec), *Slovenski glasnik* (Celovec), *Glasnik* (Maribor), *Zvon* (Dunaj), *Ljubljanski zvon* (Ljubljana), *Kres* (Celovec), *Dom in svet* (Ljubljana), *Slovenka* (Trst), strokovnih in stanovskih glasil, na primer *Časopis za zgodovino in narodopisje* (Maribor), *Učiteljski tovariš* (Ljubljana), ter verske periodike, denimo *Cerkveni glasbenik* (Ljubljana), *Zgodnja Danica* (Ljubljana), *Krščanski detoljub* (Ljubljana), *Angelček* (Ljubljana), *Družinski prijatelj* (Trst), *Svetilnik* (Trst), *Rimski katolik* (Gorica), *Drobtinice* (Gradec, Maribor, Ljubljana) in *Cvetje z vertov*

1 Martina Orožen, *Oblikovanje enotnega slovenskega knjižnega jezika v 19. stoletju* (Ljubljana: Filozofska fakulteta, Znanstveni inštitut Filozofske fakultete, 1996), 13–22.

2 Urška Perenič, *Empirično-sistemsko raziskovanje literature: Konceptualne podlage, teoretski modeli in uporabni primeri* (Ljubljana: Slavistično društvo Slovenije, 2010), 101–13, 174–80. Marijan Dović, *Slovenski pisatelj: Razvoj vloge literarnega proizvajalca v slovenskem literarnem sistemu* (Ljubljana: Založba ZRC, ZRC SAZU, 2007), 119-27.

3 Urška Perenič, »Čitalništvo v perspektivi družbenogeografskih dejavnikov.« *Slavistična revija* 60, št. 3 (2012): 365–82.

4 Vesna Mikolič, »Povezanost narodne in jezikovne zavesti,« *Jezik in slovstvo* 45, št. 5 (2000): 180–82.

sv. Frančiška (Gorica, Kamnik).[5] Uredniki (na primer Bleiweis, Einspieler, Janežič, Jurčič, Stritar, Levec, Sket, Slomšek, Škrabec idr.) so pogosto tudi lektorsko posegali v besedila skladno s svojimi nazori, zato lahko časopisje predstavlja relevanten vir za opazovanje razvoja slovenskega knjižnega jezika in postopnega sprejemanja t. i. »novih oblik«.

Jezikoslovec p. Stanislav Škrabec je svoje jezikoslovne razprave objavljal na platnicah revije *Cvetje z vertov sv. Frančiška* (v nadaljevanju *CFr*), uredniško pa je bdel nad jezikovno podobo besedil znotraj revije. O tem je p. Kunstelj zapisal: »Kar je on učil, je predvsem sam izvrševal. Zato je pa presedel dostikrat skoro po cele noči, da je prelil in predelal spise drugih po svojih zahtevah, zlasti pa, da je svoje lastne, dostikrat precej obširne sestavke, svojim nazorom primerno priredil.«[6]

Opis raziskovalnega gradiva in metod

Osrednje gradivo za raziskavo predstavlja manjši specializirani korpus starejšega slovenskega jezika CVET 1.0,[7] ki vsebuje vse objave p. Hijacinta Repiča v reviji *CFr* v letih 1881–1916 (ur. Škrabec). Pričujoči članek je nastal na osnovi prispevka avtorjev na konferenci Jezikovne tehnologije in digitalna humanistika (2024),[8] pri čemer gre za razširitev stilistične analize z dodanimi spremenljivkami za opis pripovednega sloga patra Repiča, kot jo ponuja korpusni pristop z uporabo orodij noSketch Engine in Sketch Engine. Dodan je tudi primer korpusne analize zgodovinskih leksikalnih oblik kot odraz udejanjanja Škrabčeve knjižne norme v praksi, ob primerjavi z drugimi referenčnimi jezikovnimi priročniki.

Izdelava in opis korpusa CVET 1.0

Priprava gradiva za korpus je zajemala izbor besedil (pridobljenih s strani dLib), pretvorbo v besedilne datoteke Word in kritični prepis (podrobnejši opis priprave gradiva za korpus prinaša konferenčni prispevek).[9] Vsaka datoteka Word je bila označena z metapodatki, zbranimi v preglednici Excel (identifikator besedila – ime datoteke, avtor, naslov članka, mesto objave (leto, letnik, številka), URL izvornega mesta objave v dLib in številke strani, na katerih se članek pojavi).

5 Miša Šalamun, *Slovensko primorsko časopisje. Zgodovinski pregled in bibliografski opis* (Koper: Lipa, 1961), 25, 26. Dović, *Slovenski pisatelj*, 120–23.

6 Bruno Korošak, *P. Stanislav Škrabec, frančiškan, v očeh sodobnikov* (Ljubljana: Založba Brat Frančišek, 2001), 19.

7 Diana Košir in Tomaž Erjavec, »Corpus of texts by Hijacint Repič CVET 1.0« (2024), *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1226.

8 Diana Košir in Tomaž Erjavec, »Korpus CVET 1.0: izdelava, opis in analiza zbirke starejših besedil v verski periodiki,« prispevek objavljen na letni konferenci JT-DH 2024, *Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenija, 19.–20. 9. 2024, https://doi.org/10.5281/zenodo.13936400.

9 Ibidem.

Nato je bilo gradivo iz formatov Word in Excel pretvorjeno v XML-shemo, skladno s priporočili za kodiranje besedil TEI.[10] Datoteke Word so bile pretvorjene v osnovni TEI z uporabo standardnih skript XSLT za pretvorbo v dokumente TEI in iz njih,[11] razpredelnica Excel z metapodatki o posameznih besedilih pa je bila shranjena kot datoteka TSV. Z namensko skripto so bile besedilne datoteke TEI povezane z metapodatki v eno datoteko TEI (vrhnji element <TEI>), ki vsebuje kolofon (element <teiHeader>) in besedilo (<text>). Začetek ene od teh datotek ilustrira Slika 1.

Slika 1: Primer začetka zapisa posameznega besedila v formatu TEI

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="CVET-1887_7_2_43-46" xml:lang="sl">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Vzroki in koristi terpljenja pobožnih duš. (1887) [CVET]</title>
        <author>Repič, Hijacint</author>
      </titleStmt>
      <editionStmt>
        <edition>1.0</edition>
      </editionStmt>
      <extent>
        <measure unit="words">1203</measure>
      </extent>
      <publicationStmt>
        <distributor>CLARIN.SI</distributor>
        <idno type="handle">http://hdl.handle.net/11356/2026</idno>
        <date when="2024-05-07"> 7. maj 2024</date>
      </publicationStmt>
      <sourceDesc>
        <bibl type="article">
          <author>Repič, Hijacint</author>
          <title level="a">Vzroki in koristi terpljenja pobožnih duš.</title>
          <bibl type="monogr">
            <title level="j">Cvetje z vertov sv. Frančiška</title>
            <biblScope unit="volume">7</biblScope>
            <biblScope unit="number">2</biblScope>
            <biblScope unit="page" from="11" to="14">11-14</biblScope>
            <date when="1887">1887</date>
            <idno type="URI" subtype="URN"> URN:NBN:SI:DOC-HLXFK9DN </idno>
          </bibl>
        </bibl>
      </sourceDesc>
    </fileDesc>
    ...
```

Vir: lastno delo

Korpus je formiran kot dokument XML s krovno datoteko (element <teiCorpus>), ki vsebuje kolofon korpusa in povezave (elementi XInclude) na besedilne datoteke korpusa.

Gradivo v starejši slovenščini vsebuje arhaične leksikalne oblike, zato je bilo zaradi lažje nadaljnje uporabe in jezikoslovnega označevanja zbirke avtomatsko posodobljeno z odprtokodnim orodjem za avtomatsko posodabljanje leksike cSMTiser,[12] ki je posodobilo 13.033 oziroma 7,4 odstotka besed. Izvorne besedne oblike s tem niso

10 TEI Consortium, eds., *Guidelines for Electronic Text Encoding and Interchange*, http://www.tei-c.org/P5/.

11 *GitHub - TEIC/Stylesheets: TEI XSL Stylesheets,* https://github.com/TEIC/Stylesheets.

12 Yves Scherrer in Nikola Ljubešić, »Automatic Normalisation of the Swiss German ArchiMob Corpus Using Character-Level Machine Translation,« prispevek objavljen na letni konferenci KONVENS 2016, *13th Conference on Natural Language Processing,* Bochum, Germany, 19.–21. 9. 2016.

bile izgubljene, pač pa so posodobljene oblike, kjer se razlikujejo od izvirnih, bile pripisane tem.

V postopku avtomatskega posodabljanja starejšega jezika so bile ustrezno posodobljene na primer besede, ki imajo polglasnik zapisan z »e« (npr. *miloserčnost* > *milosrčnost, serce* > *srce*, *smert* > *smrt*), končnica -vec > -lec (*bravec* > *bralec*), besede s premeno -t- > -d- v korenu (npr. *britkosti* > *bridkosti*), -z- > -g- v korenu (npr. *druzega*, *vbozih* > *drugega*, *ubogih*), izpad -i- v *stariši* > *starši* (mn.) in *kristijani* > *kristjani*, izpad -t- v *bogatstvo* > *bogastvo*, zapis v- > u- pri nekaterih glagolih (npr. *vmreti*, *vmerl* > *umreti*, *umrl*), oblika *aposteljnov* > *apostolov* in nekatere funkcijske besede: zaimek *gdo* > *kdo*, veznik *temuč* > *temveč*, veznik in člene *toraj*/*torej* > *torej*, člene *vže* > že, predlog *sè* > *s*, ž > *z*. Nekatere starejše oblike besed so bile v postopku strojnega posodabljanja leksike spregledane (npr. predlog *mej* > med, *blager* > blagor). Te bodo skupaj z nekaterimi zaznanimi napačnimi posodobitvami (pregledno jih navaja konferenčni prispevek) popravljene v prihodnji verziji korpusa.

Nato je bil korpus jezikoslovno označen z odprtokodnim orodjem CLASSLA,[13] s katerim so bile vsaki besedi dodane njena lema oziroma osnovna oblika, njene oblikoskladenjske lastnosti in skladenjska razčlenitev povedi po sistemu Universal Dependencies za slovenski jezik.[14]

Korpus vsebuje 230 besedil oziroma 3228 odstavkov, 10.109 (avtomatsko določenih) povedi in 175.907 besed. Različica 1.0 je bila objavljena v repozitoriju CLARIN.SI pod odprto licenco Creative Commons, priznanje avtorstva (CC BY).[15] Za prevzem je na voljo v štirih stisnjenih datotekah. Vsaka vsebuje direktorij, v njem pa datoteke za eno od variant korpusa (opis prinaša konferenčni prispevek). Repozitorijski vnos je povezan s konkordančniki CLARIN.SI noSketch Engine in KonText, kar omogoča poizvedbe in analize korpusa brez znanja programiranja – z različnimi vizualizacijami in možnostjo shranjevanja rezultatov poizvedb.

V nadaljevanju je korpus CVET 1.0 podlaga za a) analizo pripovednega sloga p. Hijacinta Repiča in b) komparativno analizo izbranih zgodovinskih oblik v *CFr*. Ob Škrabčevih načelih je preverjena norma v naslednjih normativnih priročnikih: Janežičevi *Slovenski slovnici za domačo in šolsko rabo* (1863), Pleteršnikovem *Slovensko-nemškem slovarju* (1894–1895), Levčevem *Slovenskem pravopisu* (1899) in Breznikovi *Slovenski slovnici za srednje šole* (1916). Prikazan je primer analize zgodovinske paradigme na glasoslovni ravni, na kateri je bilo tudi sicer zaznanih največ razlik v primerjavi s sodobno knjižno normo, ki pa jih je orodje za avtomatsko posodabljanje leksike v korpusu CVET v glavnini pravilno prepoznalo.

13 Nikola Ljubešić, Luka Terčon in Kaja Dobrovoljc, »CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages,« prispevek objavljen na konferenci JT-DH 2024, *Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenija, 19.–20. 9. 2024, https://doi.org/10.5281/zenodo.13936406.

14 Kaja Dobrovoljc, Tomaž Erjavec in Simon Krek, »The Universal Dependencies Treebank for Slovenian,« prispevek objavljen na konferenci *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, Association for Computational Linguistics, 33–38, https://doi.org/10.18653/v1/W17-1406.

15 Diana Košir in Tomaž Erjavec, »Corpus of texts by Hijacint Repič CVET 1.0« (2024), *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1226.

Stilometrija in korpusna stilistika

Na presečišču jezikoslovja in literarnih ved je literarna stilistika, opredeljena kot jezikoslovno preučevanje sloga oziroma namenske specifične rabe jezika kot osrednjega medija literature, ki nakazuje odnos med kreativnim dosežkom (učinkom) in jezikovno manifestacijo,[16] na to pa lahko vplivajo različne nejezikovne spremenljivke, kot so žanr, avtor, zgodovinsko obdobje ipd.[17] Pri stilistični analizi opazujemo sestavine sloga (angl. *features of style*), jezikovne elemente, ki v danem korpusu izstopajo in prispevajo k oblikovanju sloga.[18] Slogovne označevalce (angl. *style markers*) Leech in Short za potrebe stilistične analize razvrščata v štiri kategorije: a) leksikalne prvine, b) slovnične prvine, c) retorična sredstva (figure in tropi) ter č) kohezija in kontekst.[19]

V okviru računalniške kvantitativne stilistike se ločita pristopa stilometrije in korpusne stilistike. Stilometrija je preštevalna metoda določanja stila in temelji na vnaprej izbranih markerjih, kot so najpogostejše besede, njihova pogostnost in razpršenost v danem korpusu, raznolikost (gostota) besedišča in *hapax legomena* oziroma *hapax dislegomena*, povprečna dolžina besed oziroma povedi, črkovni in besedni n-grami, pogostost najpogostejših besednih kolokacij itd.[20] Korpusna stilistika pa je razumljena kot uporaba teorij, modelov in metod stilistike pri korpusni analizi,[21] pri čemer je poudarek na kvalitativni analizi kvantitativnih podatkov oziroma literarnovedni interpretaciji, kjer v raziskavo vstopijo naratološke teme, kulturno-družbeni in kognitivni vidiki literature.[22]

Jasno razmejitev med obema naboroma metod in pristopov je težko določiti. Analize v članku združujejo pristope korpusnega jezikoslovja, stilometrije in korpusne stilistike, slednje zlasti pri interpretaciji izbranih prvin za zvrstno-žanrski opis besedil. Za analizo so bile izbrane štiri spremenljivke, povezane s frekvenčnostjo: najpogostejših 100 besed, najmanj pogoste besede, najpogostejši besedni 2-grami in ključne besede. Frekvenčni seznami so bili pridobljeni z orodjem noSketch Engine, s plačljivo različico Sketch Engine pa so bili analizirani besedni 2-grami. CLARIN.SI konkordančnik noSketch Engine omogoča funkcijo iskanja ključnih besed (*keywords*), pridobljenih glede na referenčni korpus PriLit 1.0,[23] ki vsebuje starejša slovenska besedila (rokopise, pridige, krajšo pripovedno prozo) od sredine 17. do sredine 19. stoletja.

16 Geoffrey N. Leech in Mick Short, *Style in Fiction: A Linguistic Introduction to English Fictional Prose* (Harlow: Pearson Education Limited, 2007), 11, 55.

17 Lesley Jeffries in Daniel McIntyre, *Stylistics* (Cambridge: Cambridge University Press, 2010), 1.

18 Leech in Short, *Style in Fiction*, 44, 56.

19 Ibid., 61–66.

20 David L. Hoover, »Frequent Word Sequences and Statistical Stylistics,« *Literary and Linguistic Computing* 17, št. 2 (2002). Jack Grieve, »Quantitative Authorship Attribution: An Evaluation of Techniques,« *Literary and Linguistic Computing* 22, št. 3 (2007). Andrejka Žejn, »Računalniško podprta stilometrična analiza pripovedne literature Janeza Ciglerja in Christopha Schmida v slovenščini,« *Fluminensia: časopis za filološka istraživanja* 32, št. 2 (2020).

21 Dan McIntyre in Brian Walker, *Corpus Stylistics: Theory and Practice* (Edinburgh: Edinburgh University Press, 2019), 15.

22 Berenike J. Herrmann, Arthur M. Jacobs in Andrew Piper, »Computational Stylistics,« v: Donald Kuiken in Arthur M. Jacobs, ur., *Handbook of Empirical Literary Studies* (Berlin, Boston: De Gruyter, 2021), 460–69.

23 Andrejka Žejn in Tomaž Erjavec, »The corpus of older Slovenian narrative prose PriLit 1.0« (2021), *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1319.

Analiza opazovanih parametrov služi opisu leksike (slogovnega in normativnega vidika), vpogledu v raznolikost besedišča in tematskemu opisu korpusa, kar razodeva avtorjevo pomensko polje, njegov spoznavni in literarni svet.[24] Pri izboru slogovnih označevalcev in njihovi interpretaciji pa je do neke mere treba upoštevati tudi subjektivno presojo in interpretativni okvir raziskovalca, na kar sta opozorila tudi Leech in Short.[25]

Korpusna analiza avtorjevega pripovednega sloga

Med najpogostejšimi 100 besedami (lemami) v korpusu CVET 1.0 so: a) glagoli *biti, imeti, moči, hoteti, reči, morati, priti, storiti, prositi, govoriti, videti, delati, vedeti, ljubiti*; b) samostalniki *Bog, svet, duša,* človek, *brat, Frančišek, gospod, dan,* življenje, *greh, srce* (izv. *serce*), *oče, volja, beseda, Jezus, milost, otrok, Kristus, križ, ljubezen, mati,* samostalniški zaimki *jaz, ti*; c) pridevniki *božji, dober, velik,* svojilni zaimki *moj, tvoj, svoj, njegov, naš*; č) prislovi *jako, dobro*; d) vezniki *da, in, ter, v, ako,* oziralni in vprašalni zaimki v vlogi veznika *kateri* (izv. *keteri*), *kakor* (izv. *kaker*), *kar, kaj, kako*; e) členki *tudi, naj,* še, *ne, le, samo*. Lema *nič* se pojavi kot samostalnik (*Oh saj gre vse sčasoma v nič!*), sam. zaimek (*Zato se ne veseli v ničemer drugem, kar je pod nebom!*) in prislov (*Keder se nič več ne želi, takrat se čuti pravo, popolno veselje*), zaimek *vsak* v pridevniški (*Priporočeval se jima je serčno vsako jutro in vsak večer*) in samostalniški rabi (*Ako je vsacemu dovoljeno storiti si iz sukna tako ali tako obleko, bo li Bog imel menj pravic do svojega?*).

Med glagoli se poleg tistih, ki so značilni za vsakdanje sporazumevalne okoliščine (pomožni glagol *biti, imeti, delati, dati*), pogosto pojavijo glagoli rekanja (*reči, prositi, govoriti,* tudi *praviti, odgovoriti, povedati*) in modalni glagoli (prvo število so pojavitve, drugo pa odstotek besedil v korpusu, kjer se lema pojavi): *moči* (597 = 75 %), *hoteti* (577 = 70,87 %) in *morati* (465 = 66,52 %). Kontekst rabe teh naklonskih glagolov je pričakovan za nabožno vzgojno literaturo, npr. *Naposled, ker brez milosti božje ne moreš nič dobrega storiti, prosi Boga, da ti bode milostljiv / Ali Bog te hoče poskušati sè skušnjavami, je li res, da ga ljubiš ali ni / Nadalje se moramo vdati božji volji, ako imamo dušne ali telesne pogreške*. V primerih je opazno prehajanje med prvo- in drugoosebnimi glagolskimi oblikami oziroma načini nagovarjanja bralca. Če v kontekstu diskurza s prevladujočo vplivanjsko vlogo opazujemo različne pripovedne perspektive ('mi', 'ti' in 'vi') za glagol *morati*, ki je z vidika naklonskosti najbolj neposreden v izražanju nujnosti, ugotovimo, da se oblika *moramo* (1. os. mn.), kjer sta pripovedovalec in bralec del množinskega subjekta, pojavi 123-krat, oblika *morate* (2. os. mn.), kjer se pripovedovalec distancira od množinskega prejemnika in se postavi v superiorno vlogo, le 11-krat, oblika *moraš* (2. os. ed.), kjer gre v primerjavi s prejšnjima za najbolj neposredno obliko nagovarjanja, pa 56-krat. Med opazovanimi je najpogostejša

24 Vesna Mikolič, *Izrazi moči slovenskega jezika* (Koper: Annales ZRS, Ljubljana: Slovenska matica, 2020). Vesna Mikolič, *Ali bereš Cankarja?* (Ljubljana: Slovenska matica, 2022).

25 Leech in Short, *Style in Fiction,* 3, 34–36.

prvoosebna množinska oblika, kjer gre za približevanje sporočevalca naslovniku, razmeroma pogosto pa pripovedovalec rabi tudi bolj neposredno, edninsko obliko, ki deluje kot okrepitev sporočila (npr. *Ako hočeš torej vživati Boga, se moraš odpovedati grešnim posvetnim tolažbam*).

Za ugotovitev, ali raba naklonskega glagola *morati* v korpusu CVET statistično značilno izstopa, sta bila za primerjavo izbrana dva korpusa starejšega slovenskega jezika: korpus KDSP[26] s pripovedno prozo 19. stoletja in korpus Slovenska periodika (1771–1914).[27] V korpusu KDSP se glagol *morati* pojavi v 99,6 odstotka besedil in predstavlja 0,16 odstotka korpusa, v korpusu periodike je prisoten v 73,19 odstotka besedil in obsega 0,15 odstotka korpusa, v korpusu CVET pa se *morati* pojavi v 66,5 odstotka besedil, medtem ko vse pojavitve predstavljajo 0,22 odstotka korpusa. V korpusu periodike in v CVET je prvi podatek (odstotek besedil s pojavnico) pričakovano bistveno nižji kot v KDSP, ker vsak posamezni članek predstavlja svojo besedilno enoto. Je pa v korpusu CVET opazen najvišji odstotek pojavnosti glagola v razmerju do celotnega korpusa. Kvantitativna analiza je, upoštevajoč zvrstno raznoliko primerjalno gradivo, pokazala višjo prisotnost naklonskega glagola *morati* v pripovednem slogu patra Repiča v besedilih (polliterarne) zvrsti versko-vzgojnega periodičnega članka, kot se je to pokazalo za korpus literarnih besedil in periodične članke v glavnem necerkvene vsebine. Smiselno bi bilo zato opraviti dodatne kvalitativne analize kontekstov rabe, da bi se lahko utemeljilo razmerje med sporočanjsko in vplivanjsko vlogo besedil ter piščevo izbiro posameznih naklonskih sredstev kot stilistično karakteristiko avtorja in literarne zvrsti oziroma žanra.

Religiozni diskurz tipično opredeljuje moralno-vrednostno razmerje dobro : slabo in med najpogostejšimi besedami v korpusu CVET so predvsem te, ki semantično sodijo v vrednostno oznako »dobro« (*ljubiti*, *Bog*, *duša*, *srce*, *milost*, *ljubezen*, *dober*, *dobro*).

Večina od prvih 100 besed se pojavi v več kot 60 odstotkih vseh besedil, kar pomeni, da je njihova zastopanost razpršena po celotnem korpusu.

Po pogostosti izstopa raba zaimka *kateri*, ki se pojavi kar 1822-krat in v 90,87 odstotka besedil. Izrazita je vezniška raba, kjer uvaja podredje z oziralnim odvisnikom, v današnji knjižni slovenščini bi ga nadomestil zaimek *ki* (npr. *Ali ti, moj Jezus, keteremu je vse odkrito, in keteri si vstvaril vse v meri in številu …*). Tudi sicer hiter pregled gradiva pokaže, da so v Repičevih tekstih pogoste dolge povedi s podredji, kar potrjuje statistični podatek za povprečno dolžino povedi (izračunan na podlagi števila besed in povedi v korpusu), ki znaša 17 besed (npr. *Kar se tiče pa vsacega od nas v tacih občnih nesrečah, pomislimo sè zaupanjem na Boga, keteri zna število naših las, in keterih nam niti eden ne pade z glave brez njegove presvete volje, kar pomeni, da se nam čisto nič ne more pripetiti, kar Bog neče ali ne dopusti.*).

26 Lucija Mandić in Tomaž Erjavec, »Corpus of longer narrative Slovenian prose KDSP 1.0« (2023), *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1823.

27 Filip Dobranić, Bojan Evkoski in Nikola Ljubešić, »Corpus of Slovenian periodicals (1771–1914) sPeriodika 1.0« (2023), *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1881.

Med načinovnimi prislovi po frekvenčnosti izstopa leksem *jako* (342-krat; ob enajstih pojavitvah sopomenke *zelo*; npr. *To prepričanje nas jako tolaži, ker vemo, da nas Bog more tako lehko poklicati k sebi, ko smo v sreči in zdravju, kaker ko smo v občnih nesrečah*). Glede na konkordance in metapodatke o izvirnih besedilih, ki jih posreduje korpus, je mogoče ugotoviti, da se starejša oblika 'jako' pojavlja po vsem gradivu (v 57,83 odstotka besedil), novejša oblika 'zelo' pa ob njej nekoliko bolj konsistentno po letu 1913, kar kaže na utrjevanje nove oblike v knjižni normi. Ker se prislov *jako*, kot bo razvidno v nadaljevanju, pojavi tudi med ključnimi besedami glede na referenčni korpus starejših pripovednih besedil, ga je mogoče označiti kot značilno potezo patrovega pripovednega sloga.

Med najmanj pogostimi besedami (ena do dve pojavitvi, korpus takih lem navaja 4656) so nepričakovane tvorjenke: manjšalnice (človeček, *stvarca*), *kletvina* (ob *kletev*), *pekočina*, *bogatin* (ob *bogataš* in posamostaljenem pridevniku *bogati* (mn.)), *tolažilo* (ob *tolažba*), *zaveržek*, *pomoček*, kroatizem škatulje (mn.), *nasladnost* (ob *naslada*), *ostrašen* (ob *prestrašen*), modifikacija glagolskega vida (*peljavati*, *zmaščevati*, *oveseliti*), onomatopoetični glagol *zberbrati* (= na hitro, nerazločno oziroma nepremišljeno izreči, zmoliti); deležnik na -č: *kretajoč (se)*, *pazeč*, deležnik na -ši: *znebivši*, *zbudivši*; členek *znabiti*, izvirni besedni zvezi čudapoln *mir* in *prvak apostolov*, *zvezdoznanka*, *zgoja* (ob *vzgoja*), *zgojitelj*, *razjasnjenje*, život (ob *telo*), *nesposobnost*, *neizmernost*, *gotovost*, *enakoličnost* itd. Nepričakovane oziroma z vidika rabe redke dvojnice lahko razumemo kot odraz tedaj še nepoenotene pisne norme (npr. *vzgoja* ali *zgoja*), nekatere glagolske oblike in izvirne tvorjenke, ki lahko glede na dani kontekst okrepijo pomen, pa kažejo na bogato besedišče in piščevo ustvarjalno rabo jezika. Skozi uporabo ekspresivne, slogovno zaznamovane leksike se pisec hkrati lahko čustveno razodeva, izraža naklonjenost oziroma nenaklonjenost.

Tretja opazovana spremenljivka so besedni 2-grami (najpogostejše zveze dveh lem). Namen je pridobiti vpogled v najpogostejše in ustaljene ter redke in nepričakovane besedne zveze, ki tematsko (še bolj ilustrativno kot posamezne besede) označujejo besedilni korpus, z vidika pripovednega sloga pa je mogoče v danih kontekstih opazovati njihovo semantično plat in estetsko funkcijo (dejanski in preneseni pomen, učinek v besedilu). Izluščene so bile zveze 'samostalnik + pridevnik' in 'samostalnik + samostalnik', pri čemer so bila izločena osebna lastna imena. V analizi besednih zvez so zaradi večje preglednosti navedene posodobljene leksikalne oblike.

Med najpogostejšimi besednimi zvezami so samostalniške zveze s pridevnikom *božji/a/e* (volja, milost, dobrota, ljubezen, milosrčnost, previdnost, usmiljenje, čast, blagoslov, sodba, naredba, roka, sin, služabnik, mož, prijatelj, svetnik; miza, služba, pot, dar, beseda, zapoved, pomoč, martra), zveze z lastnostnim pridevnikom *velik/a/o* (ki se v korpusu v osnovni in stopnjevani obliki pojavi kar 727-krat) za izražanje visoke stopnje (milost, veselje, ljubezen, svetost, zaupanje, dobrota, ponižnost, milosrčnost, slava, sreča, spoštovanje, žalost; grešnik, greh, dolžnost, tolažba, čudež, uboštvo, čudodelnik, čednost, imenitnost, prijatelj, skušnjava, sladkost, popolnost, razloček, sočutje, strah), za izražanje skladnosti z zaželenim oziroma resničnim so pogoste zveze

s *pravi/a/o* (pobožnost, gorečnost, veselje; prijatelj, ponižnost, služabnik, hrana, vera), enako učinkuje tudi zveza *neovrgljiva resnica*. Podobno velja za besedne zveze z drugimi pridevniki za stopnjevanje navzgor: besedotvorni pridevniški presežniki z obrazilom (presveta volja, presveto ime, presladko ime, presladke solze, presladka beseda, velikansko dejanje) in pomenski presežniki[28] (*večen/a/o* življenje, slava, modrost, plačilo, krona, resnica, ljubezen, pogubljenje, ogenj) – v krščanskem diskurzu so pozitivni atributi vezani na Boga in posmrtno življenje v nebesih, negativni pa na pekel; *nepopisljiv/a/o* (veselje, radost, žalost, ginjenost, bridkost, sladkost, želja), *neskončna dobrota, neizmerna* (dobrota, ljubezen), *neomejeno zaupanje, izredna milost, imeniten stan, dragocen zaklad, strašna bolečina*. Pogoste so zveze s pomensko nasprotnimi pridevniki *svet(i)/a/o* (obhajilo, maša, cerkev, zakrament, pismo, razpelo, volja, življenje, apostol, mož) in *pobožen(/-ni)/na/no* (človek, kristjan, redovnik, ljudstvo, starš, mož, žena, kmet; molitev, pesem, misel, duša) ter *posveten/na/no* (čast, slava, veselje, naslada, razveseljevanje, veselica, človek, nečimrnost, tolažba) in *grešen(/-ni)/na/no* (življenje, naslada, natura, priložnost, veselica). Izluščene so bile še nekatere frekventne besedne zveze, ki kažejo na subjektivno vrednotenje (*otročje zaupanje* (= pozitivno, pristno), *neprecenjena nedolžnost, bridko trpljenje, prazna hvala, dober* (kristjan, človek, oče; glas, uspeh), *uboga žena, lepa čednost, sladke besede, goreča molitev, goreča pobožnost, stanovitna molitev*, ekspresivne besedne zveze (*napuh življenja, poželenje oči, peklenska hudoba, znamenje mlačnosti, dejanje čednosti, dušna bolezen, dušna suhota*), biblijske in izvirne metafore (*dušni pastir, božji poslanec, prvak apostolov, solzna dolina, smrtna ura, smrtna postelja, strupena kača, goreča peč* (= trpljenje); *nesrečno spanje* (= brezbožno življenje), *tečajna zvezda* (= Bog), *kukec nečimrnosti* (črv = greh), druge besedne zveze, ki preraščajo v retorične trope, na primer pretiravanje ali hiperbola (*morje bridkosti, brezno nesreč, brezno ponižnosti*) in bistroumni nesmisel ali oksimoron (*neusmiljeno usmiljenje, radovoljna raztresenost*).

Potem so bile v fokusnem korpusu CVET analizirane ključne besede skozi primerjavo z izbranim referenčnim korpusom starejše pripovedne proze PriLit, ki vsebuje starejša besedila od sredine 17. do sredine 19. stoletja in je zato relevanten vir za opazovanje razvoja tem do obdobja, ki ga zajema korpus CVET. Poleg tega sta bila korpusa podobno procesirana in posodobljena z enakim orodjem, kar omogoča večjo primerljivost gradiva.

Med 100 ključnimi besedami se za korpus CVET pojavijo naslednje leme:

– samostalniki: *Frančišek, Monald* (izv. tudi *Monaljd*),[29] zemljepisna imena *Asiz* (= Assisi), *Koper, Padova*; *redovnik, frančiškan, sobrat, tretjerednik, zavetnica, voditelj*, častilec (izv. častivec), *posvetnjak, vodilo, miloščina, milosrčnost* (izv. *miloserčnost*), *sočutje, blager* (= blagor; trikrat samostalniška raba v pomenu sreča, blagoslov), *način* (prim. *na ta(k)/en/pervi/drugi/vsak način*), *sredstvo, oblika* (prim. živeti po obliki sv. evangelija = živeti skladno z nauki in zgledi iz evangelija), *kesanje, naslada, pogrešek, vdanost*, izglagolska samostalnika *zatajevanje, občevanje*;

28 Mikolič, *Izrazi moči*, 57, 58.

29 Krajšava izv. se v članku rabi za 'izvirno', torej tako, kot je zapisano v neposodobljenem korpusu besedil.

- pridevniki: *redoven, blažen, brezmadežen, blagoslovljen, presladek, izreden, nepopisljiv,* (Marija) *Porcijunkuljska* (= nanaša se na cerkev Marije Angelske pri Assisiju, t. i. Porcijunkulo), *vsakovrsten* (izv. *vsakoversten*), *brezštevilen*;
- glagoli: *pridigati, oznanjevati, občevati* (= biti v stiku, sporazumevati se), *spolnjevati* (redko *izpolnjevati*), *spodbujati, rabiti, zanemarjati, zoperstavljati se, vničiti* (= uničiti);
- prislovi: *jako, kesneje* (= kasneje), *naposled*;
- vezniki: *koliker* (= kolikor), čiger (= čigar), *keteri* (= kateri), *mariveč* (= marveč);
- členki: *nikaker* (= nikakor), *seveda* (pisano tudi *se ve da*), *potemtakem, blager* (= blagor; 26-krat členkovna raba).

Izkazalo se je, da so bile poleg frekvenčno izstopajočih kot ključne besede prepoznane tudi besede, ki niso bile posodobljene ali pa so v postopku posodabljanja pridobile napačno obliko.

Ugotoviti je mogoče, da polnopomenske besede nedvomno pripadajo religijskemu (krščanskemu) diskurzu, nekatere med njimi še podrobneje usmerijo na red sv. Frančiška Asiškega (npr. *Frančišek, Asiz, frančiškan, vodilo, Porcijunkuljska, redovnik, redoven, sobrat, tretjerednik, voditelj*) in na lokalno okolje, od koder je pisec p. Repič prihajal (*bl. Monald Koprski, Koper*).

Med glagolskimi ključnimi besedami so pogosti glagoli rekanja, vezani na širjenje krščanske vere (liturgijo, pastoralo), ob tem pa tudi nekaj takšnih, ki v danih kontekstih pomenijo dejanja kršenja krščanskih načel.

Pridevniki in pomenski presežniki *blažen, brezmadežen, blagoslovljen, presladek, izreden, nepopisljiv, vsakovrsten, brezštevilen* ter členek *blagor* imajo ob sopojavitvi s pomensko pozitivnimi samostalniki močan pozitivni naboj, izražajo odobravanje in občudovanje koga/česa. V vsakem primeru pa z vidika intenzitete jezika omenjene jezikovne prvine okrepijo pomen ubesedenega.[30] V smer krepitve argumenta vodijo tudi pogosta raba prislova *jako* (redko *zelo*) in pomensko nasprotna si členka *nikaker* in *seveda* – slednji skupaj s členkom *potemtakem* nastopa tudi v vlogi diskurznega označevalca, ko usmerja potek diskurza z vrednotenjem, potrjevanjem in sklepanjem ter tako vpliva na bralčevo recepcijo besedila.

Primerjalna analiza jezikovne norme

V nadaljevanju je predstavljen primer korpusne analize za izbrano glasoslovno paradigmo, pri kateri je na prelomu 19. v 20. stoletje prihajalo do razlik med jezikoslovci, in sicer zapis besedotvorne končnice *-lec/-vec* v samostalnikih tipa *bralec*.[31] Orisana je norma v tedaj pomembnejših jezikoslovnih priročnikih, Škrabčeva stališča pa so preverjena v korpusnem gradivu.

30 Mikolič, *Izrazi moči*, 56, 66, 67.

31 Razhajanja so se sicer nadaljevala tudi še v 20. stoletje, npr. v polemikah ob izdaji *Slovenskega pravopisa* 1962.

V zvezi s samostalniško končnico *-vec*/*-lec* v poimenovanju delavnih ali delujočih oseb Janežič v *Slovenski slovnici* (1863) navaja, da se ta tvorijo tako, da se med osnovo in pripono *-ec* pri odprtih zlogih doda črka *v* (npr. *bravec, pevec, pivec, delavec, igravec, morivec, pisavec, poslušavec, svetovavka*). Ob tem dodaja, da nekateri pisatelji (slovenski in srbohrvaški) takšne samostalnike tvorijo iz preteklega deležnika in pišejo *delalec, igralec, poslušalec, svetovalec, tkalec, vladalec*.[32]

Škrabec se je v razpravah vprašanja lotil prek opazovanja zgodovinskega razvoja oblik in v težnji po enotni razlagi prišel do naslednjih ugotovitev: a) nekatere besede s pripono *-lec* [*-ləc*] so se razvile iz poimenovanj za orodje, navadno s samostalniško končnico *-lo* (npr. *tkalec, rilec, motovilec*); b) nekatere besede s pripono *-vec* [*-vəc*] so nastale iz glagolskega pridevnika na *-av(en)* (npr. *smrkavec, delavec, igravec*)[33]; c) besede za delujočo oziroma delavno osebo lahko imajo pripono z *l* ali *v*, s tem da prva pomeni tistega, ki je glagolsko dejanje že izvršil/prestal (nastala iz deležnika na *-l*, npr. *pogorelec*), druga pa tistega, ki ga vrši ali more vršiti; č) mogoče so tudi dvojnice v primerih, kjer gre za pomensko razločevanje (pripono z *v* naj bi imela poimenovanja, izpeljana iz glagolskih pridevnikov, »ki zaznamujejo nagnjenje, navado, službo ali lastnost osebe, ki kaj stanovitno dela«,[34] pripono z *l* pa poimenovanja oseb, ki so kaj dovršile; npr. *morivec* in *morilec*, *prebivavec*, ko je mišljen kraj, in *prebivalec*, ko se nanaša na prebivanje).[35] Stramljič Breznik poudarja, da je Škrabec s takšno utemeljitvijo dopuščanja obstoja obeh oblik, tvorjenih iz iste podstave, uveljavljal razlikovalna načela, ki jih je obenem kritiziral pri sodobnikih (npr. pri Perušku).[36] V korpusu CVET je mogoče najti naslednja primera rabe: *Ako pa vi stariši to dosežete, namreč pravo kerščansko življenje vaših otrok, srečni vi, ker ste dosegli konec zgoje, ketera je v pravem pomenu besede: skerb, da otroci postanejo nebeški* ***prebivavci***; *Ne živi kaker da bi bil* ***prebivalec*** *tega sveta.* V korpusnem gradivu semantično razlikovanje med oblikama sicer ni dosledno upoštevano, pojavljata se denimo obe besedni zvezi, *nebeški prebivavec*/*prebivalec*, kar kaže na ohlapno utemeljitev pomenskega razlikovanja obeh oblik (po letu 1898 je v rabi le oblika na *-vec*).

V korpusu CVET se pojavijo naslednje oblike:

– na *-vec: bravec, pisavec, poslušavec, obrekovavec, opravljavec, klepetavec, častivec, hvalivec, ljubivec, zapeljivec, lovec, tergovec, kupčevavec, pohujšljivec, molivec, lažnjivec, delavec, izdajavec, zaničevavec, vdovec, slepivec, godernjavec, togotljivec, delivec, hinavec, zapeljivec, vbijavec, premišljevavec, ražaljivec, skušnjavec, spolnjevavec, podiravec,* in ženske oblike *nasledovavka, napovedovavka, hinavka;*
– na *-lec: obiskovalec, izgojevalec, odgojevalec, tekalec, prilizovalec, oporekovalec, gledalec, spričevalec, posvečevalec;*

32 Anton Janežič, *Slovenska slovnica za domačo in šolsko rabo* (Celovec: Janez Leon, 1863), 119.
33 Stanislav Škrabec, *Jezikoslovna dela 1* (Nova Gorica: Frančiškanski samostan Kostanjevica, 1994), 46.
34 Irena Stramljič Breznik, »Škrabčeva obravnava priponskih obrazil (a/i) v/l (əc),« v: Jože Toporišič, ur., *Škrabčeva misel II* (Nova Gorica: Frančiškanski samostan Kostanjevica, 1997), 197.
35 Stanislav Škrabec, *Jezikoslovna dela 2* (Nova Gorica: Frančiškanski samostan Kostanjevica, 1994), 425, 426.
36 Stramljič Breznik, »Škrabčeva obravnava,« 197.

– dvojne oblike:[37] *zatajevavec* (1)/*zatajevalec* (1), *posredovavec* (2)/*posredovalec* (1), *prebivavec* (5)/*prebivalec* (3), *svetovavec* (1)/*svetovalec* (2), *spoznavavec* (4)/*spoznavalec* (3), *posnemovavec* (2)/*posnemovalec* (4), *zmagovavec* (1)/*zmagalec* (1).

V kasnejši razpravi Škrabec ugotavlja, da je prava končnica poimenovanj delavnih oseb *-vec* in ne *-lec*,[38] kar potrjuje tudi znatno večji obseg oblik na *-vec*, med dvojnicami pa je prevlada tvorjenk z *-vec* opazna po letu 1900. Škrabec je verjel, da lahko le normativni priročnik, kot je bil Pleteršnikov slovar, naredi konec jezikovni pravdi med zapisovanjem *bravec*/*bralec*.[39]

Pri tem jezikoslovnem vprašanju se je pokazal Škrabčev vpliv na Levca, ki se je odločil za prepoved zapisovanja besednih oblik tipa *bralec* z namenom, da bi ustavil »elkanje«.[40] V *Slovenskem pravopisu* (1899) pri tvorjenju samostalnikov s priponami navaja, da samostalniki na *-ec* pomenijo delujoče osebe, med korenom in pripono pa stoji soglasnik *-v-* (npr. *bravec, igravec, svetovavec, morivec, pisavec*; kot napačne, čeravno v obči rabi, navaja oblike: *delalec, igralec, morilec, pisalec, poslušalec, svetovalec* itd.). Kot pravilne oblike s pripono *-lec* pa navaja nekatere tvorjenke iz preteklega deležnika (npr. *tkalec, pogorelec, črešnjejelec, gnilec, umrlec, otelec, pismoznalec, prišlec, rilec, vrtelec, vžgalec* idr.).[41]

Tudi Pleteršnik (po Janežiču in Škrabcu) v svojem slovarju (1894–95) navaja oblike s pripono *-vec* (npr. *bravec, poslušavec, igravec, morivec, prebivavec, delavec*), enako kot predhodniki pa oblike tipa *tkalec* (*motovilec, pogorelec* itd.).

Breznik v *Slovenski slovnici* (1916) zmedo med zapisovanjem *l* ali *v* v poimenovanjih delujočih oseb pojasnjuje z enako izreko pripone *-vəc*/*-ləc* (od 17. stoletja dalje). Deloma povzema Škrabčeve ugotovitve in dopušča dvojnično rabo v primerih, kjer so se iz prvotnih končnic *-(a/i)vec* v novejši pisavi pojavile *-(a/i)lec* (npr. *bravec*/*bralec, poslušavec*/*poslušalec, igravec*/*igralec, šivavec*/*šivalec, zaničevavec*/*zaničevalec*) oziroma kjer je *l* v korenu (npr. *volivec*/*volilec, delivec*/*delilec, ponavljavec*/*ponavljalec*). Prvotni *v* pa je (kot pri Škrabčevi ugotovitvi o izpeljavi iz glagolskega pridevnika) ostal v besedah *brivec, pivec, delavec, kimavec, pevec, tajivec*. Samo *-lec* se piše za osebe, ki so kaj storile ali trpele (*umrlec, prišlec, osamelec, prebivalec*) in v izpeljankah iz samostalnikov na *-lo* (*motovilec, rilec, tkalec* itd.).[42]

37 V oklepaju je za primerjavo zapisano število pojavnic v korpusu.
38 Stanislav Škrabec, *Jezikoslovna dela 3* (Nova Gorica: Frančiškanski samostan Kostanjevica, 1995), 132, 133.
39 Ibid., 84–86.
40 Helena Dobrovoljc, *Pravopisje na Slovenskem* (Ljubljana: Založba ZRC, ZRC SAZU, 2004), 49.
41 Fran Levec, *Slovenski pravopis* (Dunaj: Cesarska kraljeva zaloga šolskih knjig, 1899), 54.
42 Anton Breznik, *Slovenska slovnica za srednje šole* (Celovec: Družba sv. Mohorja, 1916), 31.

Razprava in sklepni razmislek

V korpusu CVET so zbrana vsa besedila patra Hijacinta Repiča, ki so izvirni in prevodni zapisi po tujejezičnih predlogah (vir avtor konsistentno navaja v samem besedilu oziroma opombah). Prikazana korpusna analiza je bila usmerjena v raziskovanje avtorjevega pripovednega sloga. Uporabljene so bile funkcije, ki jih ponuja prostodostopni konkordančnik noSketch Engine na CLARIN.SI: seznam najpogostejših in najmanj pogostih besed (lem) in seznam ključnih besed (lem) glede na izbrani referenčni korpus ter analiza najpogostejših besednih zvez (2-gramov) z orodjem Sketch Engine. Analiza frekvenčnosti besedja je podala nekatere pričakovane rezultate, kot so samostalniki in pridevniki ter besedne zveze, značilni za religiozni diskurz, ter glagoli, vezani na vsakdanje praktičnosporazumevalne okoliščine, pri čemer so po pogostosti izstopali modalni glagoli. Z vidika naklonskosti bi bilo smiselno primerjati različne verske pripovednike, kjer bi primerjava jezikoslovnih in diskurzivnih razsežnosti, ki kažejo na razmerje med sporočanjsko in vplivanjsko vlogo besedil in na sporočevalčev odnos do vsebine (vzgojno-moralne problematike) ter naslovnika, lahko podala širši vpogled v razvoj krščanske misli.

Analiza najpogostejših samostalnikov in pridevnikov potrjuje vsebinski okvir besedil, skozi kontekste rabe posamezne leksike pa je mogoče dostopati do vseh pojavitev v gradivu. Pri Repiču se pojavijo nekatere besedne zveze, kjer imajo posamezne sestavine ekspresiven, dobeseden oziroma prenesen pomen (ustaljene ali izvirne metafore, prispodobe): *otročje zaupanje*, *neprecenjena nedolžnost*, *prazna hvala*, *lepa čednost*; *napuh življenja*, *poželenje oči*, *peklenska hudoba*, *znamenje mlačnosti*, *dejanje čednosti*, *dušna bolezen*, *dušna suhota*; *prvak apostolov*, *solzna dolina*, *smrtna ura*, *smrtna postelja*, *strupena kača*, *goreča peč*, *nesrečno spanje*, *tečajna zvezda*, *kukec nečimrnosti*; *morje bridkosti*, *brezno nesreč*, *brezno ponižnosti*; *neusmiljeno usmiljenje*, *radovoljna raztresenost*.

Analiza najpogostejših besednih 2-gramov je pokazala, da med zvezami 'pridevnik + samostalnik' prevladujejo: (1) pridevniki, ki označujejo veliko oziroma presežno lastnost samostalnika (*velik*, *velikanski*, *presladek*, *presvet*; *večen*, *nepopisljiv*, *neskončen*, *izreden* itd.) ali lastnost, ki je v skladu z zaželenim/resničnim (*pravi*, *neovrgljiv*, *dober*, *lep*); (2) pomensko diametralno nasprotni pridevniki (*svet*, *pobožen – posveten*, *grešen*), ki v okviru religioznega diskurza pridobijo simbolni pomen pravih/dobrih in napačnih/slabih življenjskih zgledov, nazorov in izbir ter tako nastopajo v vzgojni funkciji.

Opazovani najpogostejši prislovi, zaimki, vezniki in členki zaznamujejo tipično skladnjo (za Repičev slog so značilne dolge povedi z veliko podredji, pogosta je raba veznika *kateri*) oziroma so označevalci intenzitete, ki bodisi krepijo (npr. *jako*, *zelo*; *vsak*; *seveda*, *nikakor*) ali šibijo (*le*, *samo*) končni pomen in argumentativno moč sporočila.[43] Šele po primerjalni analizi z drugimi pripovedniki pa bi bilo mogoče sklepati, ali gre tudi za specifične lastnosti Repičevega sloga.

Primer primerjalne normativne analize v članku je osvetlil različne poglede tedanjih jezikoslovcev na zapisovanje samostalniške pripone *-vec*/*-lec*.

43 Mikolič, *Izrazi moči*, 79–82.

Ugotovljeno je bilo, da je imel Škrabec v jezikoslovnih razpravah tendenco jasno definirati in semantično razmejiti rabo ene in druge končnice v primeru dvojnic, kar se je v praksi izkazalo za 'prisiljeno' in se potrjuje tudi na korpusnem gradivu. Je pa njegovo zagovarjanje zapisa besed tipa *bravec* vplivalo na to, da je tudi Breznik v slovnici še dopuščal dvojnično rabo.

Korpus CVET 1.0 je prva takšna jezikovna zbirka besedil Škrabčevega verskega glasila *Cvetje z vertov sv. Frančiška*, velja pa omeniti, da je bilo besedje *Cvetja* sporadično zajeto že v Pleteršnikovem *Slovensko-nemškem slovarju* (1894–1895). Raziskava je pokazala, kako lahko korpusno gradivo, označeno z metapodatki o avtorstvu in objavi (leto, letnik, zvezek), služi za komparativne sinhrono-diahrone raziskave normiranja slovenskega jezika na prelomu 19. v 20. stoletje v periodičnem tisku in za raziskave udejanjanja knjižne norme, ki jo je urednik Škrabec podrobno predstavil v svojih jezikoslovnih znanstvenih razpravah na platnicah, v sami vsebini *Cvetja*. Hkrati korpus ponuja tematski vpogled v vsebino verske revije (versko-moralni nauki, krščanska vzgoja, hagiografija itd.). V nasprotju s primerljivim verskim periodičnim tiskom, denimo Slomškovimi *Drobtinicami*, ki so pričele izhajati pol stoletja prej in v podobnem formatu kot *Cvetje*, je vsebina slednjega (razen jezikoslovnih razprav na platnicah) še neraziskana. Korpus objav patra Hijacinta Repiča bi lahko z nadgradnjo postal podkorpus obsežnejšega korpusa celotnih izdaj revije *Cvetje z vertov sv. Frančiška*.

Zahvala

Članek je nastal v okviru raziskovalnega programa P5-0409 *Razsežnosti slovenstva med lokalnim in globalnim v začetku tretjega tisočletja*, ki ga financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS) iz državnega proračuna.

Viri in literatura

Literatura

Dobrovoljc, Helena. *Pravopisje na Slovenskem*. Ljubljana: Založba ZRC, ZRC SAZU, 2004.

Dobrovoljc, Kaja, Tomaž Erjavec in Simon Krek. »The Universal Dependencies Treebank for Slovenian.« V: *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, 33–38. https://doi.org/10.18653/v1/W17-1406.

Dović, Marijan. *Slovenski pisatelj. Razvoj vloge literarnega proizvajalca v slovenskem literarnem sistemu*. Ljubljana: Založba ZRC, ZRC SAZU, 2007. https://doi.org/10.3986/9789610503200.

Grieve, Jack. »Quantitative Authorship Attribution: An Evaluation of Techniques.« *Literary and Linguistic Computing* 22, št. 3 (2007): 251–70. https://doi.org/10.1093/llc/fqm020.

Herrmann, Berenike J., Arthur M. Jacobs in Andrew Piper. »Computational Stylistics.« V: *Handbook of Empirical Literary Studies*, uredila Donald Kuiken in Arthur M. Jacobs, 451–86. Berlin, Boston: De Gruyter, 2021. https://doi.org/10.1515/9783110645958.

Hoover, David L. »Frequent Word Sequences and Statistical Stylistics.« *Literary and Linguistic Computing* 17, št. 2 (2002): 157–80. https://doi.org/10.1093/llc/17.2.157.

Jeffries, Lesley in Daniel McIntyre. *Stylistics*. Cambridge: Cambridge University Press, 2010.https://doi.org/10.1017/CBO9780511762949.

Korošak, Bruno. *P. Stanislav Škrabec, frančiškan, v očeh sodobnikov*. Ljubljana: Založba Brat Frančišek, 2001.

Košir, Diana in Tomaž Erjavec. »Korpus CVET 1.0: izdelava, opis in analiza zbirke starejših besedil v verski periodiki.« Prispevek objavljen na letni konferenci JT-DH 2024, *Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenija, 19.–20. 9. 2024. https://doi.org/10.5281/zenodo.13936400.

Leech, Geoffrey N. in Mick Short. *Style in Fiction: A Linguistic Introduction to Englis Fictional Prose*. 2nd ed. Harlow: Pearson Education Limited, 2007. Pridobljeno 10. 5. 2024. https://sv-etc.nl/styleinfiction.pdf.

Ljubešić, Nikola, Luka Terčon in Kaja Dobrovoljc. »CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.« JT-DH 2024, *Conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenija, 19.–20. 9. 2024. https://doi.org/10.5281/zenodo.13936406.

McIntyre, Dan in Brian Walker. *Corpus Stylistics: Theory and Practice*. Edinburgh: Edinburgh University Press, 2019.

Mikolič, Vesna. »Povezanost narodne in jezikovne zavesti.« *Jezik in slovstvo* 45, št. 5 (2000): 173–86. Pridobljeno 15. 5. 2024. URN:NBN:SI:DOC-5BD79SLS.

Mikolič, Vesna. *Izrazi moči slovenskega jezika*. Koper: Annales ZRS; Ljubljana: Slovenska matica, 2020.

Mikolič, Vesna. *Ali bereš Cankarja?*. Ljubljana: Slovenska matica, 2022.

Orožen, Martina. *Oblikovanje enotnega slovenskega knjižnega jezika v 19. stoletju*. Ljubljana: Filozofska fakulteta, Znanstveni inštitut Filozofske fakultete, 1996.

Perenič, Urška. *Empirično-sistemsko raziskovanje literature: Konceptualne podlage, teoretski modeli in uporabni primeri*. Ljubljana: Slavistično društvo Slovenije, 2010. Pridobljeno 10. 5. 2024. https://zdsds.si/tiskovina/562/.

Perenič, Urška. »Čitalništvo v perspektivi družbenogeografskih dejavnikov.« *Slavistična revija* 60, št. 3 (2012): 365–82. Pridobljeno 15. 5. 2024. https://srl.si/ojs/srl/article/view/COBISS_ID 50413154.

Scherrer, Yves in Nikola Ljubešić. »Automatic Normalisation of the Swiss German ArchiMob Corpus Using Character-Level Machine Translation.« Prispevek objavljen na letni konferenci KONVENS 2016, *13th Conference on Natural Language Processing*, Bochum, Germany, 19.–21. 9. 2016. Pridobljeno 15. 10. 2025. https://archive-ouverte.unige.ch/unige:90846

Stramljič Breznik, Irena. »Škrabčeva obravnava priponskih obrazil (a/i) v/l (əc).« V: *Škrabčeva misel II*, uredil Jože Toporišič, 193–200. Nova Gorica: Frančiškanski samostan Kostanjevica, 1997.

Šalamun, Miša. *Slovensko primorsko časopisje. Zgodovinski pregled in bibliografski opis*. Koper: Lipa, 1961.

Žejn, Andrejka. »Računalniško podprta stilometrična analiza pripovedne literature Janeza Ciglerja in Christopha Schmida v slovenščini.« *Fluminensia: časopis za filološka istraživanja* 32, št. 2 (2020): 137–58. Pridobljeno 30. 4. 2024. https://doi.org/10.31820/f.32.2.5.

Spletni viri

Dobranić, Filip, Bojan Evkoski in Nikola Ljubešić. »Corpus of Slovenian periodicals (1771–1914) sPeriodika 1.0« (2023). *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1881.

Košir, Diana in Tomaž Erjavec. »Corpus of texts by Hijacint Repič in 'Cvetje z vertov sv. Frančiška' CVET 1.0« (2024). *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1226.

Mandić, Lucija in Tomaž Erjavec. »Corpus of longer narrative Slovenian prose KDSP 1.0«(2023) *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1823.

Pleteršnik, Maks. *Slovensko-nemški slovar*, 1894–1895. Spletna izdaja. Ljubljana: ZRC SAZU, 2010. Pridobljeno 2. 1. 2023. http://bos.zrc-sazu.si/pletersnik.html.

Žejn, Andrejka in Tomaž Erjavec. »The corpus of older Slovenian narrative prose PriLit 1.0« (2021). *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1319.

Tiskani viri

Breznik, Anton. *Slovenska slovnica za srednje šole*. Celovec: Družba sv. Mohorja, 1916. https://www.fran.si/slovnice-in-pravopisi/35/1916-breznik.

Janežič, Anton. *Slovenska slovnica za domačo in šolsko rabo*. Celovec: Janez Leon, 1863. https://www.fran.si/slovnice-in-pravopisi/26/18631854-janezic.

Levec, Fran. *Slovenski pravopis*. Dunaj: Cesarska kraljeva zaloga šolskih knjig, 1899. https://fran.si/slovnice-in-pravopisi/34/1899-levec.

Škrabec, Stanislav. *Jezikoslovna dela 1*. Nova Gorica: Frančiškanski samostan Kostanjevica, 1994.

Škrabec, Stanislav. *Jezikoslovna dela 2*. Nova Gorica: Frančiškanski samostan Kostanjevica, 1994.

Škrabec, Stanislav. *Jezikoslovna dela 3*. Nova Gorica: Frančiškanski samostan Kostanjevica, 1995.

Diana Košir, Tomaž Erjavec

CORPUS ANALYSIS OF NARRATIVE STYLE AND LINGUISTIC NORM IN AN OLDER RELIGIOUS PERIODICAL

SUMMARY

The paper presents the process of creating and linguistically tagging the CVET 1.0 corpus, which contains the texts of Father Hijacint Repič in the older Slovenian language, published in the religious journal *Cvetje z vertov sv. Frančiška* between 1881 and 1916. The texts were obtained in PDF format from the dLib portal, edited in Microsoft Word and then converted to TEI. Older words were automatically updated using an open-source normalisation tool that facilitates corpus search and further analysis of the material. The updated texts were then automatically linguistically annotated according to the Universal Dependencies Formalism for Slovenian. The TEI-encoded versions were converted into various formats and published under an open licence in the CLARIN.SI repository.

The second part of the paper presents an example of corpus analysis of the author's narrative style, using NoSketch Engine and Sketch Engine, based on four frequency variables: the most frequent 100 words, the least frequent words, the most frequent word-2-grams, and keywords. In addition to the expected lexis, the analysis revealed the frequent use of inflectional verbs (especially *morati*), noun-adjective phrases in which the adjective expresses a great or superlative quality of the noun (*velik*, *velikanski*, *presladek*, *presvet*; *večen*, *nepopisljiv*, *neskončen*, *izreden* etc.), and adjectives with diametrically opposed meanings (*svet*, *pobožen – posveten*, *grešen*), which in the context of religious discourse take on the symbolic meaning of right/good and wrong/bad examples of life, views and decisions, and thus have an educational function. The long sentence structure with subordinating conjunctions (e.g. *kateri*) has also emerged as a characteristic feature of Repič's style. The study has demonstrated how corpus material provided with bibliographic metadata can be used for comparative synchronic-diachronic studies of Škrabec's linguistic norm (e.g. word type *bralec/bravec*).

1.01 DOI: https://doi.org/10.51663/pnz.65.3.04

Katja Meden,* Ana Cvek,♦ Vid Klopčič,° Mihael Ojsteršek,• Matevž Pesek,♠ Mojca Šorn,▾ Andrej Pančur◊

Unlocking History: A Redesign and Content Analysis of the SIstory 5.0 Portal

IZVLEČEK

ODPIRANJE ZGODOVINE: PRENOVA IN ANALIZA VSEBINE PORTALA SISTORY 5.0

Portal Zgodovina Slovenije – SIstory.si predstavlja pomembno interdisciplinarno zbirko publikacij, podatkov, zbirk in metapodatkov, predvsem na področju zgodovinopisja. Zbirka zajema širok spekter zgodovinskih publikacij ter metapodatke, ki jih opisujejo. Nedavna prenova portala SIstory je bila osredotočena na prizadevanja, da podatkov ne bi ponudili le kot

* **Research Assistant, Institute of Contemporary History, Privoz 11, SI-1000, Ljubljana; PhD student, Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, SI-1000, Ljubljana; Jozef Stefan International Postgraduate School, Jamova cesta 39, SI-1000 Ljubljana, katja.meden@inz.si; ORCID: 0000-0002-0464-9240**

♦ **Assistant, Institute of Contemporary History, Privoz 11, SI-1000, Ljubljana, ana.cvek@inz.si; ORCID: 0009-0002-7927-3783**

° **Expert Associate, University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000, Ljubljana, vid.klopcic@fri.uni-lj.si**

• **Assistant, Institute of Contemporary History, Privoz 11, SI-1000, Ljubljana, mihael.ojstersek@inz.si; ORCID: 0009-0007-7233-2601**

♠ **PhD, Assistant Professor, University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000, Ljubljana, matevz.pesek@fri.uni-lj.si; ORCID: 0000-0001-9101-0471**

▾ **PhD, Research Fellow, Institute of Contemporary History, Privoz 11, SI-1000, Ljubljana, mojca.sorn@inz.si; ORCID: 0000-0002-4457-1118**

◊ **PhD, Research Fellow, Institute of Contemporary History, Privoz 11, SI-1000, Ljubljana, andrej.pancur@inz.si; ORCID: 0000-0001-6143-6877**

zbirke zgodovinskih publikacij, temveč bi omogočili tudi večjo preglednost, interoperabilnost in dostopnost raziskovalnih podatkov širšemu občinstvu, tako raziskovalcem kot splošni javnosti. Prispevek predstavlja proces prenove portala in njegove tehnične izboljšave ter poglobljeno analizo vsebin, ki jih portal ponuja v sedanji obliki.

Ključne besede: SIstory, prenova, podatkovni sistemi, metapodatki, zgodovinopisje

ABSTRACT

The portal History of Slovenia - SIstory.si is an interdisciplinary collection of historical publications, data, collections and metadata that has been operating since 2008. The portal encompasses a diverse range of historical information, including publications, images, extensive databases, and comprehensive metadata that describe the objects. The recent redesign of the SIstory portal has focused on ongoing efforts to offer the data not only as a collection of historical publications but also to enable greater transparency, interoperability, and availability of research data to a broader audience. This paper examines the portal's redesign, focusing on technical improvements, and then provides an in-depth analysis of its content.

Keywords: SIstory, redesign, information systems, metadata, history

Introduction

The Research Infrastructure of Slovenian Historiography (RI INZ) was established in September 2006. While the foundations were laid at that time, its primary aim – the digitisation and online publication of frequently used Slovenian historical content – was only defined and developed in the following years. An important aspect of the infrastructure's early development was the popularisation and promotion of historical–scientific research among the general public and the research community, which was to be achieved through a digital portal or a similar application. The online research and educational portal History of Slovenia – SIstory was launched in September 2008. A test version of the portal was presented to the Institute's researchers at the beginning of 2008, allowing them to test the portal's functionality, logical sequence of functions, content hierarchy, links, and the various search methods.[1]

The main content of the portal at the time consisted of a combination of historical literature, historical sources and technical infrastructure services. Its main goal was

1 Institute of Contemporary History, *Poročilo o doseženih ciljih in rezultatih v letu 2008* (Ljubljana: Institute of Contemporary History, 2009), 6, 31, 32, accessed on 26 February 2025, https://inz.si/wp-content/uploads/2025/06/2008.pdf.

to provide digitised and freely accessible research results and sources. Emphasis was placed on preserving older, less easily accessible sources, thereby safeguarding cultural and scientific heritage.[2] One of the first significant projects to populate the portal was the comprehensive digitisation of the entire edition of the scientific journal *Prispevki za novejšo zgodovino* (*Contributions to Contemporary History*), which includes issues from 1957 to the present day. In the following years, efforts expanded beyond internal production to include numerous Slovenian and international editorial offices, institutions, and individual collectors, in order to obtain verified historiographical works and materials. Alongside content acquisition, the necessary copyright permissions for publication were systematically obtained and recorded.[3]

The operation and design of the SIstory portal originally focused on supporting the research processes of the Institute's research community members. Following a highly successful initial response from the public and related institutions, SIstory gradually expanded beyond the boundaries of "written history," incorporating interactive presentations of historical content supported by emerging technologies.[4] With data preservation and research community integration at the core of its development, SIstory became the primary output of the newly established Slovenian national node of the DARIAH ERIC research infrastructure (DARIAH-SI),[5] with the Institute (specifically RI INZ) serving as the national coordinating institution.[6]

The redesign presented in this paper focuses on technological advances to improve the accessibility of the data. Furthermore, the data available on the portal has not yet been fully explored; therefore, part of our work consists of a content analysis. Finally, we are focusing our efforts on the transparency, reuse, and accessibility of data, as well as user-friendly interaction with the portal. The rest of the paper is structured as follows: the opening section outlines the history and placement of the portal within the field of digital humanities, along with a brief overview of the portal's technical development. The next section presents the redesign process, providing an overview of the upgrade components (technical foundations and metadata), as well as basic content statistics. The following section discusses the content analysis, highlighting notable trends and their significance for the portal. Finally, the last section provides an overview of the paper and presents some options for future work.

2 Mojca Šorn, Andrej Pančur, and Mitja Sunčič, "SIstory: arhivsko gradivo in e–humanistika," *Arhivi* 34, No. 1 (2011): 145.

3 Mojca Šorn and Katja Meden, "Portal Zgodovina Slovenije – SIstory in avtorske pravice," *Prispevki za novejšo zgodovino* 61, No. 2 (2021): 193–228.

4 Šorn et al., "SIstory," 145.

5 *Dariah-SI*, accessed on 26 February 2025, http://www.dariah.si/.

6 Andrej Pančur and Mojca Šorn, "Na začetku je bil SIstory: Raziskovalna infrastruktura slovenskega zgodovinopisja," in *Inštitut za novejšo zgodovino: 60 let mislimo preteklost* (Ljubljana, Institute of Contemporary History, 2019), 47–58, https://hdl.handle.net/11686/46230.

History of the SIstory Portal

The landscape of digital humanities in Slovenia during SIstory's early development differed significantly from its current state. At the time, the field was still in its early stages, with limited infrastructure and awareness, and only beginning to establish itself within academic and research institutions. Hadalin[7] details the state of digital humanities in Slovenia at that period, highlighting key institutions (such as ARNES and the Jožef Stefan Institute), and researchers who, despite general scepticism, advocated for the integration of this emerging field into mainstream science and curricula. Additionally, the author outlines essential services and research infrastructures such as DARIAH-SI and CLARIN.SI, which remain fundamental for open research data preservation and are vital to the field's growth. He also emphasises the SIstory portal as the "central hub for digital history" and outlines its importance as a collection of materials relevant not only to the research community of Slovenian history but also to wider research communities.[8]

It is important to note that while the core features of SIstory, particularly its primary purpose of digitising historical sources,[9] align with both traditional research data repositories and digital libraries, SIstory was never intended to be either. As evidenced by various sources from the time of its initial launch,[10] the portal was always meant to be co-created with users and the research community. Its modular design was specifically developed to facilitate direct engagement with users. This is further supported by the diversity and volume of materials, databases, and interactive technologies that have been, and in many cases remain, integral to the portal's ongoing development:

- **ZgoLj (Zgodovina Ljubljane – History of Ljubljana):** ZgoLj was a mobile application developed as part of SIstory, which supported augmented reality and enabled a virtual tour of Ljubljana's historical centre, based on old photographs, provided by the Historical Archives of Ljubljana.[11]
- **Interactive exhibitions:** Utilising similar technologies, SIstory hosted interactive exhibitions, such as "Slovenians and the First World War 1914–1918".[12] Due to the

7 Jurij Hadalin, "The Slovenian Digital Humanities Landscape? A Brief Overview," in Torsten Kahlert and Claudia Prinz, eds., *The Status Quo of Digital Humanities* (Berlin: H-Soz–Kult, 2015), 154–69, accessed on 26 February 2025, https://edoc.hu-berlin.de/server/api/core/bitstreams/febdd0e3-5c2b-4eb8-8b30-ea8f1b36cae7/content.

8 Ibidem, 165. In his paper, the author emphasises the relevance of materials to the broader community of the common Yugoslav and Austro-Hungarian state.

9 Mojca Šorn and Ana Cvek, *Vsebine in njihovo razporejanje na portalu Zgodovina Slovenije - SIstory (s poudarkom na publikacijah)* (Ljubljana: Institute of Contemporary History, 2023), accessed on 26 February 2025, https://sistory.github.io/Vsebine_SIstory/index.html.

10 STA, "SIstory – spletni portal slovenskega zgodovinopisja," *Siol.net*, accessed on 26 February 2025, https://siol.net/novice/novice/sistory-spletni-portal-slovenskega-zgodovinopisja-336634.

11 Hadalin, "The Slovenian Digital Humanities Landscape?." Andrej Pančur, *SIstory augmented reality 1.0 XML Schema, Documentation* (Ljubljana: Institute of Contemporary History, 2013), accessed on 26 February 2025, https://hdl.handle.net/11686/20385. Andrej Pančur, *SIstory nadgrajena resničnost 1.0 XML shema* (Ljubljana: Institute of Contemporary History, 2013), accessed on 26 February 2025, https://hdl.handle.net/11686/20369.

12 Marko Štepec and Mojca Turk, *Slovenians and the First World War*, 2011, accessed on 26 February 2025, https://hdl.handle.net/11686/1160.

obsolescence of the technologies used to create these exhibitions, such examples only contain their metadata records.

- **History Citation Index (HIS)**: Almost simultaneously with the development of SIstory, another important database – History Citation Index (*Zgodovinarski indeks citiranja* or ZIC) – was in development. The primary purpose of ZIC was to create a database of citations of Slovenian humanities production, filling the gap between the citation databases[13] recognised by the Slovenian Research Agency (ARIS, then ARRS) and humanities publishing practices.[14] The database was later also recognised by ARIS and remains regularly updated to this day.[15]
- **Database of WWI victims:** A national, freely accessible database of deceased individuals from the area within the borders of the Republic of Slovenia, resulting from the long-standing project *Zbiranje podatkov o vojaških* žrtvah *1. svetovne vojne na Slovenskem* (2015–2018). The time period encompasses data from the war period, but also includes deaths resulting from the war's effects after 1918.[16]
- **Database of WWII victims**: The database results from research conducted between 1997 and 2012 as part of four major research projects. It is a systematic record of military and civilian persons who had the right of residence in the present-day Republic of Slovenia during the Second World War and the immediate post-war period (May 1940 – January 1946) and lost their lives due to wartime and (revolutionary) post-war violence or the consequences of war. In 2025, both the WWI and WWII databases were redesigned to provide not only unrestricted access to previously limited data but also to enable public participation. The updated version now allows users to contribute additional information, comments and personal narratives within designated layers, promoting a more comprehensive and collaborative approach to historical documentation.[17]
- **Population censuses** – A database of digitised population census questionnaires for Ljubljana. A population census was carried out for the first time in 1830, based on an imperial patent from 1804 and new instructions issued in 1829.

13 Citation databases such as Web of Science and Scopus tend to favour research articles as the main form of publication in most academic fields. However, in the humanities, the primary form of scholarly publishing is the scientific monograph, which is often excluded from citation indexes or more difficult to track.

14 Katja Meden and Ana Cvek, "Nadgradnja zgodovinarskega indeksa citiranosti," *Slovenščina 2.0* 9, No. 1 (2021): 216–35.

15 Ibid. *ZIC – Zgodovinski indeks citiranosti*, accessed on 26 February 2025, https://zic.sistory.si/. Hadalin, "The Slovenian Digital Humanities Landscape?." *SIstory.si – Culture of Slovenia*, accessed on 26 February 2025, https://www.culture.si/en/SIstory.si.

16 Andrej Pančur, Neja Blaj Hribar, Mojca Šorn, and Mihael Ojsteršek, "Projekt Vojaške žrtve prve svetovne vojne na Slovenskem," in Darja Fišer and Tomaž Erjavec, eds., *Proceedings of the Conference on Language Technologies and Digital Humanities: September 24th–25th 2020, Ljubljana, Slovenia* (Ljubljana: Institute of Contemporary History, 2020), 136–40, https://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Pancur-et-al_Projekt-Vojaske-zrtve-prve-svetovne-vojne-na-Slovenskem.pdf. *Vojaške žrtve 1. svetovne vojne na Slovenskem* (Ljubljana: Institute of Contemporary History), *Zgodovina Slovenije – SIstory*, https://www.sistory.si/ww1, 13 November 2018.

17 Internal database of the Institute of Contemporary History (INZ): Tadeja Tominšek Čehulić, Mojca Šorn, Marta Rendla, Dunja Dobaja, Tamara Logar: Smrtne žrtve med prebivalstvom na območju Republike Slovenije med drugo svetovno vojno in neposredno po njej [Database].

The statistical questionnaires for the town of Ljubljana (*Conscriptions Aufnahms Bogen*) have been fully preserved and are now organised into eleven census units, according to Ljubljana's cadastral municipalities and house numbers.[18]

While the development and release of the abovementioned databases (ZIC, Popisi, WWI & WWII) were all carried out under the SIstory banner, they remained standalone until recently,[19] complementing the core of the SIstory portal – the publications. In this paper, we focus on the development history, update, and subsequent analysis of this central component.

Development history of the SIstory portal

The SIstory portal has a relatively long history of development. Since its initial release in 2008, several versions of the portal have been released as individual upgrades. In 2011, the first software and technological upgrade was carried out, establishing the latest standards and enabling faster and more stable system operation. This upgrade also played an essential role in establishing a national digital infrastructure for the humanities and arts[20]. The first upgrade consisted of several components:

- Content administration in Apache SOLR[21] and upgrading folder structures and file names.
- Implementation of the Dublin Core metadata standard (DC)[22]. The original schema contained all 15 basic DC elements. A year later, the original schema was upgraded with elements from the qualified DCMI Metadata Terms (DCTERMS[23]).
- Creation of a unique and permanent URN – Uniform Resource Name.
- Introduction of the Sphinx[24] metadata search engine. Two search engines were implemented: a basic and an advanced one.
- Upgrading of the portal administration.
- Design of the structure and access levels for users.

18 Andrej Pančur, "Popisi prebivalstva Slovenije 1830–1931: Orodje za transkribiranje historičnih demografskih podatkov," in Tomaž Erjavec and Darja Fišer, eds., *Zbornik Konference Jezikovne Tehnologije in Digitalna Humanistika, 29. September–1. Oktober 2016, Filozofska Fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija = Proceedings of the Conference on Language Technologies & Digital Humanities, September 29th–October 1st, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia* (Ljubljana: Ljubljana University Press, Faculty of Arts, 2016), 133–41, http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Pancur_Popisi-prebivalstva-Slovenije-1830-1931.pdf.

19 In 2024, the WWI and WWII databases were updated and directly connected to SIstory.

20 Ana Cvek, Mihael Ojsteršek, and Mojca Šorn, *Izhodišča metapodatkovnih sistemov portala Zgodovina Slovenije – SIstory* (2008–2016) (Ljubljana: Institute of Contemporary History, 2022), accessed on 26 February 2025, https://sidih.github.io/izhodisca/index.html.

21 *Overview of the Solr Admin UI | Apache Solr Reference Guide 6.6*, https://solr.apache.org/guide/6_6/overview-of-the-solr-admin-ui.html.

22 *DCMI: Dublin Core™ Metadata Element Set, Version 1.1: Reference Description*, https://www.dublincore.org/specifications/dublin-core/dces/.

23 *DCMI: DCMI Metadata Terms*, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

24 *Sphinx | Open Source Search Engine*, https://sphinxsearch.com.

In 2013, the portal was updated to introduce the SIstory metadata schema, a customised set of metadata elements developed to better reflect the nature of the content, which had outgrown the descriptive capabilities of the DCTERMS element set. A major update followed in 2016, during which a mapping between the SIstory schema and Dublin Core was established to enhance data interoperability. The schema was then expanded with elements and structures from the HOPE application profile,[25] a recognised standard in the GLAM community, resulting in the SIstory application profile.[26] This profile has since served as the portal's primary metadata standard, shaping the structure, syntax, and semantics of the metadata input tool. Besides the metadata enrichments, new system frameworks and graphical designs for both the administration and user interfaces were also installed. In addition, the search engine (filtering and sorting of results; full-text search) was also taken into account in the updates. Overall, since the portal's inception in 2008, a series of upgrades have been made, each improving the portal's functionalities and features. The entire development history is documented in Cvek et al.[27] This brings us to today and to new steps in the portal's development – the decision to redesign the portal from the ground up.

SIstory: The Redesign

As the portal was updated multiple times, the code became too extensive to manage efficiently. Moreover, the concepts and various solutions developed over the years were very ambitious and deemed necessary at the time. However, they did not prove to be as useful in daily practical operations as initially thought. Additionally, the outdated appearance of the user interface was a decisive factor in our choice to rebuild from scratch. When planning the redesign, we considered the legacy issues and solutions from previous versions of the portal to enhance system functionality and create a familiar user experience.

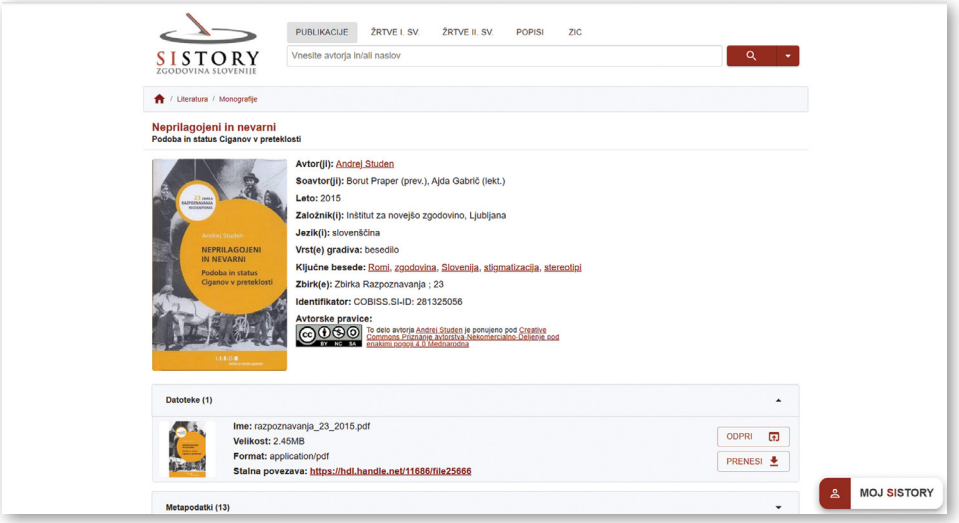
The redesign consisted of several sections, from purely technical aspects (such as the code base and integration of the OAI-PMH protocol) to simplifying the metadata schema, refining the user interface, and restructuring the content classification.

25 Bert Lemmens, Joris Janssens, Ruth V. Dyck, Alessia Bardi, Paolo Manghi, Eric Beving, Kathryn Máthé, Katalin Dobó, and Armin Straube, *Hope – The Common HOPE Metadata Structure, Including the Harmonisation Specifications (D2.2)* (Tech. Rep., HOPE, Deliverable D2.2, 2011).

26 Andrej Pančur, *Metapodatki portala Zgodovina Slovenije-SIstory* (Tech. Rep.) (Ljubljana: Institute of Contemporary HIstory, 2013). Cvek et al., Izhodišča metapodatkovnih sistemov. Katja Meden, "'Posmrtno življenje posmrtnih mask': sodelovanje Raziskovalne infrastrukture slovenskega zgodovinopisja Inštituta za novejšo zgodovino z Društvom za domače raziskave", in *Odlivanje smrti: posmrtne maske v slovenskih javnih zbirkah* (Ljubljana: Institute of Contemporary History, 2023), accessed on 26 February 2025, https://sistory.github.io/Odlivanje_smrti/ch02.html.

27 Cvek et al., *Izhodišča metapodatkovnih sistemov.*

Figure 1: The new SIstory user interface



Source: Own work

Technical design

In terms of technical composition, the redesigned SIstory 5.0 portal is based on a robust technical framework, while the backend utilises Django for efficient data management and content delivery. On the frontend, SIstory employs Next.js and React in combination with Node.js for dynamic and interactive user interfaces. This modern frontend stack enables smooth navigation and responsive design across various devices, improving accessibility for users accessing historical content. Figure 1 shows an example of the redesigned user interface. The portal's database architecture is based on PostgreSQL and provides a robust foundation for storing and retrieving large volumes of historical data with high speed and reliability. In addition, SIstory integrates Matomo, an analysis function that enables administrators to gain valuable insights into user behaviour and interaction patterns, thus forming the basis for future developments and improvements.

For efficient search functionality, SIstory incorporates Elasticsearch and Kibana, allowing users to quickly locate relevant historical documents and sources. The use of Elasticsearch ensures fast and accurate search results, improving the overall usability of the portal. In addition, SIstory employs a Handle[28] system to provide permanent identifiers, enabling reliable and permanent access to specific historical documents and sources. This allows users to reference and cite materials consistently, contributing to the scholarly integrity and reliability of the portal. Overall, SIstory's technical specifications underline the portal's commitment to providing a robust and user-friendly platform for accessing Slovenia's rich historical heritage.

28 *Handle.Net Registry,* https://www.handle.net/.

Metadata design

The portal previously based its metadata element set on the SIstory application profile (SIstory AP), which was derived from the HOPE application profile. In practice, this posed a problem as SIstory AP contained several elements (and element groups) that were not used as frequently as initially assumed. This, in turn, led to a simplification of the profile. To this end, an analysis of the existing SIstory AP was conducted to identify the metadata elements that should be retained and address those that present legacy issues.

The current state of the metadata application profile comprises 26 elements (reduced from the original 33 elements), with a focus on the DC and DCTERMS metadata elements and only a few additional elements from the previously mentioned HOPE AP. One of the main reasons for this shift in focus is to improve the interoperability of our metadata. Only a limited number of elements of the "*SIstory*" namespace have been retained[29], primarily due to the remnants of older publications described with these specific metadata elements. The overview of the major metadata groups is presented in Table 1.

Table 1: Overview of the most important metadata groups, the number of unique instances, and the total number of occurrences on the SIstory portal (at the time of writing this paper)

Metadata	Unique Values	Nr. of Usages
No. of entries	62,656	
Creator	5,319	25,461
Subject	24,236	333,667
Publisher	1,091	57,686
Collection	432	1,895
Contributor	1,322	44,088
Type	12	63,362
Language	61	74,853

Source: Own work

In total, SIstory comprises over 60,000 unique entries and more than 5,000 unique authors/physical persons (under the category "*Creator*")[30], while "*Subject*" contains keywords that describe the publications. Secondary forms of authorship are described in the category "*Contributor*" (e.g. editor, translator...), while the type of publication

29 For example, *SIstory Unstored* – a field for storing metadata that do not fit into any other metadata field due to their content.

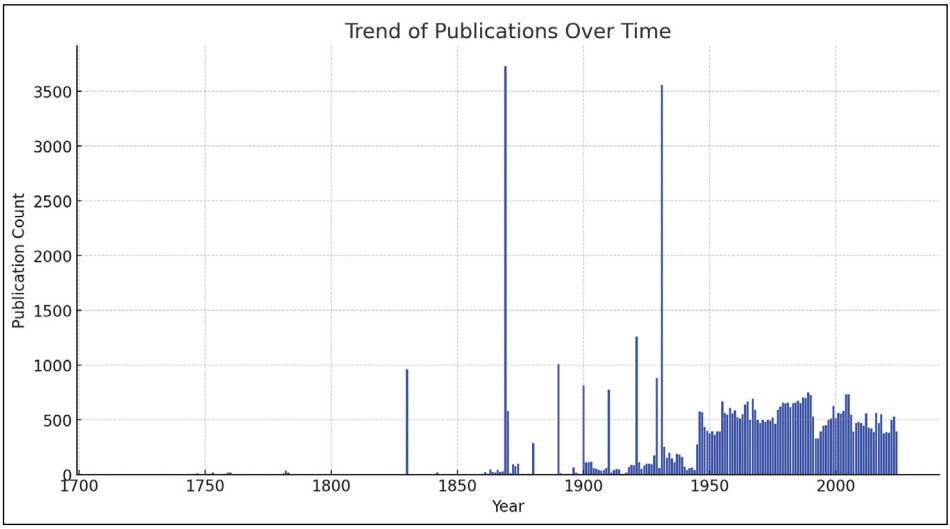
30 The metadata mask includes two separate fields for the Creator, which, according to the Dublin Core definition, can be either a physical person or a legal "organisation." In Table 1, however, only occurrences of a physical person/author are counted under the "Creator" category.

based on the controlled vocabulary (DCMI Type[31]) encompasses 12 categories. Finally, the portal includes publications in 61 different languages, which are presented in more detail in the Language trends subchapter.

SIstory Unveiled: Content Analysis

In the effort to present the redesign of the SIstory portal, it became clear that focusing mainly on technical and aesthetic improvements would not fully capture the essence of the portal – its content, or rather, its historical sources. Therefore, we expanded the scope of the work to include a comprehensive content analysis, aimed at gaining a deeper understanding of the content available on the portal.

Figure 2: Trends of publications over time



Source: Own work

The analysis focused on various aspects of the portal's content, ranging from basic statistics of the main metadata groups (excluding individual metadata elements) to a more in-depth analysis of:

- **publication date** – when a work was first published;
- **publication keywords** – e.g., *Jugoslavija, učbeniki*;
- **language of the publication** – classified according to the ISO 639-2 standard;
- **author networks** – mapping connections between contributors.

This, in turn, allows us to demonstrate not only the scope of available content but also to highlight the different types and variations of this content.

31 *DCMI: DCMI Metadata Terms*, https://www.dublincore.org/specifications/dublin-core/dcm-terms/#DCMIType.

Publication trends

One of the trends analysed in this section is the distribution of publications over time according to their publication date, to determine which years are best represented in terms of published content. The results are shown in Figure 2.

One of the first trends to emerge is the distribution of publications between approximately 1715 and the mid-20th century, which exhibits several significant spikes in the number of publications. Conversely, the post-World War II period displays a much more consistent flow. The spikes in the timeline are most likely caused by the publication of several large works or specific types of publications (e.g., textbooks, population censuses) on the portal. In contrast, the steady publication flow from the post-war period to the present day suggests a greater variety of publication types (e.g., literature, research and studies, monographs, etc.) and the absence of very large volumes of similar publications. The nature and content of these publication trends are examined in more detail in the keyword analysis, which provides additional insight and substantiates the reasons for the identified trends.

Keyword analysis

The keyword analysis of the portal content examined the ten most frequently used keywords to describe sources in the individual menus for each of the 20 years (except the period 2010–2024). This analysis was conducted for the first-level menus: *Viri* (Sources), *Literatura* (Literature), *Dogodki* (Events), *Podatki* (Data), and *DH* (Digital Humanities). The results are presented in the following sections[32].

Sources – top ten keywords

Table 2 gives an overview of the ten most frequent keywords for each twenty-year period in the "Sources" menu. This category encompasses various types of sources, including archival, oral, and printed sources, as well as digitised versions of physical objects. The latter are mainly images of physical objects, such as statues or death masks, and printed sources.

In the 18th century (1710–1790), several recurring keywords – such as *patenti* (patents), *odloki* (decrees), *norme* (norms), and *Marija Terezija* – point to one of the larger collections on the portal: *Collection of various patents, decrees, ordinances, norms, instructions, etc., issued by Charles VI, Maria Theresa, and Joseph II,*[33] acquired through collaboration with the Central Judicial Library. Additional keywords, such as "*popisi*

32 Some keywords are very similar to one another, mostly due to slight variations in the notation format. For example, *"uradni list"* and *"uradni listi"* are the singular and plural forms of the same keyword but are counted separately.

33 An example of a Josef II directed patent: https://hdl.handle.net/11686/31413.

prebivalstva" (population censuses) and "*občina"* (municipality), highlight a significant number of censuses from this period.

A similar trend can be observed in the 19th century, with census records accompanied by theatre lists from various provincial theatres (e.g., the Provincial Theatre in Ljubljana), and publications like *koledar* (the calendar of the Society of St. Mohor – an annual publication containing a calendar, religious prayers, illustrations, poetry, etc.).[34] Keywords such as *Kranjska*, *Carniola*, and *deželna avtonomija* (provincial autonomy) further indicate the prominence of minutes from the Carniolan Provincial Assembly.

The publications from the early 20th century, most frequently uploaded to the portal, include censuses (marked by keywords like *popisi prebivalstva* and *občina*), official gazettes (*uradni listi*) from various states and periods[35] (e.g., Slovenia, Yugoslavia, Serbia, Bosnia and Herzegovina), and stenographic records (*stenografski zapisniki*) from legislative and executive bodies – all of which became increasingly prevalent during this period.

Table 2: Top ten keywords by two-decade period for Viri (Sources)

Decade Range	Top Keywords
1710-1729	Karl VI., Patenti, odloki, predpisi, norme, navodila, okrožnice
1730-1749	Patenti, odloki, predpisi, norme, navodila, okrožnice, Marija Terezija, Karl VI., Karl VI., Corbinian Graf von Saurau, Marija Terezija, Anton Barbo Waxenstein, Marija Terezija, Anton Josef Auersperg, Marija Terezija, Anton Josepf Graf von Auersperg, Marija Terezija, Corbinian Graf von Saurau, Marija Terezija, Fridrich Wilhelm Graf von Haugwitz
1750-1769	Patenti, odloki, predpisi, norme, navodila, okrožnice, Marija Terezija, Karl VI., Marija Terezija, Marija Terezija, Anton Josepf Graf von Auersperg, Marija Terezija, Anton Joseph von Auersperg, Marija Terezija, Ludvik XVI., Marija Terezija I.
1790-1809	celjski grofje, drame, leposlovje, Celje, Ljubljana, hišne številke, rodbine, rokopisi, živinozdravniški recepti
1810-1829	Ljubljana, hišne številke, Ludvig van Beethoven, popis
1830-1849	Ljubljana, Slovenija, 1830-1857, popisi prebivalstva, programi, gledališča, 19. stoletje, gledališki listi, gledališče, Avstrija
1850-1869	Slovenija, 1869, popisi prebivalstva, Ljubljana, občina Dobrnič, občina
	Trebnje, občina Prečna, občina Mirna, občina Velika Loka, občina
	Črmošnjice

34 An example of the St Mohor calendar: https://hdl.handle.net/11686/27099.

35 Examples of official gazettes: https://sistory.si/menu/1/7/69.

Decade Range	Top Keywords
1870-1889	Slovenija, popisi prebivalstva, občina Vrhnika, 1870, 1880, Kranjska, deželna avtonomija, provincial autonomy, Carniola, koledar
1890-1909	Slovenija, popisi prebivalstva, občina Vrhnika, 1890, 1900, Družba sv. Mohorja, koledar, Avstro-Ogrska, popis prebivalstva, upravna razdelitev
1910-1929	Slovenija, popisi prebivalstva, Ljubljana, 1921, šolski listi, 1910, občina Vrhnika, 1929, Komunistična partija Jugoslavije, delavsko gibanje
1930-1949	Slovenija, Ljubljana, popisi prebivalstva, 1931, Jugoslavija, uradni listi, Srbija, BiH, Bosna in Hercegovina, uradni list
1950-1969	uradni listi, Jugoslavija, Ljubljana, BiH, Bosna in Hercegovina, Kosovo, Vojvodina, stenografski zapisniki, Socialistična republika Slovenija, družbeno samoupravljanje
1970-1989	Jugoslavija, uradni listi, stenografski zapisniki, predstavniška telesa, družbeno samoupravljanje, Socialistična republika Slovenija, Kosovo, BiH, Bosna in Hercegovina, Vojvodina
1990-2009	Slovenija, parlament, zakonodaja, državni zbor, Jugoslavija, uradni listi, skupščina, BiH, Bosna in Hercegovina, Vojvodina
2010-2024	popisi prebivalstva, Ljubljana, analiza, 1921, zgodovina, krajevna imena, 1900, krajevni leksikoni, toponimi, privilegiji

Source: Own work

Finally, for the more recent period (2010–2024), the keywords primarily refer to studies conducted in connection with the censuses of Slovenia from 1830 to 1931, which resulted from cooperation with the Historical Archive of Ljubljana.

Literature – top ten keywords

Table 3: Top ten keywords by two-decade period for *Literatura* (Literature)

Decade Range	Top Keywords
1810-1829	učbeniki, 19.st., abecedniki, slovenska književnost, slovensko-nemški abecednik, učbenik, učbeniki za osnovne šole, verouk
1830-1849	učbeniki, 19.st., izobraževanje, katekizem, katoliška vera, matematika, verouk
1850-1869	finance, Avstrijsko cesarstvo, učbeniki, slovnica, banke, valute, finančno vprašanje, slovenščina, valuta, nemščina
1870-1889	učbeniki, nemščina, matematika, politične stranke, organizacije in društva, čitanke, zgodovina, učbeniki za osnovne šole, berila, Kranjska, učbeniki za srednje šole

Decade Range	Top Keywords
1890-1909	politične stranke, organizacije in društva, avstrijska doba, politični programi, Književna poročila, učbeniki, katoliški tabor, liberalni tabor, Narodopisne razprave in Mala izvestja, Mala izvestja, matematika
1910-1929	Slovstvo, politične stranke, organizacije in društva, politični programi, Izvestja, avstrijska doba, Razprave, učbeniki, liberalni tabor, katoliški tabor, zgodovina
1930-1949	Slovstvo, Razprave, Izvestja, zgodovina, učbeniki, Pregled, Zapiski, učbeniki za srednje šole, geografija, Jugoslavija
1950-1969	ocene in poročila, druga svetovna vojna, Slovenija, zgodovina, NOB, Ljubljana, zgodovinski pregledi, arheologija, Slovenci, Jugoslavija
1970-1989	ocene in poročila, druga svetovna vojna, Slovenija, arhivsko gradivo, arhivi, poročila, NOB, srednji vek, Jugoslavija, zgodovina
1990-2009	ocene in poročila, Slovenija, arhivi, druga svetovna vojna, zgodovina, arhivsko gradivo, Slovenci, arhivistika, biografije, Jugoslavija
2010-2024	ocene in poročila, Slovenija, zgodovina, Jugoslavija, druga svetovna vojna, socializem, Ljubljana, prva svetovna vojna, vojaška zgodovina, ocene

Source: Own work

Similarly, Table 3 shows the ten most frequent keywords over a single 20-year period for the Literature menu, which consists of publications such as research monographs, (Slovenian) serial history publications – along with the in-house produced scientific journal *Prispevki za novejšo zgodovino* (Contributions to Contemporary History) – school and university theses, and collections of digital monographs.

The 19th century is primarily characterised by the textbooks produced as part of the projects "Šolski *listi*" and "*Schools and Imperial, National, and Transnational Identifications: Habsburg Empire, Yugoslavia, and Slovenia*".[36] These represent an extensive digitisation project of textbooks, mainly intended for schools, covering various school subjects, and identified in the table with the following keywords: *učbeniki* (textbooks), *abecedniki* (abecedarium), *matematika* (mathematics), čitanke, and *berila* (reading materials), etc. For the early to mid-20th century, however, the topics are then expanded to include additional materials on politics, political programmes, and political parties, as indicated by the keywords *politični programi* (political programmes), *katoliški tabor* (Catholic camp), and *liberalni tabor* (Liberal camp). For the second half of the 20th century, the themes shift towards World War II, more precisely to the role of Yugoslavia (and Slovenia) in World War II (keywords). Directly related to this is also a considerable amount of literature referring to archival sources (*arhivsko gradivo*), mostly in connection with a specific journal, *The Gazette of the Archival Association and*

36 *Schools and Imperial, National, and Transnational Identifications: Habsburg Empire, Yugoslavia, and Slovenia* | *Faculty of Arts, University of Ljubljana*, accessed on 26 February, https://www.ff.uni-lj.si/en/raziskovanje/sole-in-identifikacije.

Archives of Slovenia. Lastly, a very prominent keyword, *ocene in poročila* (reviews and reports), refers to a particular form of contributions to various Slovenian (scientific) journals, where authors provide reviews of various published works on topics covered by the journal (mainly history in this case).

Events – top ten keywords

While text documents are the predominant type of publication within the SIstory portal, RI INZ also offers in-house production and recording of various events, as well as digitisation of various exhibitions related to the field of history, the Institute, or related institutions.

The first key difference between Tables 2 and 3 is the significantly shorter time span, which is expected given that the portal has only existed since 2008. This also explains the limited and unrepresentative keyword coverage for the 1990–2009 period, which includes only two publications – both focused on girls' education in Ljubljana[37] and Slovenian students abroad[38]. However, the number of publications increased in the period 2010– 2019. The most common keywords, such as *Filozofska fakulteta* (Faculty of Arts), *Oddelek za zgodovino* (Department of History), *zgodovina* (history) and *Slovenija* or *Jugoslavija*, refer to the institutions, organisations and general topics that organised the events (mostly recorded lectures).

Data and Digital Humanities – top ten keywords

In contrast to the keyword analysis of sources and literature, which covers several centuries, the two following publication types, *Podatki* (Data) and *DH* (Digital Humanities data), are limited to the last decade (2010–2024). In both cases, the number of publications is relatively small, so these keywords are more representative of individual sources rather than the entire portal.

Table 4: Top ten keywords by two-decade period for *Podatki* (Data)

Decade Range	Top Keywords
2010-2024	1910, Dravska banovina, Judje, Slovenije, krajevna imena, popisi prebivalstva

Source: Own work

Table 4 mainly focuses on research data in the history field, specifically the data on old place names in Slovenia and censuses of the Jewish population in Slovenia.[39]

37 Šola naših babic: izobraževanje deklet v Ljubljani, https://hdl.handle.net/11686/37914.
38 Študenti s Kranjske na avstrijskih in nemških univerzah 1365–1917, https://hdl.handle.net/11686/31001.
39 A list of Jews in Slovenia (Dravska banovina), 1937, https://hdl.handle.net/11686/11136.

While the categories examined are generally research data, the Digital Humanities category mainly reflects the vision of the RI INZ at the time – expanding into digital humanities and providing data and tools to support research activities in these (related) fields. This eventually resulted in the creation of a separate repository for digital humanities, the SI-DIH repository, another product of RI INZ and DARIAH-SI.

Table 5: Top ten keywords by two-decade period for *DH*

Decade Range	Top Keywords
2010-2024	nadgrajena resničnost, XML shema, SIstory augmented reality XML, metapodatki, DOCX, HTML publikacija, SIstory, SIstory nadgrajena resničnost XML shema, TEI, administracija

Source: Own work

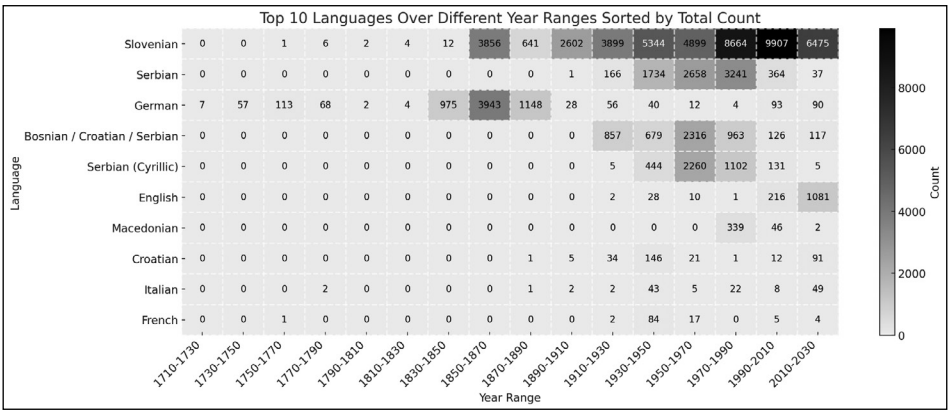
As these were the infrastructure's initial steps towards DH, there are only limited publications and tools available, but they incorporate the technologies of that period. This is also reflected in the keywords in Table 5, such as *nadgrajena resničnost* (augmented reality), *XML shema* (XML schema), *metapodatki* (metadata), *HTML*, and *TEI*.

Language trends

In addition to keyword analysis, we also examined the languages of the publications within the SIstory portal, as shown in Figure 3.

It is not surprising that the most frequent language of publications on the portal is Slovenian, with a total of 46312 occurrences, mainly in the period from 1970 to 2025, especially between 1990 and 2025. The second most common language is Serbian (8201 occurrences), although an explicit distinction must be made here, as Serbian also falls into two other language categories: Serbian (Cyrillic) for publications in Cyrillic script (3947 in total) and the Bosnian/Croatian/Serbian category for publications where the language could not be explicitly identified (mostly publications referring to the Yugoslav *Official Gazettes*), which are the most frequent among the publications.

Figure 3: Language trends – distribution of publication languages over time



Source: Own work

Publications in Serbian across all mentioned categories were most frequently issued between 1930 and 1990, mainly in connection with official gazettes. Another commonly used language is German, particularly during the 19th century – especially between 1850 and 1870 – when many publications related to Slovenian history were produced in the Austrian Empire (1804–1867) and Austria-Hungary (1867–1918), often written in both German and Slovenian. While German was less prominent in the late 18th and early 19th centuries (1710–1830), it became more notable in conjunction with Slovenian in later decades. Less frequently used languages – each appearing fewer than five times – include Spanish (Castilian), Latin, Polish, Arabic, Albanian, Ukrainian, and Esperanto.[40] Additionally, 42 publications are multilingual, and some items have no associated language. This is especially true of the digitised collection "*The Casting of Death*", a series of death masks detailed in Meden and Pančur.[41]

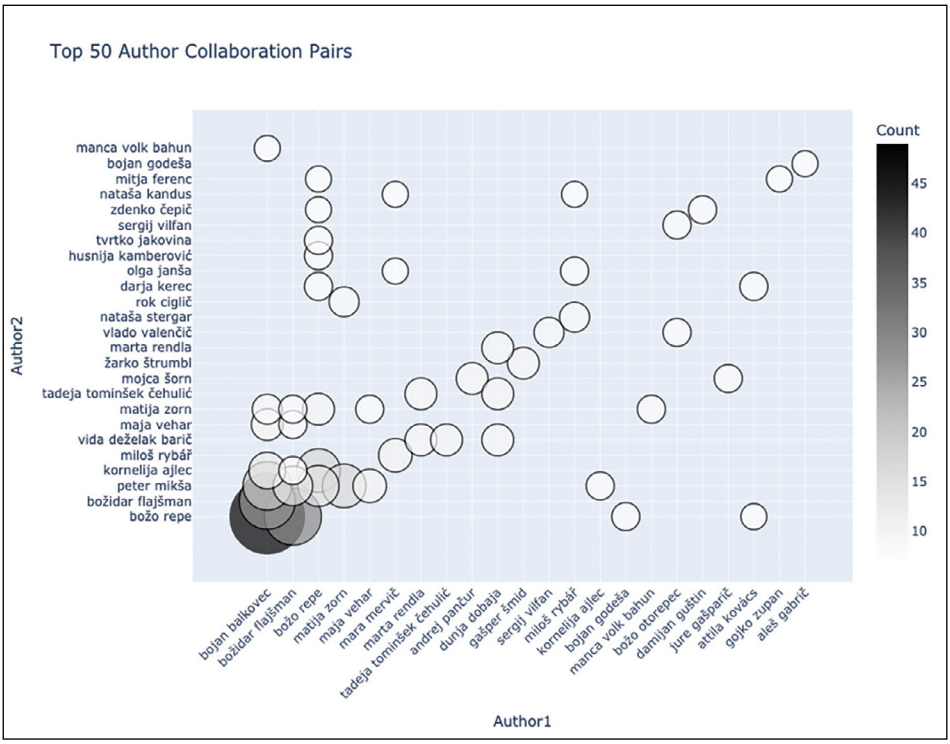
Authorship

As the final part of our comprehensive analysis of the content and characteristics of the SIstory portal, we examined authorship trends, focusing on frequent author pairs (Figure 4) and co-authorship networks (Figure 5).

40 *Esperantsko-slovenski in slovensko-esperantski slovar*, accessed on 26 February 2025, https://hdl.handle.net/11686/38289.

41 Katja Meden, "'Posmrtno življenje posmrtnih mask': sodelovanje Raziskovalne infrastrukture slovenskega zgodovinopisja Inštituta za novejšo zgodovino z Društvom za domače raziskave," in Alenka Pirman, ed., *Odlivanje smrti: posmrtne maske v slovenskih javnih zbirkah* (Ljubljana: Institute of Contemporary History, 2023), accessed on 26 February 2025, https://sistory.github.io/Odlivanje_smrti/ch02.html. Andrej Pančur, Alenka Pirman, and Maruša Dražil, "Spregledana kulturna dediščina in uporaba digitalne raziskovalne infrastrukture za humanistiko v raziskavi Odlivanje smrti," in Alenka Pirman, ed., *Odlivanje smrti: posmrtne maske v slovenskih javnih zbirkah* (Ljubljana: Institute of Contemporary History, 2023), accessed on 26 February 2025, https://sistory.github.io/Odlivanje_smrti/ch01.html.

Figure 4: Bubble plot of the 50 most frequent author pairs. The size and darker colours of the bubbles indicate higher counts of co-occurrences for specific pairs.



Source: Own work

Figure 4 shows the co-occurrence of author pairs based on the "Author" metadata field, which includes only individual publication authors. Organisational, editorial, and translation authorship were excluded from the analysis.

Of the original 6317 author pair combinations, 5511 occurred only once, representing 87.3% of all author pairings. The remaining 12.7% consist of combinations with multiple occurrences. This may suggest limited recurring collaborations and a diverse research network, as well as the fact that the production on SIstory includes a high number of unique collaborators. Supporting this idea is the very high percentage of publications with single authorship across SIstory's entire production (94.5%)[42], indicating a trend of authors favouring individual work over joint efforts.

Several authors have frequently collaborated on publications available through the portal, with specific pairs standing out for their joint work. The most frequent collaborators are Bojan Balkovec and Božo Repe, who co-authored 49 publications. They are followed by Božidar Flajšman and Božo Repe with 29 joint works, and Bojan Balkovec and Božidar Flajšman with 28. Although these author pairs exhibit high

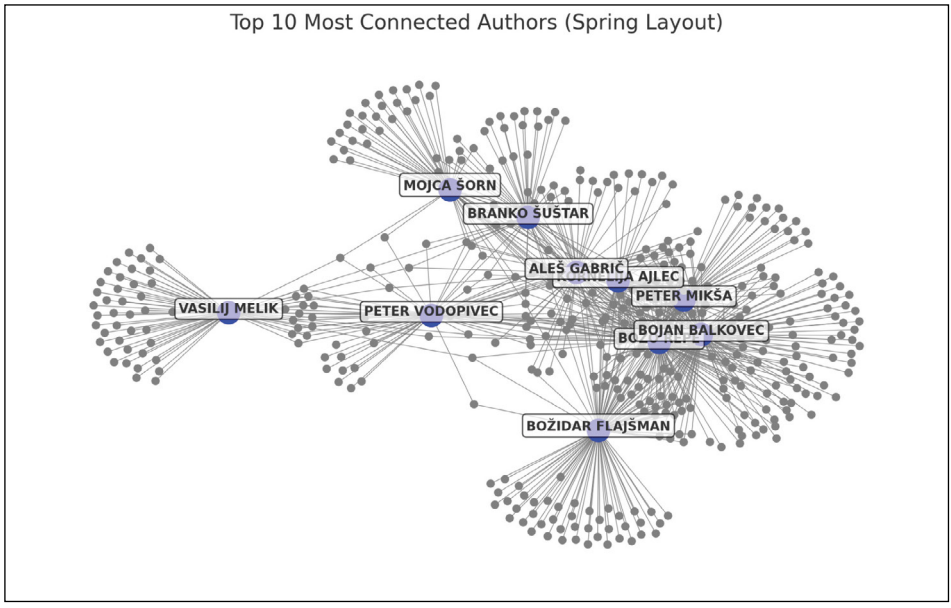
42 For this analysis, publications without an identified author (creator) have been excluded from the total publication count. These publications, primarily official gazettes, account for approximately 63% of the entire database.

co-occurrence, manual checks reveal that most of their collaborations are concentrated in the *Events* section (*Dogodki*), specifically video recordings on historical themes produced by the Department of History, Faculty of Arts, University of Ljubljana – an institution to which all three authors are affiliated.

In contrast, author pairs with fewer than ten co-occurrences tend to be more varied and are more likely to produce traditional academic outputs such as scientific papers or monographs. For example, Gašper Šmid and Žarko Štrumbl co-authored eight scientific papers, mainly in the journal Arhivi, while Mitja Ferenc and Božo Repe collaborated on six chapters in a scientific monograph. However, this pattern is not universal; manual checks of author pairs within the 5–10 co-occurrence range revealed exceptions. One such case is that of Marta Rendla and Vida Deželak Barič, who co-authored eight works – all of which are periodic interim reports for a research project.[43]

Within the scope of the authorship analysis, we also examined the author network. Figure 5 shows the ten most connected authors, based on the number of direct co-authorship links for each author.

Figure 5: Network of the ten most connected authors



Source: Own work

The co-authorship network, shown in Figure 5, highlights the ten most interconnected authors and their relationships, outlining key figures within the SIstory

43 Vida Deželak Barič et al., *Vmesno poročilo o rezultatih opravljenega raziskovalnega dela na projektu v okviru ciljnega raziskovalnega programa (CRP) "Konkurenčnost Slovenije 2006–2013": Vmesno poročilo velja za obdobje od 15. 11. 2010 do 15. 3. 2011: Pregled mrliških matičnih knjig za ugotovitev števila ter strukture žrtev druge svetovne vojne in neposredno po njej* (Ljubljana: Institute of Contemporary History, 2011), https://hdl.handle.net/11686/1120.

production. With a density metric of 0.0046, the network is relatively sparse, indicating that most authors are not directly connected. However, the average path length of 3.89 suggests that most authors can be linked within about four steps, implying that although direct connections are limited, authors are frequently linked through short co-authorship chains.

Many of the most connected authors, such as Božo Repe, Bojan Balkovec, and Božidar Flajšman, are already identified in Figure 4 as frequent collaborators from the same institution. However, additional authors, like Mojca Šorn and Vasilij Melik, although not part of recurring co-authorship pairs, are notable for their high connectivity within the network. Their position on the periphery of the top ten suggests they may not co-author often but remain important figures within the portal's production.

Conclusions

This paper presents the History of Slovenia – SIstory.si portal, detailing its background and the technical, visual, and content-related updates in the latest version (5.0). The development of the portal illustrates its evolving role in supporting historical research and its contribution to the expanding field of digital humanities. Over time, its reach has grown beyond the original academic focus, serving a broader audience interested in Slovenian history.

The decision to redesign the portal from the ground up appears to have been a step in the right direction, as initial user feedback indicates improved responsiveness and a generally smoother user experience.

To better understand the scope and structure of the portal's content, we conducted an exploratory content analysis. While the portal's content was recently the focus of a study,[44] the emphasis was on the chronological additions to the SIstory portal throughout its history. Still, no detailed (metadata) analysis has been performed to help us understand the content coverage and themes of the portal. Therefore, the redesign offered us the ideal opportunity to familiarise ourselves more thoroughly with the content we have collected and worked on so far, thus creating a valuable foundation for the future.

The initial content study, based on trends in publication over time, provided an overview of content distribution and outlined the likely reasons for this. These were then further examined through keyword analysis, which showed, to some degree, that the type of publications within the peaks of the graph aligns well with the hypothesised reasons for such a distribution (i.e., large volumes of publications of the same type, such as textbooks and censuses). This also applies to the language analysis, which primarily aimed to provide an overview of the variety of publication languages available on the portal. Given the historical context, it is not surprising that in the 18th and

44 Šorn and Cvek, *Vsebine*.

19th centuries, publications were mainly in German, with some exceptions published in Slovenian, while in the 20th century, language coverage began to expand to other South Slavic languages (again with Slovenian as the dominant language).

Lastly, the authorship analysis allowed us to familiarise ourselves with the diversity of (co-)authorship and connections/relations amongst them. The initial overview revealed a notably high number of single-authored publications. Additionally, examining author pairs revealed that co-authorship within SIstory is quite limited and primarily linked to the specific institution and type of content produced, such as extensive video production by a particular institution to which the authors belong. The subsequent analysis of the author network suggests a relatively sparse overall network, with most authors not directly connected. It also highlights the rather unsurprising fact that the authors within the most frequent author pairs exhibit the centrality of the author network.

Our future work will focus on further analysing the technical processes behind data collection and expanding the scope of metadata. The current state of the redesigned portal will serve as a foundation for more direct community involvement in the development process. We aim to incorporate user feedback from the beginning and explore the integration of visualisations that will facilitate easier interaction with the data. We will also continue to improve existing collections, ensuring the portal remains a comprehensive, dynamic, and accessible platform for historical research.

Acknowledgement

The research was conducted within the framework of the DARIAH-SI research infrastructure, the infrastructure programme I0-0013 *The Research Infrastructure of Slovenian Historiography infrastructure programme* [*Raziskovalna infrastruktura slovenskega zgodovinopisja*], and research programme P6-0436 *Digital humanities: resources, tools and methods* [*Digitalna humanistika: viri, orodja in metode*], which are co-financed by the Slovenian Research and Innovation Agency (ARIS) from the state budget and the RSF.

Sources and Literature

Cvek, Ana, Mihael Ojsteršek, and Mojca Šorn. *Izhodišča metapodatkovnih sistemov portala Zgodovina Slovenije – SIstory (2008–2016)*. Ljubljana: Institute of Contemporary History, 2022.

Deželak Barič, Vida, Tadeja Tominšek Čehulić, Dunja Dobaja, and Marta Rendla. *Vmesno poročilo o rezultatih opravljenega raziskovalnega dela na projektu v okviru ciljnega raziskovalnega programa (CRP) "Konkurenčnost Slovenije 2006–2013": Vmesno poročilo velja za obdobje od 15. 11. 2010 do 15. 3. 2011: Pregled mrliških matičnih knjig za ugotovitev števila ter strukture žrtev druge svetovne vojne in neposredno po njej.* Ljubljana: Institute of Contemporary History, 2011. Accessed February 26, 2025. https://hdl.handle.net/11686/1120.

Hadalin, Jurij. "The Slovenian Digital Humanities Landscape? A Brief Overview." In *The Status Quo of Digital Humanities,* edited by Torsten Kahlert and Claudia Prinz, 154–69. Berlin: H-Soz–Kult, 2015. Accessed February 26, 2025. https://edoc.hu-berlin.de/server/api/core/bitstreams/febdd0e3-5c2b-4eb8-8b30-ea8f1b36cae7/content.

Institute of Contemporary History. *Poročilo o doseženih ciljih in rezultatih v letu 2008*. Ljubljana: Institute of Contemporary History, 2009. Accessed February 26, 2025. https://www.inz.si/f/docs/Letna-porocila/2008.pdf.

Lemmens, Bert, Joris Janssens, Ruth V. Dyck, Alessia Bardi, Paolo Manghi, Eric Beving, Kathryn Máthé, Katalin Dobó, and Armin Straube. *Hope - The Common HOPE Metadata Structure, Including the Harmonisation Specifications (D2.2)*. Tech. Rep., HOPE, Deliverable D2.2, 2011.

Meden, Katja, and Ana Cvek. "Nadgradnja zgodovinarskega indeksa citiranosti." *Slovenščina 2.0* 9, No. 1 (2021): 216–35.

Meden, Katja. "'Posmrtno življenje posmrtnih mask': sodelovanje Raziskovalne infrastrukture slovenskega zgodovinopisja Inštituta za novejšo zgodovino z Društvom za domače raziskave." In *Odlivanje smrti: posmrtne maske v slovenskih javnih zbirkah,* edited by Alenka Pirman. Ljubljana: Institute of Contemporary History, 2023. Accessed February 26, 2025. https://sistory.github.io/Odlivanje_smrti/ch02.html.

Pančur, Andrej, Alenka Pirman, and Maruša Dražil. "Spregledana kulturna dediščina in uporaba digitalne raziskovalne infrastrukture za humanistiko v raziskavi Odlivanje smrti." In *Odlivanje smrti: posmrtne maske v slovenskih javnih zbirkah*, edited by Alenka Pirman. Ljubljana: Institute of Contemporary History, 2023. Accessed February 26, 2025. https://sistory.github.io/Odlivanje_smrti/ch01.html.

Pančur, Andrej, and Mojca Šorn. "Na začetku je bil SIstory: Raziskovalna infrastruktura slovenskega zgodovinopisja." In *Inštitut za novejšo zgodovino: 60 let mislimo preteklost,* 47–58. Ljubljana, Institute of Contemporary History, 2019. Accessed February 26, 2025. https://hdl.handle.net/11686/46230.

Pančur, Andrej, Neja Blaj Hribar, Mojca Šorn, and Mihael Ojsteršek. "Projekt Vojaške žrtve prve svetovne vojne na Slovenskem." In *Proceedings of the Conference on Language Technologies and Digital Humanities: September 24th–25th 2020, Ljubljana, Slovenia*, edited by Darja Fišer and Tomaž Erjavec, 136–40. Ljubljana: Institute of Contemporary History, 2020. Accessed February 26, 2025. https://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Pancur-et-al_Projekt-Vojaske-zrtve-prve-svetovne-vojne-na-Slovenskem.pdf.

Pančur, Andrej. "Popisi prebivalstva Slovenije 1830–1931: Orodje za transkribiranje historičnih demografskih podatkov." In *Proceedings of the Conference on Language Technologies & Digital Humanities, September 29th–October 1st, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*, edited by Tomaž Erjavec and Darja Fišer, 133–41. Ljubljana: Ljubljana University Press, Faculty of Arts, 2016. Accessed February 26, 2025. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Pancur_Popisi-prebivalstva-Slovenije-1830-1931.pdf.

Pančur, Andrej. *Metapodatki portala Zgodovina Slovenije-SIstory.* Tech. Rep. Ljubljana: Institute of Contemporary History, 2013.

Pančur, Andrej. *SIstory augmented reality 1.0 XML Schema, Documentation*. Institute of Contemporary History, 2013. Accessed February 26, 2025. https://hdl.handle.net/11686/20385.

Pančur, Andrej. *SIstory nadgrajena resničnost 1.0 XML shema*. Institute of Contemporary History, 2013. Accessed February 26, 2025. https://hdl.handle.net/11686/20369.

Rožman, Bogomir, and Gregor Marolt. *Analiza podatkov in postavitev standardov in infrastrukture za nadgradnjo portala SIstory.si.* Ljubljana: Institute of Contemporary History, 2011.

SIstory.si - Culture of Slovenia. Accessed February 26, 2025. https://www.culture.si/en/SIstory.si.

STA. "SIstory - spletni portal slovenskega zgodovinopisja." *Siol.net*. Accessed February 26, 2025. https://siol.net/novice/novice/sistory-spletni-portal-slovenskega-zgodovinopisja-336634.

Šorn, Mojca, and Ana Cvek. *Vsebine in njihovo razporejanje na portalu Zgodovina Slovenije - SIstory (s poudarkom na publikacijah)*. Ljubljana: Institute of Contemporary History, 2023.
Šorn, Mojca, and Katja Meden. "Portal Zgodovina Slovenije – SIstory in avtorske pravice." *Prispevki za novejšo zgodovino* 61, No. 2 (2021): 193–228.
Šorn, Mojca, Andrej Pančur, and Mitja Sunčič. "SIstory: arhivsko gradivo in e-humanistika." *Arhivi* 34, No. 1 (2011): 145–48.
Štepec, Marko, and Mojca Turk. *Slovenians and the First World War*, 2011. Accessed February 26, 2025. https://hdl.handle.net/11686/1160.
ZIC - Zgodovinski indeks citiranosti. Accessed February 26, 2025. https://zic.sistory.si/.

Katja Meden, Ana Cvek, Vid Klopčič, Mihael Ojsteršek, Matevž Pesek, Mojca Šorn, Andrej Pančur

ODPIRANJE ZGODOVINE: PRENOVA IN ANALIZA VSEBINE PORTALA SISTORY 5.0

POVZETEK

Portal Zgodovina Slovenije – SIstory.si že od svojih začetkov predstavlja pomembno zbirko publikacij, podatkov, zbirk in metapodatkov, ključnih za raziskovanje slovenske zgodovine. Njegova vloga sega onkraj golega arhiviranja zgodovinskih virov, saj omogoča napredno raziskovanje zgodovinopisja z vključevanjem sodobnih digitalnih pristopov. V svojem sedemnajstletnem obstoju je portal doživel več tehničnih nadgradenj, katerih cilj je bil izboljšati preglednost, interoperabilnost in dostopnost podatkov tako za raziskovalce kot širšo javnost. Zadnja tehnična posodobitev, predstavljena v prispevku, je vključevala optimizacijo podatkovne infrastrukture, izboljšanje iskalnih mehanizmov in prilagoditev metapodatkovne sheme.

Prispevek se v drugem delu posveča vsebinski analizi metapodatkov. Začetna študija vsebine, ki je temeljila na časovnih trendih objav, je omogočila pregled porazdelitve vsebine in opredelila verjetne razloge zanjo. Ti so bili nato dodatno raziskani z analizo ključnih besed, ki je delno potrdila, da vrhovi v grafu ustrezajo pričakovanim vzorcem – predvsem zaradi velikih količin publikacij iste vrste, kot so učbeniki in popisi. To velja tudi za jezikovno analizo, ki je zagotovila pregled različnih jezikov objav, dostopnih na portalu. Glede na zgodovinski kontekst ni presenetljivo, da so bile publikacije v 18. in 19. stoletju večinoma v nemščini, z nekaj izjemami v slovenščini. V 20. stoletju se je jezikovni nabor razširil na druge južnoslovanske jezike, pri čemer je slovenščina ostala prevladujoča.

Analiza avtorstva je omogočila vpogled v raznolikost (so)avtorstva ter povezave in odnose med avtorji. Analiza avtorskih parov je razkrila, da je soavtorstvo vsebin znotraj SIstory zelo omejeno in tesno povezano z določeno institucijo ter vrsto ustvarjene vsebine. Analiza mreže avtorjev kaže na relativno razpršenost celotne mreže, saj večina avtorjev ni neposredno povezana. Hkrati poudarja tudi – čeprav ne nepričakovano –, da avtorji znotraj najpogostejših avtorskih parov predstavljajo center v avtorskemu omrežju.

V prihodnje se bo razvoj portala osredotočil na podrobno analizo internih tehničnih procesov in konsolidacijo (meta)podatkov. Prenovljeni portal bo služil kot temelj za nadaljnje izboljšave, pri čemer bo še naprej aktivno vključena skupnost – pristop, ki je bil ključen že v dosedanjem razvoju portala. Ena izmed predvidenih nadgradenj je vključitev orodij za vizualizacije podatkov, ki bi raziskovalcem in drugim uporabnikom omogočila lažjo interakcijo s podatki. Poleg tega bo uredništvo portala nadaljevalo z dopolnjevanjem obstoječih zbirk in zagotavljanjem celovite ter dinamične platforme za dostopne in ponovno uporabne sodobne zgodovinske vire.

1.01 DOI: https://doi.org/10.51663/pnz.65.3.05

Luka Terčon,* Kaja Dobrovoljc,° Nikola Ljubešić♦

CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages

IZVLEČEK

CLASSLA-STANZA: NASLEDNJI KORAK ZA JEZIKOVNO PROCESIRANJE JUŽNOSLOVANSKIH JEZIKOV

V članku predstavljamo orodje CLASSLA-Stanza, cevovod za avtomatsko jezikovno označevanje južnoslovanskih jezikov, ki temelji na cevovodu za procesiranje naravnega jezika Stanza. Opišemo vse glavne izboljšave, ki jih prinaša CLASSLA-Stanza v primerjavi s Stanzo in podamo podroben opis postopka učenja modelov v različici 2.2, najnovejši različici orodja. Obenem poročamo o rezultatih delovanja cevovoda za različne jezike in jezikovne zvrsti. CLASSLA-Stanza dosega konsistentno visoke rezultate za vse podprte jezike in preseže rezultate izvornega cevovoda Stanza pri vseh podprtih jezikih. Predstavimo tudi novo funkcijo cevovoda, ki omogoča učinkovito procesiranje spletnih besedil, in opišemo učinkovitost cevovoda za označevanje transkriptov govora.

Ključne besede: južnoslovanski jeziki, avtomatsko procesiranje jezika, označevalni cevovod, jezikovno označevanje

* **Tch. Asst., University of Ljubljana, Faculty of Arts, Aškerčeva 2, SI-1000 Ljubljana, Slovenia; Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana, Slovenia, luka.tercon@ff.uni-lj.si; ORCID: 0009-0006-3237-3583**

° **PhD, Res. Assoc., University of Ljubljana, Faculty of Arts, Aškerčeva 2, SI-1000 Ljubljana, Slovenia; Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, kaja.dobrovoljc@ff.uni-lj.si; ORCID: 0000-0002-5909-7965**

♦ **PhD, Sr. Res. Assoc., Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia; University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia; Institute of Contemporary History, Privoz 11, SI-1000 Ljubljana, Slovenia, nikola.ljubesic@ijs.si; ORCID: 0000-0001-7169-9152**

ABSTRACT

We present CLASSLA-Stanza, a pipeline for automatic linguistic annotation of South Slavic languages, which is based on the Stanza natural language processing pipeline. We describe the main improvements in CLASSLA-Stanza with respect to Stanza and give a detailed description of the model training process for the latest 2.2 release of the pipeline. We also report performance scores produced by the pipeline for different languages and language varieties. CLASSLA-Stanza exhibits consistently high performance across all the supported languages and outperforms its parent pipeline Stanza at all the supported tasks. We also present the pipeline's new functionality that enables efficient processing of web data and describe the efficiency of the pipeline for annotating written transcripts of spoken data.

Keywords: South Slavic languages, automatic linguistic processing, annotation pipeline, linguistic annotation

Introduction

The South Slavic languages make up one of the three major branches of the Slavic language family. Despite being used by around 30 million people worldwide,[1] many languages of this group remain relatively low-resourced and under-represented in the field of natural language processing. Goldhahn et al.[2] include Macedonian and Bosnian in their list of languages that are significantly under-resourced despite having more than 1 million speakers.

Although much additional work is required before South Slavic languages can approach the level of support enjoyed by linguistic giants such as English, steps have been taken towards establishing common platforms for supporting the development of new resources and tools for these languages. The CLARIN Knowledge Centre for South Slavic Languages (CLASSLA)[3] was established as a result of prior cooperation in the development of language resources for Slovenian, Croatian, and Serbian and currently acts as a platform providing expertise and support for developing language resources for South Slavic languages.[4] The efforts of the knowledge centre gave rise to

1 Nikola Ljubešić et al., "Tour de CLARIN: The CLARIN Knowledge Centre for South Slavic Languages (CLASSLA)," CLARIN, published 18 November, 2021, https://www.clarin.eu/blog/tour-de-clarin-clarin-knowledge-centre-south-slavic-languages-classla.

2 Dirk Goldhahn et al., "Corpus collection for under-resourced languages with more than one million speakers," paper presented at the *Collaboration and Computing for Under-Resourced Languages (CCURL)* workshop, 2016, http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-CCURL2016_Proceedings.pdf#page=74.

3 CLASSLA: *Knowledge centre for South Slavic languages*, https://www.clarin.si/info/k-centre/.

4 Nikola Ljubešić et al., "Together we are stronger: Bootstrapping language technology infrastructure for South Slavic languages with CLARIN. SI," in Darja Fišer and Andreas Witt, eds., *CLARIN. The Infrastructure for Language Resources* (De Gruyter, 2022), 429–56.

the CLASSLA-Stanza[5] pipeline for linguistic processing, which arose as a fork of the Stanza neural pipeline.[6] CLASSLA-Stanza was created with the aim of providing state-of-the-art automatic linguistic processing for South Slavic languages[7] and currently supports Slovenian, Croatian, Serbian, Macedonian, and Bulgarian. Additionally, Slovenian, Croatian, and Serbian have support for standard, nonstandard, and internet varieties, while Slovenian also supports processing spoken language transcripts. In contrast to its parent pipeline Stanza, CLASSLA-Stanza covers the standard Macedonian language, as well as the nonstandard and internet varieties of Slovenian, Croatian, and Serbian and the spoken variety of Slovenian. Besides the expanded coverage of languages and varieties, CLASSLA-Stanza shows improvements in performance at all presented levels.

The aim of this paper is to provide both a systematic overview of the differences that CLASSLA-Stanza has to the official Stanza pipeline and a description of the model training procedure which was adopted when training models for the latest releases. The description of the training procedure is intended to serve as the main reference for future releases as well as for anyone using the CLASSA-Stanza tool to produce their own models for linguistic annotation.

In accordance with this aim, we first describe the differences between CLASSLA-Stanza and Stanza in Section Differences Between CLASSLA-Stanza and Stanza. Section Datasets then introduces the datasets used for training the most recent models. Section Model Training gives a general description of the model training process, which is followed by an analysis of the results produced by the latest models in Section Model Performance Analysis..

At present, the CLASSLA-Stanza annotation tool supports a total of six tasks: tokenization, morphosyntactic annotation, lemmatization, dependency parsing, semantic role labelling, and named-entity recognition. Tokenization is handled by one of two external rule-based tokenizers included in CLASSLA-Stanza, either the Obeliks tokenizer for standard Slovenian[8] or the ReLDI tokenizer for nonstandard Slovenian and all other languages.[9] While the basic tasks of tokenization, morphosyntactic annotation, lemmatization, and dependency parsing are covered at least for

5 *GitHub - clarinsi/classla: CLASSLA Fork of the Official Stanford NLP Python Library for Many Human Languages*, https://github.com/clarinsi/classla/.

6 Peng Qi et al., "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages," paper presented at the *58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, https://doi.org/10.18653/v1/2020.acl-demos.14.

7 Nikola Ljubešić and Kaja Dobrovoljc, "What Does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian," paper presented at the *7th Workshop on Balto-Slavic Natural Language Processing*, 2019, https://doi.org/10.18653/v1/W19-3704. Kaja Dobrovoljc et al., "Improving UD processing via satellite resources for morphology," paper presented at the *Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, 2019, https://doi.org/10.18653/v1/W19-8004.

8 Miha Grčar et al., "Obeliks: statisticni oblikoskladenjski oznacevalnik in lematizator za slovenski jezik" [Obeliks: A Statistical Morphosyntactic Annotation and Lemmatization Tool for the Slovenian Language], paper presented at the *Eighth Language Technologies Conference*, 2012, https://nl.ijs.si/isjt12/proceedings/isjt2012_17.pdf. The repository for the tool can be found at: https://github.com/clarinsi/obeliks.

9 Tanja Samardžić et al., "Regional Linguistic Data Initiative (ReLDI)," paper presented at the *5th Workshop on Balto-Slavic Natural Language Processing*, 2015, https://aclanthology.org/W15-5306/. The repository for the tool can be found at: https://github.com/clarinsi/reldi-tokeniser.

some languages in the parent Stanza pipeline, semantic role labelling and named entity recognition for South Slavic languages are available only in CLASSLA-Stanza.

The current version of the models was trained on data annotated according to three separate systems for morphosyntactic annotation: the Universal part-of-speech tags and the Universal morphosyntactic features tags, which are both part of the Universal Dependencies framework for grammatical annotation[10] and will henceforth be referred to as UPOS and UFeats, and the MULTEXT-East V6 specifications for morphosyntactic annotation,[11] which are implemented as the language-specific XPOS tags in the CoNLL-U file format,[12] the central file format used by CLASSLA-Stanza. For dependency parsing, the Universal Dependencies system for syntactic dependency annotation was used as well as the JOS syntactic dependencies system for Slovenian.[13] Additionally, the annotation schema described in Krek et al.[14] was used for semantic role label annotation, while the named entity annotation system followed the guidelines described by Zupan et al.[15]

It must be noted that not all tasks are available for every supported language and variety. For instance, semantic role labelling currently relies on the JOS annotation system for dependency parsing of Slovenian and is thus only available for annotation of Slovenian datasets but should become available for Croatian in the future as there are training data available.[16] Table 1 provides an overview of every language variety and the tasks it supports.

Table 1: Tasks supported by CLASSLA-Stanza for every language and variety. The abbreviations for each task are as follows: Tok – tokenization, Morph – morphosyntactic tagging, Lemma – lemmatization, Depparse – dependency parsing, NER – named entity recognition, SRL – semantic role labelling.

Language	Variety	Tok	Morph	Lemma	Depparse	NER	SRL
Slovenian	standard	✓	✓	✓	✓	✓	✓
	nonstandard	✓	✓	✓	x	✓	x
	spoken	✓	✓	✓	✓	x	x

10 Marie-Catherine de Marneffe et al., "Universal Dependencies," *Computational Linguistics* 47, No. 2 (07 2021): 255–308, https://doi.org/10.1162/coli_a_00402.

11 Tomaž Erjavec, "MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages," *Language Resources and Evaluation* 46, No. 1 (2012), http://www.jstor.org/stable/41486069.

12 *CoNLL-U Format*, https://universaldependencies.org/format.html.

13 Tomaž Erjavec et al., "The JOS Linguistically Tagged Corpus of Slovene," paper presented at the *Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010, https://aclanthology.org/L10-1087/.

14 Simon Krek et al., "Označevanje udeleženskih vlog v učnem korpusu za slovenščino" [Annotating Semantic Roles in a Training Corpus for Slovenian], paper presented at the *Conference on Language Technologies and Digital Humanities (JT-DH-2016)*, 2016, https://doi.org/10.5281/zenodo.14165095.

15 Katja Zupan et al., "Smernice Janes-NER za označevanje imenskih entitet v slovenskem jeziku: Različica 1.1," CJVT Wiki, https://wiki.cjvt.si/books/08-imenske-entitete/page/oznacevalne-smernice.

16 Nikola Ljubešić and Tanja Samardžić, "Croatian Linguistic Training Corpus Hr500k 2.0," 2023, http://hdl.handle.net/11356/1792.

Croatian	standard	✓	✓	✓	✓	✓	x
	nonstandard	✓	✓	✓	x	✓	x
Serbian	standard	✓	✓	✓	✓	✓	x
	nonstandard	✓	✓	✓	x	✓	x
Bulgarian	standard	✓	✓	✓	✓	✓	x
	nonstandard	x	x	x	x	x	x
Macedonian	standard	✓	✓	✓	x	x	x
	nonstandard	x	x	x	x	x	x

Source: Own work

An earlier version of this overview was already presented at the Language Technologies and Digital Humanities conference in 2024,[17] while this paper expands upon that report by describing the training of various new models that are included in the latest 2.2 release[18] of the CLASSLA-Stanza pipeline, including new standard models for Slovenian UD dependency parsing and named entity recognition and also the first Slovenian models for annotating spoken language. We also describe new experiments that compare the effectiveness of Slovenian standard, nonstandard, and spoken models on transcripts of spoken language.

Differences Between CLASSLA-Stanza and Stanza

The Stanza neural pipeline is centred around a bidirectional long short-term memory (Bi-LSTM) network architecture.[19] CLASSLA-Stanza largely preserves the design of Stanza, except in some cases, such as tokenization, where a completely different model architecture is used. CLASSLA-Stanza also expands upon the original design with specific additions that help boost model performance for the South Slavic languages. This section thus lists the main differences between the two pipelines and provides an overview of the difference in the results produced by the models for one of the supported languages.

On the level of tokenization and sentence segmentation, Stanza uses a joint tokenization and sentence segmentation model based on machine learning. We generally view such learnt tokenizers as suboptimal, since training data for the two tasks is always limited in size and thus too few tokenization and sentence-splitting phenomena can be learnt by the model during the training process. Due to this drawback,

17 Nikola Ljubešić et al., "CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages," paper presented at the *Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, 2024, https://doi.org/10.5281/zenodo.13936406.

18 Version 2.2 refers to the latest major release of the pipeline. An additional minor release—version 2.2.1—has been made available during the publishing process of this article. This release resolves compatibility issues with newer versions of the Python programming language and improves the documentation on the GitHub repository but does not add any other substantial changes to the tool.

19 Qi et al., "Stanza."

CLASSLA-Stanza implements rule-based tokenizers, which handle both the task of tokenization as well as sentence segmentation. As stated in the introduction, the two tokenizers used are the Obeliks tokenizer for standard Slovenian[20] and the ReLDI tokenizer for nonstandard Slovenian and all other languages.[21]

CLASSLA-Stanza also adds support for the use of external inflectional lexicons, which is not present in Stanza. For morphologically rich languages, applying this resource to the annotation process usually significantly increases the performance of the model.[22] The South Slavic languages all have quite rich inflectional paradigms, which is why support for inflectional lexicons is present for almost all supported languages in the pipeline.

Most languages support external lexicon use only during lemmatization, except for Slovenian, which supports lexicon use also during morphosyntactic tagging. In that case, the lexicon is put into operation during the tag prediction phase, when the model limits the possible predictions to only those tags that are present in the inflectional lexicon for the specific token. Lexicon usage during lemmatization is similar in both Stanza and CLASSLA-Stanza, the main difference being that Stanza builds a lexicon only from the Universal Dependencies training data, while CLASSLA-Stanza can additionally exploit an inflectional lexicon. Both Stanza and CLASSLA-Stanza use the lexicon for an initial lemma lookup and fall back to predicting the lemma only in case that the form with the corresponding tag is not present in the lexicon. One important difference in the lexicon lookup in CLASSLA-Stanza is that the lookup uses XPOS tags, which contain the full morphosyntactic information, while Stanza uses only the UPOS tag, which is not enough for an accurate lemma lookup in morphologically rich languages.

When training models, Stanza uses a Universal Dependencies dataset as training data for training all the tasks in the pipeline and thus does not enable the user to train models on additional datasets. For certain layers, however, such as lemmatization and morphosyntactic tagging, the South Slavic languages often have more training data than available for dependency parsing, which is exploited by CLASSLA-Stanza. Thus, for example, instead of using only the 210 thousand tokens of data that are used for training the dependency parser, the latest set of standard Croatian models in CLASSLA-Stanza includes morphosyntactic tagging and lemmatization models which were trained on an additional 290 thousand tokens, manually annotated only on these two levels of annotation.

CLASSLA-Stanza also has a special way of handling "closed-class" tokens. Closed-class control is a feature of the tokenizers and ensures that punctuation and symbols are assigned appropriate morphosyntactic tags and lemmas. It also prevents other tokens that are not defined as punctuation and symbols in the tokenizer from being

20 Grčar et al., "Obeliks." The Obeliks tokenizer, featuring an extensive set of linguistically informed rules, is the de facto standard for Slovenian text tokenization. It has been used in tokenizing the majority of Slovenian reference corpora and thus facilitates direct comparisons of newly tokenized data to established corpora.

21 Samardžić et al., "Regional Linguistic Data Initiative (ReLDI)."

22 Ljubešić and Dobrovoljc, "What does Neural Bring?"

annotated as such. In addition to punctuation and symbols, the Slovenian package also includes closed-class control for pronouns, determiners, prepositions, particles, and coordinating and subordinating conjunctions. These additional closed classes are controlled during the morphosyntactic tagging phase using the inflectional lexicon as a reference, disallowing any token to be labelled with a closed class label if it was not defined as such in the lexicon.[23]

The Stanza pipeline relies on pretrained word embeddings as an underlying resource. While it uses embedding collections based on Wikipedia data, CLASSLA-Stanza goes the extra mile by using the CLARIN.SI embeddings,[24] which are skip-gram-based embeddings of 100 dimensions, trained with the fastText tool. These embeddings were primarily prepared for CLASSLA-Stanza but are useful for other tasks as well. They were trained on text collections that are several times larger than Wikipedia and were obtained through web crawling,[25] which ensures much more diverse word embeddings and thereby also better handling of unseen words.

When working with Slovenian, Croatian, or Serbian, the pipeline can be set to any of the four predetermined settings which are used for processing different varieties of the same language. These settings are called *modes* and can be either *standard*, *nonstandard*, or *web*. For Slovenian, an additional *spoken* mode is available. The processing modes determine which model is used on every level of annotation and are associated with their respective language varieties. The reasons for introducing separate processing modes for spoken and web texts are described in Sections Model performance on spoken data and Model performance on web data. Below is an overview showing which model is used on every layer for every mode:

Table 2: Overview of processing modes in CLASSLA-Stanza. *NER tagger* stands for Named Entity Recognition tagger.

Processing mode	Tokenizer	Morpho-syntactic tagger	Lemmatizer	Dependency parser	NER tagger
standard	standard	standard	standard	standard	standard
nonstandard	nonstandard	nonstandard	nonstandard	standard	nonstandard
web	standard	nonstandard	nonstandard	standard	nonstandard
spoken	standard	spoken	spoken	spoken	nonstandard

Source: Own work

23 In-depth instructions on how to use the closed-class control functionality are included in the GitHub repository: https://github.com/clarinsi/classla/blob/master/README.closed_classes.md.

24 Luka Terčon et al., "Word Embeddings CLARIN.SI-Embed.Sl 2.0," 2023, http://hdl.handle.net/11356/1791. Luka Terčon and Nikola Ljubešić, "Word Embeddings CLARIN.SI-Embed.Hr 2.0," 2023, http://hdl.handle.net/11356/1790. Luka Terčon and Nikola Ljubešić, "Word Embeddings CLARIN.SI-Embed.Sr 2.0," 2023, http://hdl.handle.net/11356/1789. Luka Terčon and Nikola Ljubešić, "Word Embeddings CLARIN.SI-Embed.Mk 2.0," 2023, http://hdl.handle.net/11356/1788. Luka Terčon and Nikola Ljubešić, "Word Embeddings CLARIN.SI-Embed.Bg 1.0," 2023, http://hdl.handle.net/11356/1796.

25 Marta Banón et al., "MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages," paper presented at the *23rd Annual Conference of the European Association for Machine Translation (EAMT)*, 2022, https://aclanthology.org/2022.eamt-1.41/.

The reason why the nonstandard and the web processing modes use the standard dependency parsing model is primarily the lack of training data for training a model beyond standard text and spoken transcripts. The lack of motivation for building a dataset for parsing nonstandard text lies in the fact that the parsing model has upstream lemma and morphosyntactic information at its disposal and therefore requires dedicated training data to a much lesser extent than those upstream processes.

To illustrate the performance of CLASSLA-Stanza, Table 3 provides a comparison of the results produced by both Stanza and CLASSLA-Stanza when generating predictions using the Slovenian standard models on the SloBENCH evaluation dataset.[26] SloBENCH is a platform for benchmarking various natural language processing tasks for Slovenian, which also includes a dataset for evaluating the tasks supported by CLASSLA-Stanza. The performance scores are presented in the form of micro-F_1 scores, while the relative error reduction between the scores of the pipelines is presented in percentages.

Table 3: Comparison of performance on the SloBENCH evaluation dataset by both pipelines. Metrics are micro-F_1 scores. Downstream tasks use upstream predictions, not gold labels.

Task	Stanza	CLASSLA-Stanza	Relative error reduction
Sentence segmentation	0.819	0.997	98%
Tokenization	0.998	0.999	50%
Lemmatization	0.974	0.992	69%
Morphosyntactic tagging - XPOS	0.951	0.983	65%
Dependency parsing LAS	0.865	0.911	34%

Source: Own work

Despite CLASSLA-Stanza originating as a fork of Stanza, there are currently no plans to merge CLASSLA-Stanza with the original Stanza project, as CLASSLA-Stanza is intended as a separate project with a different focus. While Stanza takes a broader approach, aiming to achieve good performance across a wide range of different languages, CLASSLA-Stanza focuses more on language-specific solutions that improve performance for the South Slavic languages in particular.

26 Slavko Žitnik and Frenk Dragar, "SloBENCH Evaluation Framework," 2021, http://hdl.handle.net/11356/1469. The SloBENCH online platform can be accessed at https://slobench.cjvt.si/.

Datasets

The latest models included in the 2.2 release of CLASSLA-Stanza were trained on a variety of datasets in five different languages: Slovenian, Croatian, Serbian, Macedonian, and Bulgarian. Slovenian had three types of training datasets available—a standard training dataset, a nonstandard training dataset, and a spoken training dataset. Croatian and Serbian were associated with two training datasets—one consisting of standard-language texts and one consisting of nonstandard texts—while Bulgarian and Macedonian only had a standard-language training dataset available.

Slovenian standard language models were first trained using the 1.0 version of the SUK training corpus.[27] It contains approximately 1 million tokens of text manually annotated on the levels of tokenization, sentence segmentation, morphosyntactic tagging, and lemmatization. Some subsets also contain syntactic dependency, named entity, multi-word expression, coreference, and semantic role labelling annotations. In the second half of 2024, an updated version of the SUK training corpus was released, containing substantially improved annotation quality on the level of UD dependency relation annotations. This new version of the corpus was dubbed SUK 1.1,[28] and an updated dependency parsing model was trained using the new data and included as the default and best-performing dependency parsing model in the 2.2 release of the CLASSLA-Stanza pipeline.

Nonstandard Slovenian models were trained on a combination of the standard training corpus and the nonstandard Janes-Tag training corpus,[29] which consists of tweets, blogs, forums, and news comments, and is approximately 218 thousand tokens in size. It contains manually curated annotations on the levels of tokenization, sentence segmentation, word normalization, morphosyntactic tagging, lemmatization, and named entity annotation.

Slovenian spoken models were trained on a combination of the SUK corpus and the Spoken Slovenian UD Treebank,[30] which is composed of transcribed audio recordings of spoken Slovenian and contains approximately 98 thousand tokens annotated on the levels of tokenization, sentence segmentation, morphosyntactic tagging, lemmatization, and UD dependency relations. Some oversampling of the spoken training

27 Špela Arhar Holdt et al., "Training Corpus SUK 1.0," 2022, http://hdl.handle.net/11356/1747.

28 Špela Arhar Holdt et al., "Training Corpus SUK 1.1," 2024, http://hdl.handle.net/11356/1959.

29 Jakob Lenardič et al., "CMC Training Corpus Janes-Tag 3.0," 2022, http://hdl.handle.net/11356/1732.

30 Kaja Dobrovoljc and Joakim Nivre, "The Universal Dependencies Treebank of Spoken Slovenian," paper presented at the *Tenth International Conference on Language Resources and Evaluation (LREC`16)*, 2016, https://aclanthology.org/L16-1248/.

data had to be performed before training due to the relatively small size of the spoken training dataset compared to the standard written training dataset.[31]

Croatian standard language models were trained on the hr500k training corpus,[32] which consists of about 500 thousand tokens and is manually annotated on the levels of tokenization, sentence segmentation, morphosyntactic tagging, lemmatization, and named entities. Portions of the corpus also contain manual syntactic dependency, multi-word expression, and semantic role labelling annotations. Croatian nonstandard models were trained on a combination of the standard training corpus and the nonstandard ReLDI-NormTagNER-hr training corpus.[33] The ReLDI-NormTagNER-hr corpus contains about 90 thousand tokens of nonstandard Croatian text from tweets and is manually annotated on the levels of tokenization, sentence segmentation, word normalization, morphosyntactic tagging, lemmatization, and named entity recognition.

Serbian standard models were trained on the Serbian portion of the SETimes corpus,[34] which contains about 97 thousand tokens of news articles manually annotated on the levels of tokenization, sentence segmentation, morphosyntactic tagging, lemmatization, and dependency parsing.

Serbian nonstandard models were trained, similar to the previously introduced languages, on a combination of the standard dataset and the nonstandard ReLDI-NormTagNER-sr training corpus.[35] ReLDI-NormTagNER-sr consists of about 90 thousand tokens of Serbian tweets manually annotated on the levels of tokenization, sentence segmentation, word normalization, morphosyntactic tagging, lemmatization, and named entity recognition.

Macedonian standard models were trained on a corpus made up of the Macedonian version of the MULTEXT-East "1984" corpus[36] and the Macedonian SETimes.MK corpus. The MULTEXT-East "1984" corpus consists of the novel *1984* by George Orwell in approximately 113 thousand tokens, while the SETimes.MK corpus in its 0.1 version is made up of 13,310 tokens of news articles.[37] Both corpora are manually

31 For the morphosyntactic tagging and lemmatization training data, it was found that eleven repetitions of the spoken data combined with one instance of written data was appropriate, while for UD dependency parsing only three repetitions of the spoken dataset were necessary due to the smaller size of the UD dependency parsing dataset for written language. A subsequent test was run to determine whether any overfitting had occurred during training of the model with eleven repetitions of spoken training data. An additional morphosyntactic tagging model was trained on six repetitions of spoken data and an appropriate proportion of written data. It was found that the model trained on eleven repetitions of spoken data still performed better during evaluation on the test set than the alternative model trained on six repetitions. We therefore concluded that no overfitting had occurred, and the model with eleven repetitions was chosen as the default model to be included in the pipeline.

32 Nikola Ljubešić and Tanja Samardžić, "Croatian Linguistic Training Corpus Hr500k 2.0."

33 Nikola Ljubešić et al., "Croatian Twitter Training Corpus ReLDI-NormTagNER-Hr 3.0," 2023, http://hdl.handle.net/11356/1793.

34 Vuk Batanović et al., "Serbian Linguistic Training Corpus SETimes.SR 2.0," 2023, http://hdl.handle.net/11356/1843.

35 Nikola Ljubešić, et al., "Serbian Twitter Training Corpus ReLDI-NormTagNER-Sr 3.0," 2023, http://hdl.handle.net/11356/1794.

36 Tomaž Erjavec et al., "MULTEXT-East '1984' Annotated Corpus 4.0," 2010, http://hdl.handle.net/11356/1043.

37 Nikola Ljubešić and Biljana Stojanovska, "Macedonian Linguistic Training Corpus SETimes.MK 0.1," 2023, http://hdl.handle.net/11356/1886.

annotated on the levels of tokenization, sentence segmentation, morphosyntactic tagging, and lemmatization. Combining the corpora was performed in the following way: the 1984 corpus was first split into three parts to obtain the training, validation, and testing data splits, after which the training data was enriched with three repetitions of the SETimes corpus to ensure a sensible combination of literary and newspaper data in the training subset.

Bulgarian standard models were trained on the BulTreeBank training corpus,[38] which consists of approximately 253 thousand tokens manually annotated on the levels of tokenization, sentence segmentation, morphosyntactic tagging, and lemmatization. About 60% of the dataset also contains manual dependency parsing annotations.

Table 4 provides an overview of dataset sizes for every language, variety, and annotation layer.

Table 4: Overview of the number of tokens annotated on every annotation layer for all training datasets used. The abbreviations for each task are as follows: Morph – morphosyntactic tagging, Lemma – lemmatization, Depparse – dependency parsing, SRL – semantic role labelling.

Language	Variety	Morph	Lemma	Depparse	SRL
Slovenian	standard	1,025,639	1,025,639	267,097	209,791
	nonstandard	222,132	222,132	n/a	n/a
	spoken	98,396	98,396	98,396	n/a
Croatian	standard	499,635	499,635	199,409	n/a
	nonstandard	89,855	89,855	n/a	n/a
Serbian	standard	97,673	97,673	97,673	n/a
	nonstandard	92,271	92,271	n/a	n/a
Bulgarian	standard	253,018	253,018	156,149	n/a
Macedonian	standard	153,091	153,091	n/a	n/a

Source: Own work

Model Training

In this section, the model training process is described in detail. Only a descriptive account of the process is provided here. For a list of the specific commands and oversampling scripts used, refer to the GitHub repository of the training procedure.[39]

In this paper, we give the general overview of the process which is common to all supported languages. For the specific steps that are unique to each language, please

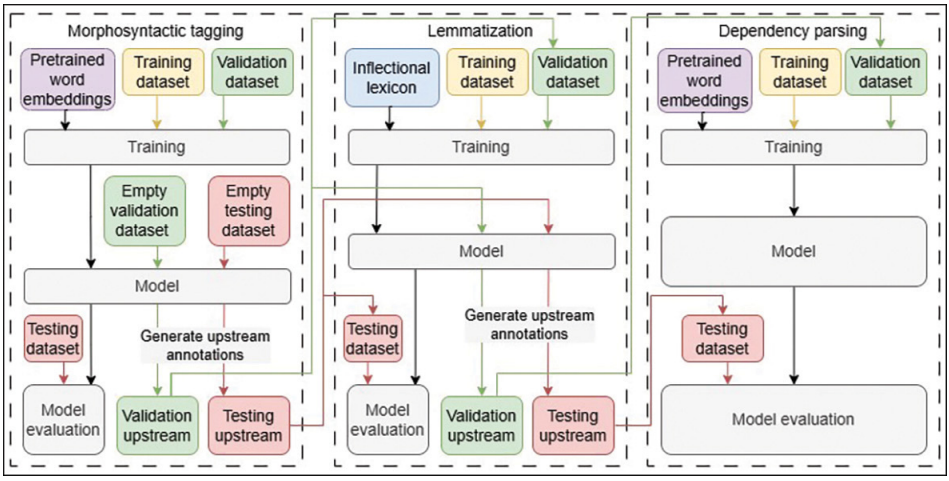
38 Osenova, Petya and Kiril Simov, "Universalizing BulTreeBank: A Linguistic Tale about Glocalization," paper presented at the 5th *Workshop on Balto-Slavic Natural Language Processing*, 2015, https://aclanthology.org/W15-5313/.

39 *GitHub - clarinsi/classla-training: Training scripts for the CLASSLA pipeline*, https://github.com/clarinsi/classla-training.

refer to the CLASSLA-Stanza technical report, a longer and older version of this paper available on arXiv.[40] The language-specific steps were necessary due to some features and levels of annotation (semantic role labelling, oversampling of the training data, etc.) which are unique to certain languages, while all languages share the steps described below.

The illustration of the basic procedure that was used to train standard models for the levels of morphosyntactic tagging, lemmatization, and dependency parsing for the latest release of CLASSLA-Stanza is shown in Figure 1.

Figure 1: Diagram of the basic model training process for standard morphosyntactic tagging, lemmatization, and dependency parsing models



Source: Own work

As stated in the introduction, all tokenizers used by CLASSLA-Stanza are rule-based and thus do not need to be trained. Model training is thus performed on pre-tokenized data, typically beginning on the level of morphosyntactic tagging and continuing on through the subsequent annotation layers.

To ensure realistic evaluation results, automatically generated upstream annotations, rather than manually assigned annotations, were used as validation and test dataset inputs on each layer. For this, empty validation and test datasets first had to be generated by stripping all annotations from the test and validation datasets on all levels except for tokenization. These empty files were filled with model-generated annotations on each level, so that validation and model evaluation on subsequent layers could be performed on automatically generated upstream labels. Training datasets were not annotated with automatically generated upstream labels, since it is unclear whether this would lead to any performance gains and would require a more complicated type

40 Luka Terčon and Nikola Ljubešić, "CLASSLA-Stanza: The next step for linguistic processing of South Slavic Languages," *arXiv preprint* (2023), https://doi.org/10.48550/arXiv.2308.04255.

of cross-validation method such as jackknifing (splitting the data into N bins, training a model on N-1 bins, and annotating the N-th bin, repeating the process N times).

For each language, standard models were first trained. For morphosyntactic tagging, the training and validation datasets from the prepared three-way data split along with the pretrained word embeddings were used as inputs to the tagger module. After training, the tagger was used in predict mode to generate predictions on the empty test dataset and evaluate the performance of the tagger. After predictions were made for the test set, predictions were generated in the empty validation dataset as well to produce a validation file with automatically generated morphosyntactic labels, that can be used later during training of subsequent annotation layer models, such as those for lemmatization and dependency parsing.

Once morphosyntactic predictions and evaluation results were obtained, the lemmatizer was trained. The validation and training datasets were used as inputs. In addition, for most languages, the inflectional lexicon is also provided to the lemmatizer as an underlying resource. During training, the lexicon is stored in the lemmatization model file to act as an additional controlling element during lemmatization. After training, the lemmatizer was run in predict mode to obtain evaluation results and add lemma predictions to the validation and test datasets for the training of the dependency parser model.

The dependency parser model was trained after lemmatizer training was finalized. CLASSLA-Stanza currently supports two types of annotation systems for syntactic dependency annotation: the UD dependency parsing annotation system, which is available for all supported languages except Macedonian, and the JOS parsing system, which is only available for Slovenian.[41] The parser was run in training mode using the training and validation datasets[42] as inputs along with the pretrained word embeddings. After training, the parser was run in predict mode to obtain evaluation results.

The process for training models for named-entity recognition was quite similar to the other tasks. The tagger training here accepts pretrained word embeddings and training and validation datasets as underlying resources. After training, the named entity recognition tagger can be run in predict mode to obtain evaluation results.

The spoken models were trained using the same process as the standard models, and a similar process was also followed for the nonstandard models with a few notable exceptions. Firstly, no syntactic dependency annotations are present in the nonstandard datasets. As a result, no nonstandard dependency parsing models were trained. Before training the nonstandard models, approximately 20% of diacritics were removed from the training datasets to ensure that the models will learn to effectively handle dediacritized forms, which occur prominently in online communication.

41 In comparison to UD, the JOS parsing system features a more concise set of dependency relations focusing on core syntactic constructs and has thus been preferred over UD in some applications.

42 For most languages, only a portion of the original datasets contained dependency parsing annotations. In these cases, a separate set of training, validation, and testing datasets consisting of only this portion of the original data had to be extracted.

Model Performance Analysis

As noted in Section Differences Between CLASSLA-STANZA and Stanza, CLASSLA-Stanza significantly outperforms Stanza on the Slovenian benchmark, with the relative error reduction between 34% and 98%, depending on the processing layer.

However, in order to fully assess the performance of the newly-trained models, we conduct a series of additional performance analyses in this section. In Section Model performance on UPOS and UD labels, we give a detailed rundown of the performance of the models for various UPOS and syntactic dependency labels for each language. In Section Model performance on spoken data, we present experiments using various Slovenian models to annotate spoken data and discuss why the training of models specifically dedicated to annotating spoken transcripts was justified. Finally, in Section Model performance on web data we present a more qualitative investigation into the performance of the models on web-specific data.

Model performance on UPOS and UD labels

To obtain a sense of which categories a model struggles with and which ones it handles well, model predictions for specific UPOS and UD syntactic relations were inspected. An accuracy score was calculated for all 17 UPOS labels and the 12 most frequent UD syntactic relations in the Croatian hr500k training corpus.[43] The accuracy score was obtained by taking the number of correct predictions for a single label in the test dataset and dividing it by the total number of occurrences of that label in the test dataset. The resulting accuracies for all the UPOS tags are contained in Table 5, while Table 6 contains accuracies for each UD dependency relation.

Table 5: Table of per-relation accuracies for all UPOS tags. The language abbreviations are followed by "st" for *standard*, "nonst" for *nonstandard*, or "spok" for *spoken*.

UPOS tag	Accuracy									
	sl-st	sl-nonst	sl-spok	hr-st	hr-nonst	sr-st	sr-nonst	mk-st	bg-st	Average
ADJ	99.31	90.71	97.72	97.93	92.27	99.27	94.58	97.74	98.28	96.26
ADP	99.90	98.54	100	99.96	99.82	100.00	99.84	99.75	99.92	99.72
ADV	95.98	91.89	94.70	95.35	91.59	95.42	87.93	95.14	97.60	93.86
AUX	98.62	96.31	96.13	99.60	99.59	100.00	98.81	99.50	92.75	98.15
CCONJ	98.01	97.03	98.93	96.53	97.21	98.95	97.21	97.94	97.87	97.59
DET	99.29	93.29	98.75	95.68	94.08	98.88	96.74	100.00	87.79	95.72
INTJ	80.00	75.82	99.49	71.43	90.22	n/a	87.65	71.43	47.58	74.88

43 The Croatian standard corpus was chosen because no language-specific relation subtype appeared among the 12 most frequent relations, thus ensuring a cross-linguistically valid comparison.

UPOS tag	Accuracy									
	sl-st	sl-nonst	sl-spok	hr-st	hr-nonst	sr-st	sr-nonst	mk-st	bg-st	Average
NOUN	98.88	93.75	98.40	98.33	93.98	99.23	97.66	99.55	98.53	97.49
NUM	99.74	98.41	87.78	98.87	100.00	98.71	100.00	100.00	98.17	99.24
PART	99.46	95.12	98.59	85.16	90.64	94.12	89.39	90.16	79.94	90.50
PRON	99.47	97.25	97.53	98.68	98.19	97.64	98.47	98.84	99.15	98.46
PROPN	98.71	78.23	96.45	93.65	77.81	97.31	83.68	97.97	98.14	90.69
PUNCT	100.00	99.79	100	100.00	99.73	100.00	99.82	100.00	100.00	99.92
SCONJ	99.78	97.99	100	95.72	94.79	99.52	98.25	94.70	99.61	97.55
SYM	100.00	99.85	n/a	90.91	99.10	100.00	99.38	n/a	n/a	98.21
VERB	97.05	94.12	95.98	99.30	97.84	99.18	98.76	99.74	96.79	97.85
X	59.13	75.67	83.53	77.15	80.10	43.33	62.86	n/a	0.00	56.89

Source: Own work

The highest accuracies among UPOS tags are generally found with tags that represent function word classes, such as *AUX* (auxiliaries), *ADP* (adpositions), and *PRON* (pronouns), and closed-class tags, such as *PUNCT* (punctuation) and *SYM* (symbols), which are handled by the pipeline, inter alia, through rules in the tokenizer, as described in Section Differences Between CLASSLA-Stanza and Stanza. Conversely, the lowest accuracies are found with the infrequent *INTJ* tag (interjections)—of which there were only 5 instances in total in the Slovenian standard test dataset and no instances at all in the Serbian standard test dataset—and the loosely delineated *X* tag, which is used for certain abbreviations,[44] URLs, foreign language tokens, and everything else that does not fit into any of the other categories.

Table 6: Table of per-relation accuracies for the 12 most frequent UD relations in the hr500k corpus. The relations are sorted by decreasing frequency in the hr500k corpus. "sl-spok" refers to the Slovenian spoken dependency parser.

UD relation	Accuracy					
	sl	sl-spok	hr	sr	bg	Average
punct	99.97	100	100.00	100.00	99.91	99.98
amod	98.43	97.96	95.97	97.38	98.66	97.66
case	99.71	99.73	99.32	99.21	99.86	99.51
nmod	92.95	86.29	91.22	90.99	91.49	91.61
nsubj	90.85	85.29	93.39	94.30	91.10	92.32

44 Within the UD system, abbreviations are usually marked with the part-of-speech category that the unabbreviated form falls under. However, for some languages, such as Slovenian, certain types of abbreviations are given the X tag. One such example is "dr.", the abbreviated form of the title "doktor".

UD relation	Accuracy					
	sl	sl-spok	hr	sr	bg	Average
obl	91.64	87.89	85.31	87.24	77.17	85.43
conj	92.24	83.80	90.92	93.06	93.95	92.61
root	92.75	85.71	94.98	95.77	95.97	94.97
obj	91.06	89.03	82.84	91.39	90.18	89.44
aux	99.74	99.34	97.88	97.57	90.46	96.35
cc	98.16	95.89	97.63	97.96	99.14	98.14
advmod	97.10	94.67	93.58	91.82	97.91	95.01

Source: Own work

A similar trend is found among the UD syntactic relations. Relations such as *case* (which usually connects nominal heads with adpositions), *cc* (connects conjunct heads with coordinating conjunctions), and *aux* (connects verbal heads with auxiliary verbs) are used for fixed grammatical patterns that permit little variation. These display consistently high accuracies across all languages. Somewhat lower accuracies are displayed by the *obl* relation, mostly used for oblique nominal arguments which play a less central role in the sentence structure than the core verbal arguments. It has been found that previous versions of dependency parsing models for CLASSLA-Stanza often incorrectly assigned the *obj* relation (used for direct objects) to instances which should receive the *obl* relation and vice versa.[45] Upon inspection of the outputs produced by the newly-trained Slovenian and Croatian parsers, it was found that this error persists also in the current version, which is a likely reason for the performance drops of the *obl* and *obj* relations in other languages as well.

The Slovenian spoken parsing model noticeably stands out, as there is a clear drop with the *nmod, nsubj, obl, conj, root,* and to some degree also the *obj* relations. A subsequent inspection of the model's predictions in these cases revealed that these errors can be ascribed to the highly fragmentary nature of spoken language, causing the model to produce errors when trying to annotate fragments for which the exact role in the wider sentence structure is more difficult to determine unambiguously. This prompted the question of how the performance of the Slovenian spoken models compares to that of the standard and nonstandard models when tasked with annotating transcripts of spoken language. We explore this question in the following section.

45 Kaja Dobrovoljc et al, "Universal Dependencies za slovenščino: nove smernice, ročno označeni podatki in razčlenjevalni model" [Universal Dependencies for Slovenian: New Guidelines, manually annotated data, and parsing model], *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave* 11, No. 1 (2023): 218–46, https://doi.org/10.4312/slo2.0.2023.1.218-246.

Model performance on spoken data

Even though transcripts of spoken Slovenian, manually annotated on the levels of morphosyntactic tagging, lemmatization, and dependency parsing, have been available for quite some time already in the form of the Spoken Slovenian UD Treebank,[46] this resource has recently received a considerable upgrade,[47] and it now contains more than twice the amount of data than what was available in its previous editions. This new version of the resource, which is included in the 2.16 release of the Universal Dependencies treebank collection[48] and the ROG training corpus,[49] served as a basis on which new models specifically adapted to processing spoken language transcripts were trained for the first time and included in the 2.2 release of the CLASSLA-Stanza annotation pipeline.

However, it remained to be tested whether the newly trained models truly offer any considerable advantages to the already available standard and nonstandard models when used to annotate transcripts of spoken Slovenian. To investigate this, all three models currently available for Slovenian were evaluated on the test set of the Spoken Slovenian Treebank. The models were evaluated on the levels of morphosyntactic tagging, lemmatization, and UD dependency parsing. The results of this comparison experiment are shown in Table 7.[50]

Table 7: Comparison of the performance of the spoken, standard, and nonstandard Slovenian models on spoken language data for various annotation levels. The best performing model scores are given in bold.

Model	Morphosyntactic tagging	Lemmatization	UD dependency parsing
Spoken	95.6	99.23	81.91
Standard	90.08	98.68	69.81
Nonstandard	90.46	98.35	n/a

Source: Own work

46 Dobrovoljc and Nivre, "The Universal Dependencies Treebank of Spoken Slovenian."

47 Kaja Dobrovoljc, "Extending the Spoken Slovenian Treebank," paper presented at the *Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, 2024, https://doi.org/10.5281/zenodo.13936394. Jaka Čibej and Tina Munda, "Metoda polavtomatskega popravljanja lem in oblikoskladenjskih oznak na primeru učnega korpusa govorjene slovenščine ROG" [A Method for Semi-automatic Corrections of Lemmas and Morphosyntactic Tags: The Case of the ROG Training Corpus of Spoken Slovene], paper presented at the *Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, 2024, https://doi.org/10.5281/zenodo.13936390.

48 Daniel Zeman et al., "Universal Dependencies 2.16," 2025, http://hdl.handle.net/11234/1-5901.

49 Darinka Verdonik et al., "Training corpus of spoken Slovenian ROG 1.0," 2024, http://hdl.handle.net/11356/1992.

50 We summarize the performance of the morphosyntactic tagging model using the micro F1 score for all three types of morphosyntactic labels combined (UPOS, XPOS, and UFeats), for the lemmatization model using the micro F1 score for all lemmas, and for the dependency parsing model using the micro F1 of the commonly employed labelled attachment score, or LAS score. The LAS score gives the percentage of tokens with both a correctly assigned head token and a correctly assigned dependency label.

The results show that the spoken models clearly outperform the other two sets of models in all three tasks. This increase in performance is particularly evident with the dependency parsing models, suggesting that the differences between spoken and written language in Slovenian are most pronounced at the syntactic level, which is a finding also reported by Dobrovoljc and Čibej.[51] It therefore appears that the training of dedicated models for spoken language annotation was justified and should in the future be expanded to other languages supported by CLASSLA-Stanza as well.

To facilitate the use of spoken models, a special *spoken* processing mode was added to the Slovenian pipeline in version 2.2 that combines the standard tokenizer and spoken models for all subsequent layers of annotation.

Model performance on web data

The model evaluations described in Section Model performance on UPOS and UD labels provide a good summary of how well the CLASSLA-Stanza pipeline performs on both purely standard and purely nonstandard data. However, modern corpus construction techniques—especially for low-resource languages—often rely on crawling data from online conversations, articles, blogs, etc.,[52] which typically consist of a mixture of different language styles and varieties. To illustrate how well the new CLASSLA-Stanza models handle language originating from the internet, this section provides a brief manual qualitative analysis of their performance on a corpus of web data.

The CLASSLA-Stanza tool was used with the newly-trained models to add linguistic annotations to the CLASSLA-web corpora, which consist of texts crawled from the internet domains of the corresponding languages.[53] In preparation for the annotation process, a short test was conducted with the goal of determining which of the two sets of models that primarily handle written language—the standard or the nonstandard—is best suited to be used for annotating the CLASSLA-web corpora. Shorter portions of the corpora were annotated on the levels of tokenization, sentence segmentation, morphosyntactic tagging, and lemmatization, once using the standard and once using the nonstandard model. The two outputs were then compared, and a qualitative analysis of the differences was conducted.

Quite a few of the analysed differences in the model outputs were connected to the processes of sentence segmentation and tokenization. In the CLASSLA-Stanza annotation pipeline, both processes are controlled by the tokenizer. As stated in Section Differences Between CLASSLA-Stanza and Stanza, the pipeline uses two different tokenizers depending on the language and the

51 Kaja Dobrovoljc and Jaka Čibej, "Spoken Slovenian Treebank: New annotated data, parsing models and linguistic insights," paper presented at the *UniDive 3rd General Meeting: Universality, diversity and idiosyncrasy in language technology*, 2025, https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:3rd_unidive_general_meeting:59_spoken_slovenian_treebank_n.pdf.

52 Goldhahn et al., "Corpus collection for under-resourced languages with more than one million speakers."

53 Nikola Ljubešić et al., "Slovenian Web Corpus CLASSLA-web.sl 1.0," 2024, http://hdl.handle.net/11356/1882. Nikola Ljubešić et al., "Croatian Web Corpus CLASSLA-web.hr 1.0," 2024, http://hdl.handle.net/11356/1929.

processing mode used.[54] The analysis showed that sentence segmentation was performed much more accurately by Obeliks and the standard mode of the ReLDI tokenizer. The nonstandard mode of the ReLDI tokenizer appears to have a tendency towards producing shorter segments, since it is optimized for processing social media texts such as tweets. Thus, the nonstandard tokenizer very consistently produces a new sentence after periods, question marks, exclamation marks, and other punctuation, even when these characters do not signify the end of a sentence. The following Croatian example in a simplified CoNLL-U format shows one such case of incorrect sentence segmentation due to the use of reported speech. The original string *„Svaku našu riječ treba da* čuvamo *kao najveće blago."* was split into two segments—the first ending on the period character, while the quotation mark was moved to a separate sentence:

```
# newpar id = 76
# sent_id = 76.1
# text = „ Svaku našu riječ treba da čuvamo kao
najveće blago.
1 „
2 Svaku
3 našu
4 riječ
5 treba
6 da
7 čuvamo
8 kao
9 najveće
10 blago
11 .

# sent_id = 76.2
# text = "
1 "[55]
```

The nonstandard models handled nonstandard word forms quite a bit better than the standard models. Particularly problematic for the standard Slovenian models were forms with missing diacritics, such as "sel" instead of šel, "cist" instead of čisto, "hoce" instead of *hoče*, and "clovek" instead of človek. These were often assigned incorrect lemmas and morphosyntactic tags. An example of the standard lemmatiser output for the word form "hoce" (which corresponds to *hoče* in standard Slovenian (Eng. "he/

54 The ReLDI tokenizer can be used in two different settings: standard and nonstandard. The Obeliks tokenizer, on the other hand, only supports tokenization of standard text.

55 This particular example is found in the CLASSLA-Web.hr corpus at the sentence ID CLASSLA-web.hr.4158219.39.1.

she/it wants")) is displayed below. The model invents a nonexistent lemma "hocati", while the correct form should be the standard Slovenian *hoteti*:

```
# sent_id = 53.1
# text = lev je lev pa naj govori kar kdo hoce
1 lev                 lev
2 je          biti
3 lev                 lev
4 pa                  pa
5 naj                 naj
6 govori      govoriti
7 kar                 kar
8 kdo                 kdo
9 hoce                hocati
```
[56]

Nonstandard forms which do not differ much from their standard counterparts, such as "zdej" as opposed to "zdaj" and "morš" as opposed to "moraš", were generally handled well by both sets of models and did not cause many discrepancies in the outputs.

The analysis of such differences in the model outputs showed that the best results for the web corpus were achieved on the one hand by the standard tokenizer, and on the other by the nonstandard models for all subsequent levels of annotation. In light of this, a new *web* mode was implemented for the CLASSLA-Stanza pipeline. This new mode combines the standard tokenizer and nonstandard models for the other layers in a single package and is intended specifically for the annotation of texts originating on the Internet.

Conclusion

In this paper, we provided an overview of the CLASSLA-Stanza pipeline for linguistic processing of the South Slavic languages and described the training process for the models included in the latest release of the pipeline. We described the main design differences to the Stanza neural pipeline, from which CLASSLA-Stanza arose as a forked project. We provided a summary of the model training process, while the technical documentation[57] should be consulted for a more detailed description of the training process for each language. We also presented per-label performance scores for UPOS labels from standard and nonstandard models and most frequent UD labels from standard models.

56 The sentence ID of this particular example in the CLASSLA-Web.sl corpus is CLASSLA-web.sl.225330.7.1.

57 Terčon and Ljubešić, "CLASSLA-Stanza."

CLASSLA-Stanza gives consistent results across all supported languages and outperforms the Stanza pipeline on all supported NLP tasks, as illustrated in Sections Differences Between CLASSLA-Stanza and Stanza and Model Training. However, low accuracies are still seen for infrequent labels and pairs of labels that are not easily disambiguated. It remains to be seen whether larger and more diverse training datasets can contribute to improving model performance in these specific cases, or rather the move to contextual embeddings, i.e., transformer models. The newly included spoken models perform much better on transcripts of spoken language, and they can be easily deployed using the special *spoken* processing mode implemented within CLASSLA-Stanza. Additionally, when processing texts obtained from the Internet, special care must be taken to use the combination of models that is best suited for the task, which is why we also described the special *web* processing mode.

The release of a specialized pipeline for linguistic processing of South Slavic languages is an important new milestone in the development of digital resources and tools for this relatively under-resourced group of languages. However, there is still much left to be achieved and improved upon. Full support for all annotation tasks and modalities, such as, for instance, semantic role labelling and the spoken modality, remains to be extended to other languages as well. As larger training datasets become available, more capable models can be trained for the currently supported languages. In addition, the aim is also to extend support to other members of the South Slavic language group, provided that training datasets of sufficient size are eventually produced for those languages as well. Finally, the performance of the CLASSLA-Stanza pipeline should also be compared to other recent state-of-the-art tools for automatic linguistic annotation, such as Trankit,[58] which was shown to outperform Stanza over a large number of languages and datasets.

Acknowledgements

The work described by this paper was made possible by the *Development of Slovene in a Digital Environment* project (*Razvoj slovenščine v digitalnem okolju*, project ID: C3340-20-278001), financed by the Ministry of Culture of the Republic of Slovenia and the European Regional Development Fund, the *Language Resources and Technologies for Slovene* research program (project ID: P6-0411), the MEZZANINE project (*Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Languag*e, project ID: J7-4642), the SPOT project (*A Treebank-Driven Approach to the Study of Spoken Slovenian,* Z6-4617), and the *Large Language Models for Digital Humanities* project (Grant GC-0002), all financed by the Slovenian Research Agency, and the CLARIN.SI research infrastructure.

58 Minh Van Nguyen et al., "Trankit: A light-weight transformer-based toolkit for multilingual natural language processing," arXiv preprint (2021), https://doi.org/10.48550/arXiv.2101.03289.

Sources and Literature

Literature

Banón, Marta, Miquel Espla-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, et al. "MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages." Paper presented at the *23rd Annual Conference of the European Association for Machine Translation, (EAMT)*, 2022. https://aclanthology.org/2022.eamt-1.41/.

Čibej, Jaka, and Tina Munda. "Metoda polavtomatskega popravljanja lem in oblikoskladenjskih oznak na primeru učnega korpusa govorjene slovenščine ROG" [A Method for Semi-automatic Corrections of Lemmas and Morphosyntactic Tags: The Case of the ROG Training Corpus of Spoken Slovene]. Paper presented at the *Conference on Language Technologies and Digital Humanities (JT-DH-2024),* 2024. https://doi.org/10.5281/zenodo.13936390.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. "Universal Dependencies." *Computational Linguistics* 47, No. 2 (07 2021): 255–308. https://doi.org/10.1162/coli_a_00402.

Dobrovoljc, Kaja. "Extending the Spoken Slovenian Treebank." Paper presented at the *Conference on Language Technologies and Digital Humanities (JT-DH-2024),* 2024. https://doi.org/10.5281/zenodo.13936394.

Dobrovoljc, Kaja, and Jaka Čibej. "Spoken Slovenian Treebank: New annotated data, parsing models and linguistic insights." Paper presented at the *UniDive 3rd General Meeting: Universality, diversity and idiosyncrasy in language technology*, 2025. https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:general_meetings:3rd_unidive_general_meeting:59_spoken_slovenian_treebank_n.pdf.

Dobrovoljc, Kaja, and Joakim Nivre. "The Universal Dependencies Treebank of Spoken Slovenian." Paper presented at the *Tenth International Conference on Language Resources and Evaluation (LREC`16),* 2016. https://aclanthology.org/L16-1248/.

Dobrovoljc, Kaja, Tomaž Erjavec, and Nikola Ljubešić. "Improving UD processing via satellite resources for morphology." Paper presented at the *Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019),* 2019. https://doi.org/10.18653/v1/W19-8004.

Dobrovoljc, Kaja, Luka Terčon, and Nikola Ljubešić. "Universal Dependencies za slovenščino: nove smernice, ročno označeni podatki in razčlenjevalni model" [Universal Dependencies for Slovenian: New Guidelines, manually annotated data, and parsing model]. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave* 11, No. 1 (2023): 218–46. https://doi.org/10.4312/slo2.0.2023.1.218-246.

Erjavec, Tomaž, Darja Fišer, Simon Krek, and Nina Ledinek. "The JOS Linguistically Tagged Corpus of Slovene." Paper presented at the *Seventh International Conference on Language Resources and Evaluation (LREC`10),* 2010. https://aclanthology.org/L10-1087/.

Erjavec, Tomaž. "MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages." *Language Resources and Evaluation* 46, No. 1 (2012): 131–42. https://doi.org/10.1007/s10579-011-9174-8.

Goldhahn, Dirk, Maciej Sumalvico, and Uwe Quasthoff. "Corpus collection for under-resourced languages with more than one million speakers." Paper presented at the *Collaboration and Computing for Under-Resourced Languages (CCURL)* workshop, 2016. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-CCURL2016_Proceedings.pdf#page=74.

Grčar, Miha, Simon Krek, and Kaja Dobrovoljc. "Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik" [Obeliks: A Statistical Morphosyntactic Annotation and

Lemmatization Tool for the Slovenian Language]. Paper presented at the *Eighth Language Technologies Conference*, 2012. https://doi.org/10.5281/zenodo.14165686.

Krek, Simon, Polona Gantar, Kaja Dobrovoljc, and Iza Škrjanec. "Označevanje udeleženskih vlog v učnem korpusu za slovenščino" [Annotating Semantic Roles in a Training Corpus for Slovenian]. Paper presented at the *Conference on Language Technologies and Digital Humanities (JT-DH-2016)*, 2016. https://doi.org/10.5281/zenodo.14165095.

Ljubešić, Nikola, and Kaja Dobrovoljc. "What Does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian." Paper presented at the *7th Workshop on Balto-Slavic Natural Language Processing*, 2019. https://doi.org/10.18653/v1/W19-3704.

Ljubešić, Nikola, Tomaž Erjavec, Maja Miličević Petrović, and Tanja Samardžić. "Together we are stronger: Bootstrapping language technology infrastructure for South Slavic languages with CLARIN. SI." In *CLARIN. The Infrastructure for Language Resources*, edited by Darja Fišer and Andreas Witt. De Gruyter, 2022. https://doi.org/10.1515/9783110767377-017.

Ljubešić, Nikola, Luka Terčon, and Kaja Dobrovoljc. "CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages." Paper presented at the *Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, 2024. https://doi.org/10.5281/zenodo.13936406.

Osenova, Petya, and Kiril Simov. "Universalizing BulTreeBank: A Linguistic Tale about Glocalization." Paper presented at the *5th Workshop on Balto-Slavic Natural Language Processing*, 2015. https://aclanthology.org/W15-5313/.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." Paper presented at the *58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. https://doi.org/10.18653/v1/2020.acl-demos.14.

Samardžić, Tanja, Nikola Ljubešić, and Maja Miličević. "Regional Linguistic Data Initiative (ReLDI)." Paper presented at the *5th Workshop on Balto-Slavic Natural Language Processing*, 2015. https://aclanthology.org/W15-5306/.

Terčon, Luka, and Nikola Ljubešić. "CLASSLA-Stanza: The next step for linguistic processing of South Slavic Languages." *arXiv preprint* (2023). https://doi.org/10.48550/arXiv.2308.04255.

Van Nguyen, Minh, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. "Trankit: A light-weight transformer-based toolkit for multilingual natural language processing." *arXiv preprint* (2021). https://doi.org/10.48550/arXiv.2101.03289.

Online sources

Arhar Holdt, Špela, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej, Eva Pori, et al. "Training Corpus SUK 1.0." 2022. http://hdl.handle.net/11356/1747.

Arhar Holdt, Špela, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej, Eva Pori, et al. "Training Corpus SUK 1.1." 2024. http://hdl.handle.net/11356/1959.

Batanović, Vuk, Nikola Ljubešić, Tanja Samardžić, and Tomaž Erjavec. "Serbian Linguistic Training Corpus SETimes.SR 2.0." 2023. http://hdl.handle.net/11356/1843.

Erjavec, Tomaž, Ana-Maria Barbu, Ivan Derzhanski, Ludmila Dimitrova, Radovan Garabík, Nancy Ide, Heiki-Jaan Kaalep, et al. "MULTEXT-East '1984' Annotated Corpus 4.0." 2010. http://hdl.handle.net/11356/1043.

Lenardič, Jakob, Jaka Čibej, Špela Arhar Holdt, Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Katja Zupan, and Kaja Dobrovoljc. "CMC Training Corpus Janes-Tag 3.0." 2022. http://hdl.handle.net/11356/1732.

Ljubešić, Nikola and Biljana Stojanovska. "Macedonian Linguistic Training Corpus SETimes.MK 0.1." 2023. http://hdl.handle.net/11356/1886.

Ljubešić, Nikola, Taja Kuzman, Tomaž Erjavec, and Petya Osenova. "Tour de CLARIN: The CLARIN Knowledge Centre for South Slavic Languages (CLASSLA)." CLARIN. Published 18 November, 2021. https://www.clarin.eu/blog/tour-de-clarin-clarin-knowledge-centre-south-slavic-languages-classla.

Ljubešić, Nikola, and Tanja Samardžić. "Croatian Linguistic Training Corpus Hr500k 2.0." 2023. http://hdl.handle.net/11356/1792.

Ljubešić, Nikola, Peter Rupnik, and Taja Kuzman. "Croatian Web Corpus CLASSLA-web.hr 1.0." 2024. http://hdl.handle.net/11356/1929.

Ljubešić, Nikola, Peter Rupnik, and Taja Kuzman. "Slovenian Web Corpus CLASSLA-web.sl 1.0." 2024. http://hdl.handle.net/11356/1882.

Ljubešić, Nikola, Tomaž Erjavec, Vuk Batanović, Maja Miličević, and Tanja Samardžić. "Croatian Twitter Training Corpus ReLDI-NormTagNER-Hr 3.0." 2023. http://hdl.handle.net/11356/1793.

Ljubešić, Nikola, Tomaž Erjavec, Vuk Batanović, Maja Miličević, and Tanja Samardžić. "Serbian Twitter Training Corpus ReLDI-NormTagNER-Sr 3.0." 2023. http://hdl.handle.net/11356/1794.

Terčon, Luka, and Nikola Ljubešić. "Word Embeddings CLARIN.SI-Embed.Hr 2.0." 2023. http://hdl.handle.net/11356/1790.

Terčon, Luka, and Nikola Ljubešić. "Word Embeddings CLARIN.SI-Embed.Sr 2.0." 2023. http://hdl.handle.net/11356/1789.

Terčon, Luka, and Nikola Ljubešić. "Word Embeddings CLARIN.SI-Embed.Mk 2.0." 2023. http://hdl.handle.net/11356/1788.

Terčon, Luka, and Nikola Ljubešić. "Word Embeddings CLARIN.SI-Embed.Bg 1.0." 2023. http://hdl.handle.net/11356/1796.

Terčon, Luka, Nikola Ljubešić, and Tomaž Erjavec. "Word Embeddings CLARIN.SI-Embed.Sl 2.0." 2023. http://hdl.handle.net/11356/1791.

Verdonik, Darinka, Kaja Dobrovoljc, Peter Rupnik, Nikola Ljubešić, Simona Majhenič, Jaka Čibej, and Thomas Schmidt. "Training corpus of spoken Slovenian ROG 1.0." 2024. http://hdl.handle.net/11356/1992.

Zeman, Daniel, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Jephtey Adolphe, Noëmi Aepli, Hamid Aghaei, et al. "Universal Dependencies 2.16." 2025. http://hdl.handle.net/11234/1-5901.

Zupan, Katja, Nikola Ljubešić, and Tomaž Erjavec. "Smernice Janes-NER za označevanje imenskih entitet v slovenskem jeziku: Različica 1.1." CJVT Wiki. Accessed 2 February, 2025. https://wiki.cjvt.si/books/08-imenske-entitete/page/oznacevalne-smernice.

Žitnik, Slavko, and Frenk Dragar. "SloBENCH Evaluation Framework." 2021. http://hdl.handle.net/11356/1469.

Luka Terčon, Kaja Dobrovoljc, Nikola Ljubešić

CLASSLA-STANZA: NASLEDNJI KORAK ZA JEZIKOVNO PROCESIRANJE JUŽNOSLOVANSKIH JEZIKOV

POVZETEK

Predstavljamo CLASSLA-Stanza, orodje za učinkovito jezikovno obdelavo besedil v naravnem jeziku, ki podpira več južnoslovanskih jezikov in je posebej prilagojeno zanje. Najnovejša različica orodja podpira obdelavo besedil v slovenščini, hrvaščini, srbščini, bolgarščini in makedonščini. Orodje je nastalo kot razvejitev cevovoda Stanza za jezikovno obdelavo, z vrsto novih izboljšav. V tem članku opisujemo glavne razlike med CLASSLA-Stanza in Stanza, podajamo pregled procesa usposabljanja za izdelavo modelov, vključenih v najnovejšo različico orodja 2.2, podajamo pregled zmogljivosti najnovejših modelov in razpravljamo o učinkovitosti cevovoda za označevanje govorjenih besedil in besedil, ki izvirajo z interneta.

CLASSLA-Stanza ohranja večino arhitekture in zasnove Stanza, vendar uvaja nekaj ključnih sprememb, med drugim specializiran niz pravilnih tokenizatorjev, ki obdelujejo segmentacijo in tokenizacijo stavkov, podporo za uporabo zunanjih fleksijskih leksikonov kot dodatnega kontrolnega elementa med napovedovanjem, specializiran način obdelave besed zaprtega razreda in podporo za več dodatnih jezikovnih različic, kot so nestandardni jezik, govorjeni jezik in besedila, pridobljena z interneta.

Splošni proces usposabljanja modela za CLASSLA-Stanza sledi zaporednemu postopku, pri katerem se usposabljanje izvaja na predhodno tokeniziranih podatkih, model pa se usposablja za vsako plast označevanja. Po usposabljanju modela za vsako plast se ta model uporabi za samodejno generiranje navzgornjih označb za validacijske in testne podatkovne nize, ki se nato uporabijo med usposabljanjem in ocenjevanjem modelov na naslednjih plasteh označevanja. Med zadnjim krogom usposabljanja modelov so bili modeli najprej usposobljeni za morfosintaktično označevanje, nato za lematizacijo in nazadnje za razčlenjevanje odvisnosti. Med usposabljanjem modela za lematizacijo je bil modulu za usposabljanje na voljo fleksijski leksikon, medtem ko slovenščina podpira tudi uporabo fleksijskega leksikona med usposabljanjem morfosintaktičnega označevalca. Usposabljanje nestandardnih in govornih modelov je potekalo po podobnem postopku z nekaj manjšimi odstopanji.

Predstavljamo ocene natančnosti modelov za vse jezike na različnih oznakah v naborih oznak UD Part-of-Speech in UD dependency relation. Pri oznakah Part-of-Speech so najvišje natančnosti na splošno ugotovljene pri razredih funkcionalnih besed, najnižje pa pri redki oznaki INTJ in oznaki X, ki je slabo opredeljena. Pri odvisnostnih odnosih odnosi case, cc in aux kažejo dosledno visoko natančnost v vseh jezikih, medtem ko nekoliko nižjo natančnost kaže odnos obl, ki se večinoma

uporablja za posredne nominalne argumente, ki imajo v stavčni strukturi manj osrednjo vlogo kot glavni verbalni argumenti.

Prav tako ponujamo analizo na novo usposobljenih govornih modelov, ki so na voljo za slovenski jezik v najnovejši različici CLASSLA-Stanza. Primerjali smo slovenske govorne morfosintaktične modele označevanja, lematizacije in odvisnostnega razčlenjevanja z zmogljivostjo ustreznih standardnih in nestandardnih modelov pri označevanju transkriptov govorjenega jezika. Rezultati kažejo, da govorni modeli znatno presegajo standardne in nestandardne modele. Usposabljanje modelov, ki so posebej prilagojeni govorjenemu jeziku, je bilo zato upravičeno in bi ga bilo treba v prihodnosti razširiti na druge jezike.

Opravili smo tudi analizo, da bi ocenili, kako dobro modeli CLASSLA-Stanza obdelujejo besedila, ki izvirajo z interneta. Ugotovili smo, da so standardni tokenizatorji bolje obdelovali segmentacijo stavkov, nestandardni modeli pa nestandardne oblike, ki se pojavljajo v internetnem jeziku, na vseh drugih ravneh označevanja. Posledično je bil implementiran nov način obdelave spleta, ki je vključen v najnovejšo različico procesa.

1.01 DOI: https://doi.org/10.51663/pnz.65.3.06

Jaka Čibej,* Tina Munda♦

Leveraging a Morphological Lexicon for a Semi-Automatic Approach to Correcting Lemmas and Morphosyntactic Tags

IZVLEČEK

UPORABA OBLIKOSLOVNEGA LEKSIKONA PRI POLAVTOMATSKEM PRISTOPU K POPRAVLJANJU LEM IN OBLIKOSKLADENJSKIH OZNAK

V prispevku predstavljamo nov polavtomatski pristop k popravljanju lem in oblikoskladenjskih oznak. Za razliko od predhodnih pristopov k ročnemu označevanju slovenskih korpusov nova metoda vsebuje dodaten korak, v katerem pojavnice ter njihove strojno pripisane leme in oblikoskladenjske oznake navzkrižno primerjamo z naborom oblik v Slovenskem oblikoslovnem leksikonu Sloleks. Na podlagi primerjave vsako pojavnico uvrstimo v enega od označevalnih scenarijev. Novi pristop občutno zmanjša količino časa in sredstev, ki jih je treba vložiti v označevanje, tako da odstrani veliko število odvečnih označevalnih nalog. Med prednostmi te metode je tudi možnost, da označevalne naloge razdelimo v sklope s podobnimi označevalnimi problemi (npr. razločevanje slovničnih enakopisnic). Ob ustrezni pripravi podatkov lahko metoda tudi drastično zmanjša potrebo po tem, da se označevalci seznanijo z obširnim označevalnim sistemom Multext-East za slovenščino, kar je v sorodnih označevalnih kampanjah predstavljalo ozko grlo. Metodo smo preizkusili med označevanjem Učnega korpusa govorjene slovenščine ROG. Algoritem pripisovanja označevalnih scenarijev preizkusimo tudi na Učnem korpusu pisne slovenščine SUK, ki je bil označen s tradicionalnim označevalnim

* **Phd, Research Associate, University of Ljubljana, Faculty of Arts, Centre for Language Resources and Technologies, Aškerčeva 2, SI-1000 Ljubljana, jaka.cibej@ff.uni-lj.si; ORCID: 0000-0002-3037-6848**

♦ **Junior Researcher, University of Ljubljana, Faculty of Arts, Centre for Language Resources and Technologies, Aškerčeva 2, SI-1000 Ljubljana, tina.munda@cjvt.si; ORCID: 0009-0001-1152-7823**

pristopom (poved za povedjo, pojavnica za pojavnico). Predstavimo rezultate primerjave in zagovarjamo, da bi bilo metodo treba uporabiti pri prihodnjih označevalnih kampanjah, da z njo prihranimo čas in stroške ter nasploh izboljšamo doslednost označevanja, pri čemer razpravljamo tudi o nekaterih slabostih in pasteh predlaganega pristopa.

Ključne besede: lematizacija, oblikoskladenjsko označevanje, govorjena slovenščina, korpusi govorjene slovenščine, ročno označeni korpusi

ABSTRACT

In the paper, we present a new semi-automatic approach to correcting lemmas and morphosyntactic tags. Unlike previous manual annotation approaches for Slovene corpora, the new method contains an additional step in which tokens and their automatically assigned lemmas and morphosyntactic tags are cross-referenced with the set of forms included in the Sloleks Morphological Lexicon of Slovene. Based on the comparison, each token is classified into one of several annotation scenarios. The new approach has noticeably reduced the time and resources invested into annotation by eliminating a large number of redundant tasks. The advantages of this method include the possibility of dividing annotation tasks into groups consisting of similar annotation problems (e.g. disambiguation of grammatical homographs). With adequate data preparation, it also drastically reduces the necessity for annotators to be familiar with the extensive Multext-East morphosyntactic tag set for Slovene, a restriction that created a bottleneck in the annotation process in similar annotation campaigns. The method was tested during the annotation process for the ROG Training Corpus of Spoken Slovene. In addition, we also test the scenario classification algorithm on the SUK Training Corpus of Written Slovene, which was annotated using the traditional sentence-by-sentence, token-by-token approach. We present the results and argue that the method should be used in future annotation campaigns to save resources and improve overall annotation consistency, while also discussing some of the caveats and disadvantages of the proposed approach.

Keywords: lemmatization, morphosyntactic tagging, training corpora, morphological lexicon, corpus annotation

Introduction

The latest tools and models for lemmatization and morphosyntactic tagging of Slovene have achieved impressive results, with the latest performances of CLASSLA-Stanza[1] amounting to an F1-score of 99.11 for lemmatization[2] and 98.27 for morphosyntactic tagging.[3] However, automatic processing is not sufficient when compiling high-quality training corpora or other benchmark datasets. Manual corrections are required, particularly if the models are applied to texts of a different genre or medium compared to what the models were trained on. The CLASSLA-Stanza models for Slovene were trained mostly on written texts, so their application on transcriptions of spoken Slovene yields less accurate results.

In recent years, two projects have highlighted the need for a high-quality training corpus dedicated to spoken Slovene, similar to the *SUK Training Corpus of Written Slovene*.[4] The MEZZANINE[5] project focuses on the development of open-access resources for spoken Slovene. Among other goals, the project aims to provide datasets annotated with speech acts and disfluencies. At the same time, one of the goals of the SPOT[6] project[7] is to compile a corpus of spoken Slovene manually annotated with dependency relations. The joint efforts of both projects thus jumpstarted the compilation of the ROG Training Corpus of Spoken Slovene.[8] However, the compilation of a training corpus of spoken Slovene along the lines of SUK requires manual corrections of annotations for lemmas and morphosyntactic tags, which can be a cumbersome and complex task that traditionally requires a large investment in time and resources with a relatively low cost-benefit (more on this in Section Related Work), even despite the fact that the planned size of ROG was relatively manageable (100,000 tokens in ROG compared to 1,000,000 tokens in SUK).

To facilitate the annotation process, a new method was developed. It adds an additional preprocessing phase before manual annotation: all tokens are first

1 Nikola Ljubešić and Kaja Dobrovoljc, "What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian," *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (Florence, Italy: Association for Computational Linguistics, 2019), 29–34.

2 Luka Terčon, Jaka Čibej, and Nikola Ljubešić, "The CLASSLA-Stanza model for lemmatisation of standard Slovenian 2.0," *Slovenian language resource repository "CLARIN.SI*, ISSN 2820-4042 (2023), http://hdl.handle.net/11356/1768.

3 Nikola Ljubešić, Luka Terčon, and 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.

4 Špela Arhar Holdt, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej et al., "Training corpus SUK 1.1," *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042 (2024), http://hdl.handle.net/11356/1959.

5 MEZZANINE (*Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language*, J7-4642, 2022–2025), https://mezzanine.um.si/.

6 SPOT (*Treebank-Driven Approach to the Study of Spoken Slovenian*, Z6-4617; 2022–2024), https://spot.ff.uni-lj.si/.

7 Kaja Dobrovoljc, "Skladenjska drevesnica govorjene slovenščine: stanje in perspektive," *Stanje in perspektive uporabe govornih virov v raziskavah govora* (2024): 41–62.

8 Darinka Verdonik, Kaja Dobrovoljc, Peter Rupnik, Nikola Ljubešić, Simona Majhenič, Jaka Čibej, and Thomas Schmidt, "Training corpus of spoken Slovenian ROG 1.0," *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042, (2024), http://hdl.handle.net/11356/1992.

cross-referenced with the *Sloleks Morphological Lexicon of Slovene*.[9] The annotation data is then divided into several packages that focus on similar annotation problems (e.g. discrimination between different cases). This approach drastically accelerates the annotation process, improves the consistency of annotation decisions, and reduces the number of redundant reviews (e.g. by skipping unambiguous units) and total annotation costs.

This paper is an extended version of a previous paper in Slovene.[10] In this version, we provide a more detailed description of the approach (Section Methodology). We focus less on the Slovene-specific dilemmas and more on the general benefit of the method to make the approach more understandable for the international audience. In addition to the evaluations of the method originally performed on the *ROG Training Corpus of Spoken Slovene*, we also evaluate the method on the *SUK Training Corpus of Written Slovene* (Section Evaluation on the Spoken Slovenian Treebank) to confirm that the method is reliable enough for other potential benchmark datasets. We also perform a more in-depth analysis on the unambiguous tokens from ROG (Section Results), which were skipped in the original paper. We take the first steps toward a more fine-grained analysis of different annotation tasks in terms of their complexity and annotation difficulty (Section First Steps in a Fine-Grained Analysis of Annotation Tasks).

The paper is structured as follows: in Section Related Work, we provide a short overview of related work and describe the experience of past annotation campaigns. In Section Methodology, we present the new semi-automatic approach and the manner of categorizing tokens by annotation scenarios. We continue by describing the data preparation and annotation phases, as well as the evaluation of the method on both ROG and SUK datasets (Section Data and Annotation). In Section Results we describe the results of the annotation on the ROG dataset and compare them with the results of the evaluation. In Section First Steps in a Fine-Grained Analysis of Annotation Tasks, we describe the most frequent annotation tasks in terms of their complexity. We conclude the paper in Section Conclusion with plans for future work.

9 Jaka Čibej, Kaja Gantar, Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec et al., "Morphological lexicon Sloleks 3.0," *Slovenian language resource repository CLARIN.SI* (2022), http://hdl.handle.net/11356/1745.

10 Jaka Čibej and Tina Munda, "Metoda polavtomatskega popravljanja lem in oblikoskladenjskih oznak na primeru učnega korpusa govorjene slovenščine ROG," *Language Technologies and Digital Humanities: Proceedings of the Conference*: 19-20 September 2022 (Ljubljana, Slovenia, 2024), 66–86.

Related Work

The most extensive annotation campaigns on the levels of lemmas and morphosyntactic tags for Slovene were carried out for the training sets JANES-Tag[11] and JANES-Norm[12] as part of the JANES project,[13] and the SUK 1.0 Training Corpus of Slovene[14] and its subcorpora, such as SentiCoref.[15]

In both campaigns, the annotation process was similar: the texts were first automatically tokenized, segmented into sentences, morphosyntactically tagged and lemmatized. Automatic annotations were then manually corrected by a group of annotators and checked by curators who accepted the final decisions. The campaigns from the JANES project used the WebAnno annotation platform,[16] which allows for multiple annotations of the same text by different annotators and facilitates curation in examples of disagreement. For the subcorpora of SUK 1.0, the annotation process took place in Google Sheets.

Both the SUK and JANES campaigns were large-scale and required a great deal of organization and resources in terms of time and human input. The corrections of tokenization, sentence segmentation and normalization of the first part of the JANES-Norm corpus included a total of 11 annotators and took 7 weeks to complete,[17] with a total of 270 hours of annotator work and an additional 45 hours of curation. Lemmatization and morphosyntactic tags for JANES-Tag (also with 11 annotators) was carried out between March 2016 and October 2016.[18] Correcting the SUK corpus[19] with 24 annotators took approximately 4 months. A significant factor contributing to the length of both campaigns was annotator training, which particularly in the case of the Multext-East v6 (MTE-6)[20] morphosyntactic annotation scheme for Slovene requires much preparation and is the reason for a steep learning curve for new annotators. Controlling inter-annotator agreement and curating the final decisions also prolong the process.

11 Tomaž Erjavec, Darja Fišer, Jaka Čibej, and Špela Arhar Holdt, "CMC training corpus JANES-Tag 1.1," *Slovenian language resource repository CLARIN.SI* (2016b), http://hdl.handle.net/11356/1081.

12 Tomaž Erjavec, Darja Fišer, Jaka Čibej, and Špela Arhar Holdt, "CMC training corpus JANES-Norm 1.2," *Slovenian language resource repository CLARIN.SI* (2016a), http://hdl.handle.net/11356/1084.

13 Darja Fišer, Nikola Ljubešić, and Tomaž Erjavec, "The JANES Project: Language Resources and Tools for Slovene User-Generated Content," *Language Resources Evaluation* 54 (2020): 223–46.

14 Arhar Holdt, Krek, Dobrovoljc, Erjavec, Gantar, Čibej et al., "Training corpus SUK 1.1."

15 Eva Pori, Jaka Čibej, Tina Munda, Luka Terčon, and Špela Arhar Holdt, "Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref," *Konferenca Jezikovne tehnologije in digitalna humanistika* (2022): 162–68.

16 Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann, "A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures," *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (Osaka, Japan: The COLING 2016 Organizing Committee, 2016), 76–84.

17 Jaka Čibej, Darja Fišer, and Tomaž Erjavec, "*Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets," Normalisation and Analysis of Social Media Texts (NORMSOME) – LREC 2016* (2016): 5–10.

18 Jaka Čibej, Špela Arhar Holdt, Darja Fišer, and Tomaž Erjavec, "Ročno označeni korpusi JANES za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave," *Viri, orodja in metode za analizo spletne slovenščine* (2018), 44–73.

19 Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Simon Krek et al., "Nadgradnja učnega korpusa ssj550k v SUK 1.0," *Razvoj slovenščine v digitalnem okolju* (2023): 119–56.

20 Multext East v6 Morphosyntactic Specifications for Slovene: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

All the listed campaigns implemented the approach of correcting individual sequential tokens in the text, which is cognitively taxing especially for morphosyntactic annotation, as it requires the annotators to mentally switch between varying problems depending on the part-of-speech of the relevant token. The SentiCoref annotation campaign[21] decided to alleviate this by dividing the annotators into separate groups, each dedicated to the annotation of different parts-of-speech.

The results of the most recent annotation campaign as part of the RSDO project[22] have shown that the accuracy of automatic annotations for Slovene is high enough to forego comprehensive manual reviews and instead rely on semi-automatic approaches that focus on the most problematic annotation dilemmas. For instance, in the SentiCoref corpus, the lemmas of only 1.3% of all tokens were corrected (which is in line with the expected accuracy of the lemmatization model), and only 2.9% of all automatic morphosyntactic tags were changed. The analysis of these corrections has also shown that approximately 25% of all corrections can be attributed to problems discriminating between common and proper nouns (*Delo* vs. *delo*) and disambiguating grammatical homographs (e.g. between the accusative and nominative cases with inanimate masculine nouns).

Methodology

The new annotation process is based on the *Sloleks Morphological Lexicon of Slovene*. In our research, we used version 3.0,[23] particularly the approximately 100,800 manually validated lexemes (their cca. 2,800,000 inflected forms). The *Sloleks* lexicon forms the morphological part of the *Digital Dictionary Database of Slovene*[24] and is the largest open-access machine-readable database of Slovene words. For each lexeme in the lexicon (e.g. *miza* 'table'), all its forms (inflected by case, number, tense, etc.) are listed as well (e.g. *mize* – genitive singular, *mizi* – dative singular, *mizo* – accusative singular), along with their corresponding morphosyntactic tags using the Multext-East v6 (MTE-6) system. In MTE-6, all morphosyntactic features for a given word are listed in a string of symbols (e.g. *Sozei – samostalnik* 'noun', *občni* 'common', *ženski* spol 'feminine', *ednina* 'singular', *imenovalnik* 'nominative').

The proposed method is based on two basic assumptions: (1) for certain tokens in a given corpus, no manual validation of automatic lemmas and morphosyntactic tags is required as these tokens are unambiguous in the lexicon; (2) for some tokens, only lemmas or only morphosyntactic tags need to be manually validated, and even in that case, the set of potential annotation options according to the lexicon is limited.

21 Pori, Čibej, Munda, Terčon, and Arhar Holdt, "Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref."
22 Arhar Holdt, Čibej, Dobrovoljc, Erjavec, Gantar, Krek et al., "Nadgradnja učnega korpusa."
23 Čibej, Gantar, Dobrovoljc, Krek, Holozan, Erjavec et al., "Morphological lexicon Sloleks 3.0."
24 Iztok Kosem, 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*. (2021), 81–83.

Instead of approaching the annotation completely from scratch for each token, a cross-comparison with the lexicon allows the annotator to select from e.g. a set of three options among morphosyntactic tags instead of the full set of approximately 1,900 tags.

The new approach cross-references each token with the forms in the lexicon and checks the following criteria: (a) is the form present in the lexicon? (b) can the analyzed form be assigned a single lemma or multiple different lemmas according to the lexicon? (c) can the combination of the form and the lemma be assigned a single morphosyntactic tag or multiple different morphosyntactic tags according to the lexicon?

Based on the results of the cross-reference, the algorithm assigns a specific annotation scenario to each token. The set of different annotation scenarios is shown in Table 1, and each scenario is described in more detail in the following section.

Table 1: Annotation scenarios

Scenario	Description	Example
1.1.1	single form, single lemma, single tag	zdaj – zdaj – Rsn
1.1.2	single form, single lemma, multiple tag options	slik – slika – Sozdr|Sozmr
1.2	single form, multiple lemma options	lahko – lahek|lahko
1.2.1	single form, disambiguated lemma, single tag	lahko – lahko – Rsn
1.2.2	single form, disambiguated lemma, multiple tag options	lahko – lahek – Ppnzet|Ppnzeo|Ppnsei|
2.1	the form is not present in the lexicon, but the lemma is	/
2.2	neither the form nor the lemma is present in the lexicon; the token needs to be annotated entirely manually	hozentregerji
0	unclassified token	e.g. punctuation, symbols

Source: Own work

Annotation scenarios

Scenario 1.1.1 includes tokens which according to the lexicon can be assigned an unambiguous lemma and a single unambiguous morphosyntactic tag. For instance, the form *zdaj* 'now' only occurs in the lexicon with the lemma *zdaj* and the morphosyntactic tag *Rsn* (adverb, general, positive), so no further disambiguation is required.

In scenario 1.1.2, the combination of the form and the lemma is unambiguous but can be assigned one of multiple morphosyntactic tags. For instance, the form *slik* only occurs under the lemma *slika* 'image' but is a grammatical homograph with either the tag *Sozdr* (noun, common, feminine, dual number, genitive case) or *Sozmr* (noun, common, feminine, plural number, genitive case). The annotation task can thus be limited to the disambiguation between the differing morphosyntactic features (dual vs. plural number).

Scenario 1.2 is only the first step in a chain that includes subscenarios. Scenario 1.2 contains tokens that first require the lemma to be disambiguated; after that, the morphosyntactic tag may require disambiguation as well. For instance, the form *lahko* can be lemmatized either as *lahko* 'may, can' (adverb) or *lahek* 'light, easy' (adjective). If the lemma is disambiguated as *lahko* in 1.2, the combination of form and lemma (*lahko – lahko*) is then again cross-referenced with the lexicon; the algorithm classifies it as scenario 1.2.1, where no further disambiguation of the morphosyntactic tag is required: the form *lahko* with the lemma *lahko* only occurs with the tag *Rsn* (adverb, general, positive). On the other hand, if the lemma is disambiguated as *lahek* in 1.2, the second cross-reference categorizes it as part of scenario 1.2.2: the combination of the form *lahko* and the lemma *lahek* is a grammatical homograph and can be assigned one of four morphosyntactic tags (*Ppnzet*, *Ppnzeo*, *Ppnsei*, *Ppnset*), which differ in gender (feminine vs. neuter) and case (accusative vs. instrumental vs. nominative).

Scenario 2.1 is unlikely when processing automatically annotated data but is useful for consistency checks after manual annotation. It contains tokens where forms are not present in the lexicon, but the assigned lemma is. This occurs either with typos or legitimate variant forms that are not included in the current version of the lexicon. No such examples were found during our analysis.

Scenario 2.2 is the only scenario that requires entirely manual annotation with no automatic suggestions, as it contains tokens where neither the form nor the lemma are included in the lexicon. An example from the *ROG Training Corpus of Spoken Slovene* is the form *hozentregerji* 'suspenders', a noun that is typically used only in colloquial (non-standard) Slovene and is absent from the current version of the morphological lexicon, which is based mostly on data from corpora of written standard Slovene.

The last of the top-level scenarios is 0, which contains tokens that require no manual annotation (such as punctuation symbols). In addition to the main annotation scenarios, it should be noted that the set also includes a number of subscenarios for 1.1.1, 1.1.2, 1.2.1, and 1.2.2. Two additional subcategories exist: M (for mismatch) and L (for lowercase), resulting in subscenarios such as 1.1.1.M, 1.1.1.L, and 1.1.2.M.

The L subcategories are equal to their parent scenarios in terms of criteria, the only difference being that the cross-referencing with the lexicon takes into account the lower-case form of the word. This is particularly useful for words occurring at the beginning of the sentence or utterance, as the title-case version (e.g. the form *Zdaj* 'now') does not occur in the lexicon. Instead of categorizing it directly as an out-of-vocabulary word in scenario 2.2, the algorithm first checks whether it occurs in the lexicon without the capitalization (*zdaj*). The form *Zdaj* is thus classified as part of scenario 1.1.1.L, i.e. a completely unambiguous form if its lower-case version is considered.

The M subcategories include examples where the combination of the form and the lemma is assigned a morphosyntactic tag that is not among the options listed in the lexicon. This occurs in cases where the model annotated the token with a tag not present in the lexicon – an example from the ROG corpus is *samo* 'only', which is listed

only as a particle (L) in Sloleks 3.0 but can also occur as a subordinating conjunction (Vp), particularly in spoken Slovene. The M subcategories are useful for identifying the discrepancies between the automatic tagger (which is based on a training corpus and the morphological lexicon) and the morphological lexicon itself. In addition, the M subcategories are useful for intermediate consistency checks. For instance, if the annotators in the phase of disambiguating lemmas (scenario 1.2) change the lemma from the adverb *odlično* to the adjective *odličen*, the automatic adverbial morphosyntactic tag (*Rsn*) is not included in the set of adjectival tags from the lexicon, and the combination *odlično – odličen – Rsn* is classified as 1.2.2.M, e.g. a form with an ambiguous lemma and multiple morphosyntactic tag options where the current morphosyntactic tag is not included in the lexicon. This is either due to an error in manual annotation or a missing form/tag combination in the lexicon.

An example of a sentence from the ROG Training Corpus in which tokens have been annotated with corresponding scenarios is shown in Table 2.

Table 2: An example of a sentence annotated with scenarios

Form	Lemma	Tag	Scenario
Drage	drag	Ppnzmi	1.2
prijateljice	prijateljica	Sozmi	1.1.2
,	,	U	0
dragi	drag	Ppnmmi	1.2
prijatelji	prijatelj	Sommi	1.1.2
govorjene	govorjen	Pdnzer	1.1.2
slovenščine	slovenščina	Sozer	1.1.2
.	.	U	0

Source: Own work

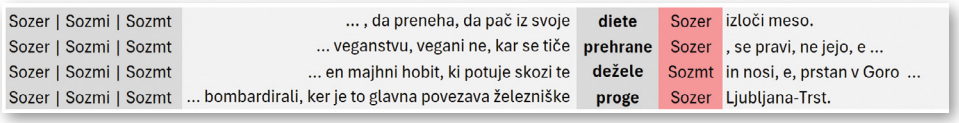
Division into annotation tasks

Based on the assigned annotation scenarios, the tokens from the corpus can then be divided into sets of tasks of varying complexity. Within the same scenario, tokens can be sorted and divided into groups consisting of similar annotation dilemmas (based on the set of morphosyntactic tags available as options from the lexicon).

The annotation tasks may differ somewhat depending on the scenario, but in general, an individual task according to this approach consists of a single token in context and the potential values that can be assigned to it. Figure 1 shows an example of a task in which the annotator is expected to determine whether the listed feminine nouns (focus forms surrounded by their context) occur in the singular genitive (*Sozer*), plural nominative (*Sozmi*) or the plural accusative form (*Sozmt*). In this case, the red column represents the final annotation, while the initial gray column lists all the possible

options from the lexicon (during the annotation of the ROG Training Corpus, several other columns were available to help the annotator – they are presented in more detail in Section Annotation workflow).

Figure 1: Examples of annotation tasks from scenario 1.1.2 (disambiguation of case and number for feminine nouns)



Source: Own image

When annotating the *ROG Training Corpus*, we only used two expert annotators (more on this in Section Data and Annotation), so no custom interface was developed as it was decided that *Microsoft Excel* files would be sufficient for such a small-scale experiment. For larger annotation campaigns, however, it would be sensible to invest more time into developing a user-friendly interface in one of the flexible annotation platforms (such as *PyBossa*[25] or *LabelStudio*[26]), which would further streamline the process and potentially even eliminate the need to train inexperienced annotators with the extensive MTE-6 tagset. A custom interface would also enable real-time consistency checks – any invalid input due to typos or human errors could be checked to ensure maximum annotation consistency.

Another important thing to note with this approach is the paradigm shift from annotating each unit (sentence or utterance) token-by-token (horizontal view) to annotating similar tokens that are part of disparate units but share some of the morphosyntactic features and have the same annotation options (vertical view, similar to the view provided by concordancers when querying corpora). This removes much of the cognitive effort present in the horizontal token-by-token approach, in which the annotator is forced to mentally switch between different parts-of-speech and the corresponding morphosyntactic features (case, gender, number for nouns; aspect and number for verbs, etc.). By grouping similar tasks together, the annotator can focus on a single type of dilemma and resolve it throughout the entire corpus.

Advantages and disadvantages

The proposed approach does pose some disadvantages or at least caveats. First, the method is the most effective if the corpus has already been tokenized and accurately segmented into units. As the annotation method focuses on individual tokens, any changes to tokenization in this approach requires the annotator to add a comment, while the actual changes are done manually by the curator at the end of the campaign. Any changes to tokenization should thus be carried out before annotation scenarios

25 PyBossa, https://docs.pybossa.com/.
26 Label Studio, https://labelstud.io/.

have been assigned. It should be noted, however, that tokenization changes pose a similar problem with the horizontal approach as well.

Another concern is the treatment of multiword expressions. It is possible that the algorithm divides the tokens of a single multiword expression into different scenarios, e.g. *lindy hop*, where *lindy* falls under 2.2 (out-of-vocabulary word) and *hop* falls under 1.1.2 (a grammatical homograph with an unambiguous lemma). In some cases, the annotation of one component greatly depends on the other, so annotators need to pay close attention to such examples, otherwise they may not be annotated consistently.

The systematic division of tokens into scenarios may also result in some lemmatization or tagging errors being lost in the scenarios that require no manual validation, particularly in the case of homographs that are treated as unambiguous in the lexicon, but the language use in the corpus proves they are in fact ambiguous. In ROG, one such example is the form *šalam*, which in the lexicon only occurs with the lemma *šala* 'joke' and the morphosyntactic tag *Sozmd* (noun, common, feminine, plural, dative case). However, in ROG, the form represents the common masculine noun *šalam*, a non-standard variant of the common feminine noun *salama* 'salami'. Because the lexeme *šalam* is missing from the lexicon, the token is mistagged and sorted into the unambiguous scenario, which is incorrect. However, this occurs rarely (see Section Data and Annotation), and the benefits of the new annotation approach far outweigh the disadvantages of a handful of mistagged examples. It should also be noted that with future updates to the lexicon, these types of errors will become even less frequent.

On the other hand, the method provides a number of advantages. First, it cuts down on redundant work as it allows us to skip annotation in the case of unambiguous morphosyntactic tags (this covers as much as 20% of all tokens). Second, when disambiguation is required, the algorithm narrows down the set of annotation options and allows annotators to discriminate among a limited set of tags or features (e.g. disambiguation of cases). This is especially important if instead of full MTE-6 morphosyntactic tags we decide to use morphosyntactic features (singular, dual, plural; nominative, genitive, dative; and so on), which everyone is already familiar with. This removes most of the need for cumbersome annotator training, as well as the need to cross-check multiple annotations to ensure inter-annotator agreement since annotations with simple features (e.g. singular vs. plural) are much easier compared to annotations with full MTE-6 tags.

Another important improvement compared to the horizontal approach concerns updates to annotation guidelines. In the token-by-token and sentence-by-sentence approach, problematic examples were discovered gradually, which often resulted in annotation guidelines being updated and changed more toward the end of the annotation process. This required some additional consistency checks and separate exports of specific tokens for cross-reference. The advantage of the vertical approach is that all similar examples are already grouped and can be analyzed together, which facilitates the updates to annotation guidelines and reduces the waiting time for examples to be collected.

Data and Annotation

In this section, we first briefly present the data included in the ROG Training Corpus of Spoken Slovene, then perform two evaluations of the proposed semi-automatic approach on two existing gold-standard datasets. We describe the division of ROG into annotation scenarios and the annotation workflow.

Contents of the ROG Training Corpus of Spoken Slovene

The data for ROG were sampled from the *GOS Corpus of Spoken Slovene*, versions 1.1[27] (approximately 40,000 tokens) and 2.0[28] (approximately 50,000 tokens). We expected no additional tokenizaton corrections since the data consists of manually transcribed speech that has also been manually segmented into utterances and tokens. The sampling criteria and several other preprocessing steps (such as the unification of segmentation criteria across different subcorpora of GOS) are described in more detail by Verdonik et al. (2024).[29]

A third sample was also included in ROG – the *Spoken Slovenian Treebank*[30] (SST), in which lemmas and morphosyntactic tags had already been manually corrected in a previous endeavor. We used this sample to evaluate the validity of the proposed method (see Section Evaluation on the Spoken Slovenian Treebank).

Evaluation on the Spoken Slovenian Treebank

We were cognizant of the difference between the ROG annotation campaign (which covers spoken Slovene) and all previously conducted campaigns, which focused on either written standard Slovene or (non-standard) internet Slovene. Any insights from previous experience might not be directly transferrable, which is why we first performed an evaluation of the semi-automatic method on the *Spoken Slovenian Treebank* (SST; 30,000 tokens). The division of its manually annotated tokens into annotation scenarios was important to demonstrate how much disagreement (and especially errors) we could expect if we approach the annotation process using the new method. The results of the SST division are shown in Table 3.

27 Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej, and Tomaž Erjavec, "Spoken corpus GOS 1.1," *Slovenian language resource repository CLARIN.SI.* (2021), http://hdl.handle.net/11356/1438.

28 Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej, Tomaž Erjavec, Darinka Verdonik et al., "Spoken corpus GOS 2.0 (transcriptions)," *Slovenian language resource repository CLARIN.SI.* (2023), http://hdl.handle.net/11356/1771.

29 Darinka Verdonik, Nikola Ljubešić, Peter Rupnik, Kaja Dobrovoljc, and Jaka Čibej, "Izbor in urejanje gradiv za učni korpus govorjene slovenščine ROG," *Konferenca jezikovne tehnologije in digitalna humanistika* (2024), 472–88.

30 Kaja Dobrovoljc and Joakim Nivre, "The Universal Dependencies Treebank of Spoken Slovenian," *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (2016): 1566–73.

Table 3: Division of the SST subset into annotation scenarios

Form	Lemma	Tag
1.1.1	8,300	29.12%
1.1.2	11,047	38.76%
1.2	6,234	21.87%
2.2	537	1.88%
1.1.1.L	11	0.04%
1.1.1.M	11	0.04%
1.1.2.L	66	0.23%
1.1.2.M	104	0.36%
0	2,192	7.69%
Total	**28,502**	**100.00%**

Source: Own work

The most problematic tokens are the ones included in the 1.1.1.M scenario. If the corpus were automatically annotated, the algorithm would classify them as 1.1.1 (entirely unambiguous). In reality, they were annotated with a morphosyntactic tag that differs from the options available in the lexicon. Because the method facilitates the annotation process by skipping the unambiguous tokens, the 1.1.1.M tokens would be mistagged in the final version of the corpus. A slightly less problematic scenario is 1.1.2.M, where the tokens have an unambiguous lemma, but multiple lexicon options for morphosyntactic tags (none of which is correct). The annotators would still check all of these tokens, but might be tempted to assign one of the lexicon options instead of opting for the correct tag. Most of these problems stem from inconsistencies or gaps in the lexicon, however, as in the case of the form *gremo*, which is only listed in the lexicon as the first person present plural form of the verb *iti* 'to go' (*Ggvspm*; verb, main, biaspectual, present, first person, plural); in non-standard or spoken Slovene, however, it can also signify the first person imperative plural form (*Ggvvpm*; verb, main, biaspectual, imperative, first person, plural). The SST subset contains only 0.4% of such tokens, however, which indicates that the division into annotation scenarios is accurate enough to be implemented in the annotation of the rest of ROG.

Evaluation on the SUK Training Corpus of written Slovene

We also performed an additional evaluation of the method on the SUK Training Corpus, which had been previously annotated from scratch with a horizontal approach and contains mostly written texts. The division of SUK into annotation scenarios is shown in Table 4. Note that the 1.2 scenario is not further subdivided in this case as it is among the least problematic since all its tokens are included in at least one phase of manual validation.

Table 4: Division of the SUK corpus into annotation scenarios

Scenario	Frequency	Percentage
1.1.1	197,240	19.23%
1.1.1.L	12,120	1.18%
1.1.1.M	474	0.05%
1.1.1.LM	36	<0.01%
1.1.2	453,449	44.21%
1.1.2.L	27,486	2.68%
1.1.2.M	10,447	1.02%
1.1.2.LM	1,818	0.18%
1.2	147,281	14.36%
1.2.L	7,202	0.70%
2.2	24,115	2.35%
0	143,971	14.04%
Total	**1,025,639**	**100.00%**

Source: Own work

The results on the SUK corpus are similar to the evaluation on SST. The most problematic tokens from 1.1.1.M that could potentially be lost in the unambiguous 1.1.1 scenario account for just 0.05% of the entire corpus. The similar, but less problematic 1.1.2.M scenario (along with 1.1.2.LM) is somewhat more frequent compared to SST (1.20% vs. 0.36%), but still within a manageable range, which further confirms that the vertical annotation approach, while certainly less thorough than the horizontal approach, provides a very good compromise between efficiency and accuracy. However, it should be noted that the mismatched annotations should be further analyzed in more detail as they might indicate gaps or inconsistencies in the lexicon that should be filled in to make the method more accurate in the future. For instance, the verb *pojokcati* 'to cry, to complain' is wrongly listed as biaspectual in the lexicon but correctly annotated as perfective in the corpus.

Division of ROG into annotation scenarios

Table 5 shows the division of tokens into scenarios for the other two samples included in ROG (V1 – 10,000 tokens from GOS 1.1; V2 – 50,000 tokens from GOS 2.0). Asterisk symbols (***) mark the second-phase scenarios of scenario 1.2, in which we first disambiguate the lemma, then divide the tokens again into different scenarios.

Table 5: Division of the rest of ROG into annotation scenarios

Scenario	Frequency – V1	Percentage – V1	Frequency – V2	Percentage – V2
1.1.1	3,962	31.31%	10,335	21.25%
1.1.1.L	5	0.04%	54	0.11%
1.1.1.M	2	0.02%	26	0.05%
1.1.2	4,391	34.70%	17,679	36.36%
1.1.2.L	17	0.13%	213	0.44%
1.1.2.M	54	0.43%	737	1.52%
1.2	3,000	23.71%	8,141	16.74%
***1.2.1	1,543	12.19%	3,879	7.98%
***1.2.1.M	22	0.17%	110	0.23%
***1.2.2	1,369	10.82%	4,028	8.28%
***1.2.2.M	66	0.52%	124	0.26%
2.2	233	1.84%	497	1.02%
0	990	7.82%	10,942	22.50%
Total	**12,654**	**100.00%**	**48,624**	**100.00%**

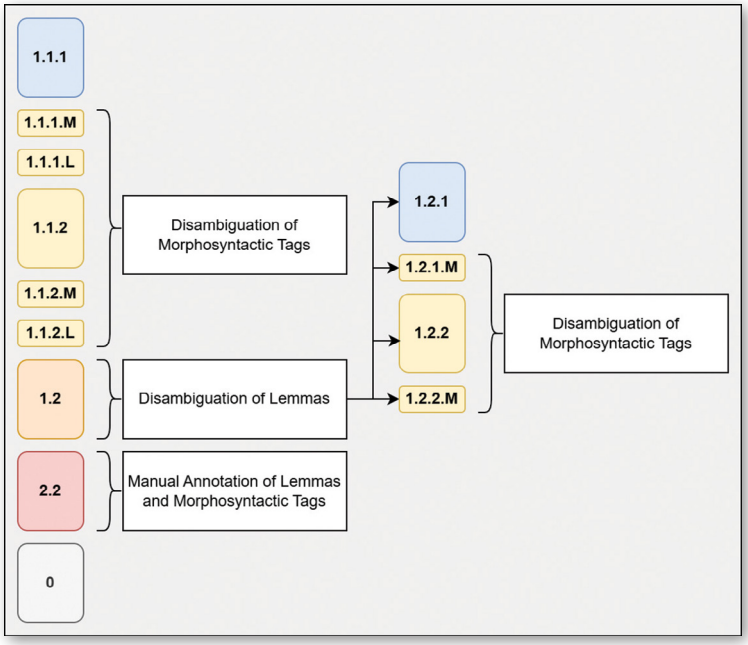
Source: Own work

All the tasks were included in at least one phase of manual annotation, except for scenarios 0 (punctuation), 1.1.1 (unambiguous tokens), and 1.2.1 (tokens that have an unambiguous morphosyntactic tag once the lemma has been disambiguated). Two annotators were used, both involved in previous annotation campaigns and familiar with both the annotation guidelines and the MTE-6 scheme. The first annotator was charged with correcting lemmas, while the second focused on morphosyntactic tags.

Annotation workflow

Figure 2 represents the annotation workflow in ROG. Tokens from different scenarios were included in different review phases. Scenarios 1.1.1 and 0 were skipped entirely. For 1.2.1, only lemmas were disambiguated. For most scenarios and tokens (e.g. 1.1.2, the largest scenario in terms of tokens), only morphosyntactic tags needed to be disambiguated.

Figure 2: Annotation workflow for correcting lemmas and morphosyntactic tags in the ROG corpus



Source: Own image

An example of an annotation task was shown previously in Figure 1, but it should also be noted that the annotation tasks contained some additional information. Besides the short context (up to 5 tokens to each side of the focus token), a separate column contained an extended version of the utterance, as well as a link to the GOS 2.1 corpus in the NoSketchEngine concordancer. In addition, three links to speech recordings from the corpus were listed (the previous segment, the focus segment, and the subsequent segment). The token IDs from the original corpus were also kept in the annotation files to ensure maximum traceability and facilitate the inclusion of the corrections in the final version of the corpus.

Results

In this section, we present the results of the manual corrections of lemmas and morphosyntactic tags using the semi-automatic approach. We also focus more on scenario 1.1.1, which is most at risk for being the source of errors in the final corpus due to lexicon inconsistencies.

Lemma corrections

Lemma corrections were rare – in the end, lemma changes occurred in only 396 tokens in the V2 sample (0.81% of the entire sample) and 175 tokens in the V1 sample (1.38% of the entire sample). Lemma corrections were the most frequent in scenario 2.2 (42% of all lemma corrections), which contains tokens for which neither the form nor the lemma are present in the lexicon. Lower accuracy of the lemmatization model in such examples is expected. In sample V2, the lemma was corrected for 164 tokens (out of 497 in scenario 2.2; 33%). In sample V1, the lemma was corrected for 73 tokens (out of 233 in scenario 2.2; 31%). Approximately a third of out-of-vocabulary tokens in both samples were incorrectly lemmatized. For example, determining the lemma seems to cause problems with proper nouns (*Netflix* – **Netflixu*, *Šerbi* – **Šerba*, *Lidl* – **Lidel*) or nouns with ambiguous morphological patterns, such as the *-j-* lengthening (*espe* – **espej*, *mikronivo* – **mikronivoj*).

On the other hand, words that do appear in the lexicon but are still a significant source of lemma corrections belong to scenario 1.2 (lemma disambiguation) and include problematic homographs (approximately 328 tokens in total, or 57% of all lemma corrections). The most frequent corrections pertain to adjective-adverb disambiguation (*mogoč* 'possible' – *mogoče* 'possibly', *dober* 'good' – *dobro* 'well').

Another noteworthy insight is that in scenario 1.1.2 (disambiguation of morphosyntactic tags), the lemma was changed in only 6 examples, which confirms that separating the lemma disambiguation task and morphosyntactic tagging is a sensible course of action.

Morphosyntactic tag corrections

Corrections of morphosyntactic tags were somewhat more frequent than lemma corrections, but they still account for only a small fraction of tokens. The tag was changed for only 2,029 tokens in the V2 sample (4.17% of the entire sample) and 627 tokens in the V1 sample (4.95% of the sample).

As expected, 1,782 corrections (67.09% of all tag corrections) were made within scenario 1.1.2 (including 1.1.2.M and 1.1.2.L), which is focused on the disambiguation of grammatical homographs with an unambiguous lemma. Similarly, 578 corrections (21.76%) were made within 1.2 (lemma disambiguation) and its subscenarios, where a lemma correction often results in a tag correction as well. Even though only a small percentage of total corrections (296 tokens or 11.15%) were made in 2.2 (out-of-vocabulary tokens), an analysis of the percentage of corrections within 2.2 shows that the V2 sample accounted for 37.83% of corrected tokens and the V1 sample for 46.35% of tokens, meaning that out-of-vocabulary tokens present the most problematic category, even if less frequent compared to grammatical homographs. In other scenarios, this percentage was much smaller (around 7%), which emphasizes the need

for an up-to-date morphological lexicon to ensure maximum accuracy in morphosyntactic tagging.

Table 6 shows the morphosyntactic features of the automatic tags that were most frequently corrected (sorted by frequency). While general adjectives are the most frequent in total, the most problematic features are revealed by the percentages of corrected tokens within each category. In relative terms, the most frequently corrected tokens were proper masculine nouns, which required corrections in cca. 25% of examples. A similar percentage can be observed in cardinal letter numerals and interrogative pronouns. Interestingly, automatic tagging seems to be almost completely unproblematic in the case of verbs, which accounted for only 84 corrections (between 0.5% and 1.3%, depending on the aspect).

Table 6: Morphosyntactic features of the most frequently corrected tokens (with a frequency of at least 100)

Features	Corrected	All Tokens	Percentage
Pp (adjective, general)	384	2,998	12.81%
Som (noun, common, masculine)	281	3,412	8.24%
Soz (noun, common, feminine)	267	3,287	8.12%
Rs (adverb, general)	261	5,103	5.11%
Zk (pronoun, demonstrative)	215	1,860	11.56%
Zo (pronoun, personal)	140	1,341	10.44%
Slm (noun, proper, masculine)	122	473	25.79%
Sos (noun, common, neuter)	110	1,361	8.08%
Kbg (numeral, letter, cardinal)	109	486	22.43%
Vp (conjunction, coordinating)	106	3,265	3.25%
Zv (pronoun, interrogative)	103	497	20.72%

Source: Own work

Table 7 shows the most frequent corrections of morphosyntactic features (with a frequency of at least 50). These account for more than half of all corrections (53%), while almost a third of them (28%) concern the disambiguation between the nominative and the accusative cases (notable grammatical homographs in Slovene).

Table 7: The most frequent corrections of morphosyntactic features (with a frequency of at least 50)

Correction	Frequency	Percentage	Examples
nominative, accusative	561	21.12%	Somei → Sometn (stol), Kbgmi→ Kbg-mt (tisoč), Zk-mei → Zk-met (ta)
accusative, nominative	190	7.15%	Sometn → Somei (video), Zkset → Zk-sei (tisto), Kbg-mt → Kbg-mi (devetsto)
adverb, particle	136	5.12%	Rsn → L (a)
masculine, feminine	122	4.59%	Zotmmt–k → Zotzmt–k (jih), Ppnmmr → Ppnzmr (naslednjih)
nominative plural, genitive singular	82	3.09%	Sozmi → Sozer (preiskave), Ppnzmi → Ppnzer (radijske), Sosmi → Soser (zdravila)
general adjective, general adverb	80	3.01%	Ppnsei → Rsn (mogoče), Ppnzet → Rsn (primerno)
masculine, neuter	67	2.52%	Zotmet–k → Zotset–k (ga), Ppnmeo → Ppnseo (zdravim), Kbvmei → Kbvsei (devetnajststo)
coordinating conjunction, general adverb	64	2.41%	Vp → Rsn (zato)
common, proper	55	2.07%	Somei → Slmei (Piano), Somem → Slmem (Lidlu), Sozer → Slzer (Jute)
interrogative pronoun, general adverb	50	1.88%	Zv-sei → Rsn (kako), Zv-set → Rsn (kaj)

Source: Own work

Analysis of scenario 1.1.1

In our previous paper, we only skimmed through the tokens of scenario 1.1.1 since the evaluations on gold standard datasets (see Sections Evaluation on the Spoken Slovenian Treebank and Evaluation on the SUK Training Corpus of Written Slovene) have shown that only a small fraction of tokens slip through the cracks. Here, we performed a more thorough analysis of those tokens as well.

Only 22 different types account for more than half of approximately 14,400 tokens from scenario 1.1.1 and its subscenarios (see Section Division of ROG into annotation scenarios). These are very frequent functional words such as conjunctions (*pa* 'and', *ki* 'which', *ker* 'because'), particles (*tudi* 'also', *še* 'still'), forms of the auxiliary verb *biti* 'to be'

(*so* 'they are', *bi* 'would'), and adverbs (*zelo* 'very'). While some of these can theoretically occur in another role (for instance, *bi* as a shortened non-standard version of *biseksualen* 'bisexual', *pa* as an interjection in *pa pa* 'bye-bye'), this is very infrequent compared to their predominant context and begs the question of whether it is worth checking an additional 14,000 tokens for a handful of marginal examples. In any case, should the lexicon be updated with these marginal uses, the tokens would end up in a different scenario (e.g. 1.2 or 1.1.2).

The other half of scenario 1.1.1 contains forms that truly are unambiguous. It is practically impossible for them to signify anything else than what is already included in the lexicon, such as the forms *ljudem* (plural dative of the common masculine noun *človek* 'human'), *knjiga* (nominative singular of the common feminine noun *knjiga* 'book'), and *rešitvami* (instrumental plural of the common feminine noun *rešitev* 'solution'). The only example we found in scenario 1.1.1 that is completely mistagged is the already mentioned non-standard form *šalam* 'salami', which was mislemmatized as *šala* 'joke'.

First Steps in a Fine-Grained Analysis of Annotation Tasks

As shown in Section Morphosyntactic tag corrwctions, there seems to be a concentration of frequent corrections in certain morphosyntactic features. However, a closer look shows that in some cases, this pertains to an even narrower type of task: the combination of a specific lemma and its morphosyntactic features. A good example of this is the form *to*, which is lemmatized as *ta* 'this', but needs to be morphosyntactically disambiguated (a choice between four options: feminine+singular+accusative, feminine+singular+instrumental, neuter+singular+nominative, neuter+singular+accusative). In the future, the division into scenarios can be further updated with an even more granular approach to create a list of fine-grained tasks which can then be categorized according to their complexity and difficulty. In any future annotation campaigns, the list can be used to divide the disambiguation tasks between less experienced annotators (or even crowdsourcers) on the one hand (for tasks of lower complexity) and experts on the other. This would allow for a much more sensible division of human resources.

As a first step, we provide the 10 most frequent disambiguation tasks in scenario 1.1.2 (Table 8), annotated with a subjective complexity rating (low, middle, or high complexity) based on the annotator's opinion of how demanding and time-consuming the task is.

Table 8: The most frequent disambiguation tasks from scenario 1.1.2 annotated with complexity ratings

Disambiguation Task and Relevant Forms	Frequency	Complexity
adverb | coordinating conjunction (in, ali, torej, vendar, zato)	1,265	high
noun | preposition, instrumental | preposition, accusative (v)	886	low
nominative | accusative (with singular masculine nouns)	686	low-to-middle
particle | coordinating conjunction (ne, sicer)	635	middle
interjection | preposition, accusative | preposition, locative (na)	612	low
singular genitive | plural nominative | plural accusative (with common feminine nouns)	549	low
feminine singular accusative | feminine singular instrumental | neuter singular nominative | neuter singular accusative (to)	544	middle
singular accusative | singular instrumental (with common feminine nouns)	447	low
preposition, instrumental | preposition, genitive | preposition, accusative | adverb (za)	352	low-to-middle
9 combinations of gender, number, and case (with general adjectives)	345	middle-to-high

Source: Own work

In the future, task complexity can be calculated bottom-up based on the time spent and taking into account the number and types of morphosyntactic features that need to be disambiguated. In this paper, we only provide manual estimations. As shown in Table 8, tasks may vary in complexity even within the same annotation scenario depending on how much context the annotator requires to disambiguate the examples. While low-complexity tasks require a context of only a single word (e.g. the annotator only needs to look at the preceding preposition to determine the case of the noun), high-complexity tasks require a wider context and are more time-consuming (as is the case of disambiguating adverb-conjunction homographs).

Conclusion

In the paper, we presented a new semi-automatic approach to correcting lemmas and morphosyntactic tags using the example of the ROG Training Corpus of Spoken Slovene. The results are encouraging particularly when comparing the expected duration of the annotation campaign using the traditional approach: based on previous experience, lemma and tag annotation for each token takes approximately 12 seconds. In the case of approximately 60,000 tokens for ROG, using 6 annotators, collecting 3 responses per token, and enforcing a 10-hour weekly quota, the campaign would take 9–10 weeks, a total of 500 hours of annotator work (or 160 hours if only a single response per token were collected). This does not include any additional curation and data preparation. For the annotation of ROG, it took 105 hours (25 hours for lemmas and 80 hours for morphosyntactic tags), while the final percentage of corrected tokens is comparable to the traditional approach.

In the future, the method can also be used to identify inconsistencies in previously annotated corpora such as SUK. The scenarios can also be analyzed after an update to the lexicon as any changes may show a potential inconsistency in annotations. Scenarios and more fine-grained tasks could also be useful as potential weighted features for more accurate evaluations of models (e.g. a case error in a grammatical homograph is arguably less serious than an error in part-of-speech).

Another step that can be taken in the future is to implement lexicon updates along with corpus annotation to ensure that both the lexicon and training datasets are synchronized. The annotation process can be made even more efficient by generating a list of rarely problematic low-priority disambiguities (e.g. disambiguating *kaj* 'what', a highly frequent pronoun vs. *kaja*, an archaic word for *kajenje* 'smoking').

We have shown that fully manual approaches to annotating lemmas and morphosyntactic tags can be successfully substituted by a semi-automatic method that offers several additional opportunities for optimization. We will explore these in our future work.

Acknowledgement

The research presented in this paper was carried out as part of the research project *Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language* (MEZZANINE, J7-4642), the research project *Treebank-Driven Approach to the Study of Spoken Slovenian* (SPOT, Z6-4617), and the research program *Language Resources and Technologies for Slovene* (P6-0411), all funded by Slovenian Research and Innovation Agency (ARIS).

The authors would like to thank Matija Škofljanec for lemma corrections and Kaja Dobrovoljc for additional suggestions on how to improve the semi-automatic method presented in this paper. A sincere word of gratitude also goes to the anonymous reviewers for their constructive comments.

Sources and Literature

Arhar Holdt, Špela, Jaka Čibej, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Simon Krek et al. "Nadgradnja učnega korpusa ssj550k v SUK 1.0." *Razvoj slovenščine v digitalnem okolju* (2023): 119–56.

Arhar Holdt, Špela, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej et al. "Training corpus SUK 1.1." *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042 (2024). http://hdl.handle.net/11356/1959.

Čibej, Jaka, and Tina Munda. «Metoda polavtomatskega popravljanja lem in oblikoskladenjskih oznak na primeru učnega korpusa govorjene slovenščine ROG." *Language Technologies and Digital Humanities: Proceedings of the Conference*: 19–20 September 2024. Ljubljana, Slovenia. (2024): 66–86. https://www.sdjt.si/wp/wp-content/uploads/2024/09/JT-DH_2024_Cibej_Munda.pdf.

Čibej, Jaka, Darja Fišer, and Tomaž Erjavec. *Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets. Normalisation and Analysis of Social Media Texts (NORMSOME) – LREC 2016 (2016): 5–10. Portorož, Slovenia*. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NormSoMe_Proceedings.pdf#page=10.

Čibej, Jaka, Kaja Gantar, Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec et al. "Morphological lexicon Sloleks 3.0." *Slovenian language resource repository CLARIN.SI* (2022). http://hdl.handle.net/11356/1745.

Čibej, Jaka, Špela Arhar Holdt, Darja Fišer, and Tomaž Erjavec. "Ročno označeni korpusi JANES za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave." *Viri, orodja in metode za analizo spletne slovenščine* (2018): 44–73. https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/111/203/2416.

Dobrovoljc, Kaja, and Joakim Nivre. "The Universal Dependencies Treebank of Spoken Slovenian." *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), 2016, 1566–73. https://aclanthology.org/L16-1248.

Dobrovoljc, Kaja. "Skladenjska drevesnica govorjene slovenščine: stanje in perspektive." *Stanje in perspektive uporabe govornih virov v raziskavah govora*, 2024, 41–62.

Eckart de Castilho, Richard, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. "A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures." *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. Osaka, Japan: The COLING 2016 Organizing Committee (2016), 76–84. https://www.aclweb.org/anthology/W16-4011.

Erjavec, Tomaž, Darja Fišer, Jaka Čibej, and Špela Arhar Holdt. "CMC training corpus JANES-Norm 1.2." *Slovenian language resource repository CLARIN.SI* (2016a). http://hdl.handle.net/11356/1084.

Erjavec, Tomaž, Darja Fišer, Jaka Čibej, and Špela Arhar Holdt. "CMC training corpus JANES-Tag 1.1." *Slovenian language resource repository CLARIN.SI* (2016b). http://hdl.handle.net/11356/1081.

Fišer, Darja, Nikola Ljubešić, and Tomaž Erjavec. "The JANES Project: Language Resources and Tools for Slovene User-Generated Content." *Language Resources Evaluation* 54 (2020): 223–46. https://doi.org/10.1007/s10579-018-9425-z.

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.

Label Studio, https://labelstud.io/.

Ljubešić, Nikola, and Kaja Dobrovoljc. "What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian." *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Florence, Italy. Association for Computational Linguistics, 2019, 29–34. https://aclanthology.org/W19-3704/.

Ljubešić, Nikola, Luka Terčon, and 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.

Pori, Eva, Jaka Čibej, Tina Munda, Luka Terčon, and Špela Arhar Holdt. "Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref." *Konferenca Jezikovne tehnologije in digitalna humanistika* (2022): 162–68. Ljubljana, Slovenija. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Pori-et-al_Lematizacija-in-oblikoskladenjsko-oznacevanje-korpusa-SentiCoref.pdf.

PyBossa. https://docs.pybossa.com/.

Terčon, Luka, Jaka Čibej, and Nikola Ljubešić. "The CLASSLA-Stanza model for lemmatisation of standard Slovenian 2.0." *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042 (2023). http://hdl.handle.net/11356/1768.

Verdonik, Darinka, Andreja Bizjak, Mirjam Sepesy Maučec et al. "ASR database ARTUR 1.0 (transcriptions)." *Slovenian language resource repository CLARIN.SI* (2023). http://hdl.handle.net/11356/1772.

Verdonik, Darinka, Kaja Dobrovoljc, Peter Rupnik, Nikola Ljubešić, Simona Majhenič, Jaka Čibej, and Thomas Schmidt. "Training corpus of spoken Slovenian ROG 1.0." *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042 (2024). http://hdl.handle.net/11356/1992.

Verdonik, Darinka, Nikola Ljubešić, Peter Rupnik, Kaja Dobrovoljc, and Jaka Čibej. "Izbor in urejanje gradiv za učni korpus govorjene slovenščine ROG." *Konferenca jezikovne tehnologije in digitalna humanistika*. Ljubljana, Slovenija. (2024), 472–88.

Zwitter Vitez, Ana, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej, and Tomaž Erjavec. "Spoken corpus GOS 1.1." *Slovenian language resource repository CLARIN.SI*. (2021). http://hdl.handle.net/11356/1438.

Zwitter Vitez, Ana, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej, Tomaž Erjavec, Darinka Verdonik et al. "Spoken corpus GOS 2.0 (transcriptions)." *Slovenian language resource repository CLARIN.SI* (2023). http://hdl.handle.net/11356/1771.

Jaka Čibej, Tina Munda

UPORABA OBLIKOSLOVNEGA LEKSIKONA PRI POLAVTOMATSKEM PRISTOPU K POPRAVLJANJU LEM IN OBLIKOSKLADENJSKIH OZNAK

POVZETEK

V prispevku smo zasnovali nov polavtomatski pristop k popravljanju lem in oblikoskladenjskih oznak, ki se od predhodnih ročnih pristopov razlikuje po dodatni fazi navzkrižne primerjave s Slovenskim oblikoslovnim leksikonom Sloleks. V tem koraku so pojavnice in njihove strojno pripisane oblikoskladenjske značilnosti ter leme razvrščene v označevalne scenarije, na podlagi katerih je delo mogoče razdeliti v ločene sklope. Na ta način potrebujemo precej manj časa za proučevanje označevalnih smernic po sistemu Multext-East za slovenščino, delitev na sklope podobnih nalog pa omogoča tudi, da različno izkušenih označevalcem dodelimo delo primerne težavnosti. Metodo smo preizkusili pri označevanju Učnega korpusa govorjene slovenščine ROG ter dodatno stestirali na Učnem korpusu pisne slovenščine SUK. Rezultati kažejo, da je novi pristop hitrejši in v primerjavi s predhodnimi metodami zmanjša časovni vložek s približno 500 ur na 105 ur dela (na primeru korpusa ROG), pri čemer je končni odstotek popravljenih lem in oblikoskladenjskih oznak primerljiv (4-5 % za oblikoskladenjske oznake ter 1,3 % za leme). Pri tem so problematične predvsem enakopisnice na eni strani (zlasti če še niso popisane v leksikonu) ter neleksikonske pojavnice na drugi. S posodabljanjem Slovenskega oblikoslovnega leksikona Sloleks bo metoda v prihodnje še zanesljivejša, v prihodnje pa lahko postopek še nadgradimo s proučevanjem posameznih mikronalog – opazujemo lahko, kako se strojno označevanje obnese pri določenih enakopisnicah, ter popišemo, katere so manj verjeten vir napak, kar lahko upoštevamo pri načrtovanju označevanja.

1.01 DOI: https://doi.org/10.51663/pnz.65.3.07

Mojca Brglez ,* Veronika Bajt,♦ Senja Pollak,° Špela Rot,• Matej Martinc♠

Od kamnitega do spletnega portala: samodejno zaznavanje sprememb v rabi besed

IZVLEČEK

V prispevku prikažemo sistem za zaznavanje sprememb v rabi besed v slovenščini, ki omogoča samodejno zaznavanje pomenskih premikov v različnih časovnih obdobjih. Najprej predstavimo tehnično zasnovo in zahteve sistema, metodologijo za odkrivanje sprememb in grafični uporabniški vmesnik, ki omogoča uporabniku prijazno uporabo, nato pa demonstriramo, kako je sistem mogoče implementirati na referenčnem korpusu slovenščine Gigafida 2.0 in ga uporabiti za iskanje in analizo sprememb v rabi besed v različnih časovnih obdobjih. Rezultate sistema evalviramo s pomočjo kognitivno-jezikoslovne in leksikalne analize najbolj spremenjenih pridevnikov in samostalnikov, kjer raziščemo in kategoriziramo pomene in rabe besed v zaznanih gručah glede na njihovo semantično motiviranost in zastopanost v slovarju. Nazadnje sistem uporabimo na primeru reprezentacije migracij v časovnih obdobjih z ročno določenimi ločnicami, ki so signifikantno vplivale na odnos do migracije in migrantov v Sloveniji, ter tako preverimo njegovo uporabnost za sociolingvistične raziskave. Z jezikoslovnega vidika ugotavljamo, da sistem razločuje pomensko, skladenjsko in drugače kontekstualno različne rabe,

* **Asist., Univerza v Ljubljani, Filozofska fakulteta, Aškerčeva cesta 2, Ljubljana; Institut »Jožef Stefan«, Jamova cesta 39, Ljubljana, mojca.brglez@ff.uni-lj.si; ORCID: 0000-0002-8806-0942**

♦ **Dr., znan. sod., Mirovni inštitut, Metelkova 6, Ljubljana, veronika.bajt@mirovni-institut.si; ORCID: 0000-0002-6917-3255**

° **Doc. dr., Institut »Jožef Stefan«, Jamova cesta 39, Ljubljana, senja.pollak@ijs.si; ORCID: 0000-0002-4380-0863**

• **Mag. psih., Univerza v Ljubljani, Filozofska fakulteta, Oddelek za psihologijo, spelca.rot@gmail.com**

♠ **Asist. dr., Institut »Jožef Stefan«, Jamova cesta 39, Ljubljana, matej.martinc@ijs.si; ORCID: 0000-0002-7384-8112**

in pokažemo, da omogoča zaznavo tako kratkoročnih kot dolgoročnih sprememb. Po drugi strani ugotavljamo, da sistem jasno prikaže vpliv zunanjih dejavnikov v specifičnih časovnih obdobjih na jezik in diskurz in je tako uporabno orodje za sociolingvistično analizo.

Ključne besede: zaznavanje sprememb v rabi besed, semantika, pomenski premiki, sociolingvistika

ABSTRACT

A SYSTEM FOR WORD USAGE CHANGE DETECTION: ITS USE IN LINGUISTIC AND SOCIOLINGUISTIC STUDIES

This paper presents a system for detecting changes in Slovene word usage, enabling the automatic identification of semantic and other shifts across different time periods. We first introduce the system's technical design and requirements, the methodology for detecting changes, and the graphical user interface, which ensures a user-friendly experience. We then demonstrate how the system can be implemented on the reference corpus of Slovene, Gigafida 2.0, and used to search for and analyse changes in word usage across various time periods. The system's results are evaluated through a cognitive-linguistic and lexical analysis of the most changed adjectives and nouns, where we examine and categorise word meanings and usages within the detected clusters based on their semantic motivation and representation in dictionaries. Finally, we apply the system to a case study of migration representation in different time periods with manually defined boundaries, which have significantly influenced attitudes toward migration and migrants in Slovenia, thereby testing its applicability for sociolinguistic research. From a linguistic perspective, we observe that the system distinguishes between semantic, syntactic, and other contextually distinct usages, demonstrating its ability to detect both short-term and long-term changes. Furthermore, we observe that the system clearly illustrates the impact of external factors on language and discourse in specific time periods, making it a valuable tool for sociolinguistic analysis.

Keywords: word usage change detection, semantics, meaning shifts, sociolinguistics

Uvod

Jezik je dinamičen sistem, ki se z uporabo v družbenih interakcijah, spremembami kulturnih praks in razvojem tehnologije nenehno spreminja.[1] Spremembe so lahko vidne na fonološki, skladenjski, leksikalni ali semantični ravni, torej zadevajo od sprememb v izgovorjavi do spremembe pomenov besed. Preučevanje semantičnih

1 Jean Aitchison, *Language Change: Progress or Decay?* (Cambridge University Press, 2001), 133–83.

sprememb se je pričelo še pred pojavom sodobnega jezikoslovja v poznem 19. in zgodnjem 20. stoletju, področje pa vse od takrat napreduje.[2] Zaznavanje teh sprememb je pomembno za različne sinhrone in diahrone jezikoslovne raziskave, prispeva pa tudi k širši družboslovni analizi in omogoča vpogled v različne dejavnike sprememb.[3] Z vidika kognitivnega jezikoslovja jezik poleg zunanjih odraža tudi notranje dejavnike, tj. procese zaznavanja in razumevanja sveta okrog nas.[4] Med kognitivnimi mehanizmi, ki botrujejo pomenskim prenosom, sta ključni metonimija, ki temelji na sorodnosti, in metafora, ki temelji na podobnosti.[5]

Raziskave razvoja jezika se bodisi osredotočajo na dolgoročne spremembe pomena v diahronih korpusih ali pa na precej pogoste kratkoročne pojave, kot je na primer pojavitev besede v novem kontekstu. Pri slednjem ni nujno, da gre za spremembo ali razširitev pomena, saj pomen v kontekstu ustreza enemu od pomenov v slovarju.[6] Ko v pričujočem članku govorimo o »spremembah v rabi besed«, se nanašamo na vse vrste sprememb – kratkoročne ali dolgoročne, ki poleg jasnih pomenskih premikov vključujejo tudi spremembe kontekstov rabe besed.

Samodejno zaznavanje sprememb v rabi besed je zelo aktivno raziskovalno področje. Medtem ko so bili prvi sistemi za samodejno zaznavanje semantičnih sprememb razviti pred več kot desetletjem,[7] so raziskave v zadnjem času dobile dodaten zagon z idejo o uporabi besednih vložitev. Te so visokodimenzionalni matematični vektorji, ki predstavljajo besede po načelu distribucijske semantike: pomen besed je odvisen od njihove uporabe v kontekstu oziroma sopojavljanja z drugimi besedami.[8] Najsodobnejši sistemi za zaznavanje sprememb uporabljajo različne vrste besednih vložitev, za sistematično primerjavo različnih metod pa je bilo v zadnjih letih organiziranih tudi več tekmovanj in delavnic.[9] Delavnice so sicer večinoma namenjene zaznavanju sprememb v jezikih z veliko viri in govorci, kot so angleščina, ruščina, nemščina, italijanščina in španščina, jezikom z manj viri in govorci, med katerimi je tudi slovenščina, pa se doslej ni posvečalo veliko pozornosti.

2 Nina Tahmasebi, Lars Borin, Adam Jatowt et al., ur., *Computational Approaches to Semantic Change* (Language Science Press, 2021), https://doi.org/10.5281/zenodo.5040241.

3 Nabeel Gillani in Roger Levy, »Simple dynamic word embeddings for mapping perceptions in the public sphere,« v: *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science* (2019), 94–99, https://doi.org/10.18653/v1/W19-2111. Polona Gantar, Špela Arhar Holdt in Senja Pollak, »Leksikalne novosti v besedilih računalniško posredovane komunikacije,« *Slavistična revija* 66, št. 4 (2018): 459–72.

4 George Lakoff in Johnson, Mark, *Metaphors We Live By* (University of Chicago Press, 1980).

5 Eve Sweetser, *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure* (Cambridge University Press, 1990).

6 Syrielle Montariol, Matej Martinc, Lidia Pivovarova et al., »Scalable and interpretable semantic change detection,« v: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistis, 2021), 4642–52.

7 Martin Hilpert in Stefan Th. Gries, »Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition,« v: *Literary and Linguistic Computing* 24, št. 4 (2009): 385–401, https://doi.org/10.1093/llc/fqn012. Patrick Juola, »The time course of language change,« *Computers and the Humanities* 37, št. 1 (2003): 77–96, https://doi.org/ 10.1023/A:1021839220474.

8 Zellig S. Harris, »Distributional Structure,« *WORD* 10, št. 2-3 (1954): 146–62.

9 Med drugimi je bila v letu 2020 organizirana delavnica *SemEval-2020 Task 1: Unsupervised lexical semantic change detection* za zaznavanje sprememb v rabi besed za angleščino, nemščino, švedščino in latinščino, v letu 2022 pa *LSCDiscovery: A shared task on semantic change discovery and detection in Spanish* za španščino.

Pričujoči članek temelji na konferenčnem prispevku, ki so ga pripravili Martinc in sod.,[10] v katerem sta predstavljena izdelava prvega javno dostopnega sistema za zaznavanje sprememb v rabi posameznih besed za slovenščino in uporabniku prijazen spletni vmesnik.[11] Medtem ko omenjeni konferenčni članek zgolj na kratko demonstrira, kako je sistem mogoče uporabiti za jezikoslovne analize, v tem prispevku poleg predstavitve celotnega cevovoda ponudimo tudi podrobnejšo evalvacijo rezultatov. Sistem ovrednotimo predvsem z vidika njegove uporabnosti za razpoznavanje pomenskih premikov, pri čemer iščemo razširitve in/ali zožitve osnovnega pomena, ki so običajno metaforično ali metonimično motivirane. Poleg tega prikažemo uporabnost sistema za sociolingvistične analize z analizo izbrane leksike s področja migracij, kar omogoča vpogled v odnos lokalnega prebivalstva do priseljevanja v različnih obdobjih in naslavljanje širših družbenopolitičnih posledic polarizirajočih javnih razprav o migracijah.

Sorodne raziskave

V zadnjem času področje avtomatskega zaznavanja sprememb v rabi besed postaja vse pomembnejše, saj je uporabno ne le v jezikoslovju, na primer v diahronih korpusih za raziskave zgodovinskega razvoja jezika[12] ali specifičnih semantičnih premikov, kot je metafora,[13] temveč tudi v sinhronih korpusih pri različnih sociooloških in kulturoloških raziskavah. Med temi lahko omenimo na primer zaznavanje kratkoročnih sprememb v diskurzu, ki jih povzročijo krizni dogodki, kot je pojav neologizmov ob epidemiji virusa covid-19,[14] ali pa zaznavanje ideološko pogojenih razlik v diskurzu.[15]

Prvi sistemi za samodejno zaznavanje sprememb v rabi so bili razviti pred več kot desetletjem. Temeljili so na metodah, ki vzorčijo in analizirajo predvsem pogostost besed v različnih časovnih obdobjih.[16] S takimi metodami lahko v diahronih korpusih, ki zajemajo različna obdobja, zgolj na podlagi spremembe v številu pojavitev odkrivamo neologizme ali nove pomene besed, na primer pojav besede *medmrežje* ob nove

10 Matej Martinc, Veronika Bajt, Špela Rot et al., »Sistem za zaznavanje sprememb v rabi besed in njegova uporaba za sociolingvistično analizo,« v: *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2024* (Inštitut za novejšo zgodovino, 2024), 298–318, https://doi.org/10.5281/zenodo.13936410.

11 Uporabniški vmesnik je javno dostopen na spletnem naslovu http://kt-nlp-demo.ijs.si:8080

12 Yuting Wei, Meiling Li, Yangfu Zhu, Yuanxing Xu, Yuqing Li in Bin Wu, »A diachronic language model for long-time span classical Chinese,« *Information Processing & Management* 62, št 1 (2025), 103925, https://doi.org/10.1016/j.ipm.2024.103925.

13 Marco Del Tredici, Malvina Nissim in Andrea Zaninello, »Tracing metaphors in time through self-distance in vector spaces,« v: *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-It 2016,* Accademia University Press, 2016, 117–22, https://doi.org/10.4000/books.aaccademia.1760.

14 Quirin Würschinger in Barbara McGillivray, »Semantic change and socio-semantic variation: the case of COVID-related neologisms on Reddit,« *Linguistics Vanguard* (2024), https://doi.org/10.1515/lingvan-2023-0106.

15 Isabelle Gribomont, »From Diachronic to Contextual Lexical Semantic Change: Introducing Semantic Difference Keywords (SDKs) for Discourse Studies,« v: *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, Association for Computational Linguistics, 2023, 153–60. Matej Martinc, Nina Perger, Andraž, Pelicon, Matej Ulčar, Andreja Vezovnik in Senja Pollak, »EMBEDDIA hackathon report: Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+,« v: *Proceedings of the EACL Hackashop on news media content analysis and automated report generation* (2021), 121–26.

16 Hilpert in Gries, »Assessing frequency changes.« Juola, »The time course.«

tehnologije konec 20. stoletja, pa tudi upad nekaterih jezikovnih oblik, kot je deležnik preteklega časa (*videvši*, *pozabivši*). Podroben opis metod, ki temeljijo na pogostosti, je mogoče najti na primer v preglednem članku Tahmasebi, Borin in Jatowt.[17] Ti pristopi se danes le redko uporabljajo, saj so se s pojavom besednih vložitev razvile mnogo učinkovitejše metode. Vložitve so eden od načinov, s katerimi lahko informacije v jeziku matematično predstavimo in katerih izgradnja temelji na načelu distribucijske semantike: pomen besed je odvisen od njihove uporabe v kontekstu oziroma sopojavljanja z drugimi besedami.[18] Besedne vložitve so reprezentacije posamičnih besed v vektorskem prostoru z veliko dimenzijami, običajno od 100 do 1000. Ustvarimo jih s pomočjo jezikovnih modelov, ki se učijo napovedovati sosednje ali manjkajoče besede na veliki količini besedil. Za razliko od prejšnjih metod, ki so temeljile le na pogostosti pojavitev, besedne vložitve vsebujejo tudi skladenjske in pomenske informacije.[19] V ustvarjenem vektorskem prostoru imajo pomensko in skladenjsko podobne besede tudi podobne vložitve, z ustvarjenimi vektorji pa lahko izvajamo različne računske operacije, kot je »računanje« analogij.[20]

Sodobni sistemi za samodejno zaznavanje sprememb v rabi besed temeljijo na izgradnji vložitev za vsako posamično časovno obdobje (rezino korpusa) posebej, pri čemer so te lahko ustvarjene na dva načina. Pri prvem nastanejo t. i. *statične vložitve*, saj se za vsako besedo ustvari le ena vložitev, ki je nekakšno povprečje vseh njenih rab v učnem korpusu. Novejši tip vložitev, ki jih pridobimo na primer z jezikovnimi modeli tipa BERT,[21] pa so t. i. *dinamične* ali *kontekstualne vložitve*: za besedo dobimo drugačno vložitev glede na specifično sobesedilo (npr. poved), v katerem je uporabljena. To omogoča razločevanje različnih pomenov in rab besed, denimo med besedo *golf*, uporabljeno v pomenu športne discipline, ali besedo *golf*, s katero označujemo model avtomobila.

Pri metodah za samodejno zaznavanje sprememb v rabi, ki uporabljajo statične vložitve, so te vložitve običajno najprej naučene na vsaki časovni rezini korpusa posebej in zatem poravnane, da postanejo med seboj primerljive. V prispevku Kim in sod.[22] je bila ta metoda uporabljena za zaznavanje angleških besed, ki so znatno spremenile rabo med letoma 1900 in 2009 (npr. besedi *gay* in *cell*). Ker je posamična beseda

17 Nina Tahmasebi, Lars Borin in Adam Jatowt, »Survey of computational approaches to lexical semantic change detection,« v: Tahmasebi et al., ur., *Computational Approaches to Semantic Change* (Language Science Press, 2021, 1–91), https://doi.org/10.5281/zenodo.5040302.

18 Harris, »Distributional Structure.«

19 Tomas Mikolov, Ilya Sutskever, Kai Chen et al., »Distributed representations of words and phrases and their compositionality,« v: *Advances in Neural Information Processing Systems* 26 (2013): 3111–19.

20 Eden bolj znanih primerov izračuna semantične analogije je *moški – kralj + ženska = x*, pri čemer je rezultatu x najbližje vložitev besede *kraljica*.

21 Jacob Devlin, Ming-Wei Chang, Kenton Lee et al., »BERT: Pre-training of deep bidirectional transformers for language understanding,« v: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (ACL, 2019): 4171–86.

22 Yoon Kim, Yi-I Chiu, Kentaro Hanaki et al., »Temporal analysis of language through neural language models,« v: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science* (2014): 61–65. William L. Hamilton, Jure Leskovec in Dan Jurafsky, »Diachronic word embeddings reveal statistical laws of semantic change,« v: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (ACL, 2016): 1489–501.

(oziroma vse njene rabe) znotraj časovne rezine predstavljena samo z eno vektorsko reprezentacijo, so metode, ki temeljijo na statičnih vložitvah, manj natančne, prav tako pa rezultate težje interpretiramo. Omejitev je mogoče odpraviti z uporabo kontekstualnih vložitev, ki omogočajo modeliranje različnih pomenov in rab. Vsi taki pristopi k zaznavanju sprememb v rabi vsebujejo tudi postopek agregacije, v katerem so kontekstualne vložitve posameznih pojavitev besed v določenem časovnem obdobju v korpusu združene v smiselne časovne reprezentacije. Za agregacijo se uporabljajo različne metode, od preprostega povprečenja[23] in primerjave parov vektorjev[24] do združevanja v gruče.[25] Pri zadnjem se predvideva, da posamezna gruča reprezentacij združuje eno rabo oziroma pomen dane besede. Najbolj priljubljena metoda za primerjavo gruč iz različnih časovnih obdobij, in s tem pridobitev kvantitativne ocene spremembe v rabi določene besede, je Jensen-Shannonova divergenca (JSD),[26] ki so jo uporabili na primer Giulianelli in sod.[27] ter Martinc in sod.[28] Pri tej primerjamo distribucije različnih gruč (ki naj bi ustrezale pomenom in rabam) v različnih časovnih obdobjih in tako ugotovimo, ali se je distribucija pomenov/rab v dveh ali več obdobjih spremenila. To metodo so Montariol in sod.[29] uporabili za identifikacijo kratkoročnih (mesečnih) sprememb v rabi angleških besed med pandemijo COVID. Tako na primer beseda *strain*, ki se je v prvih dveh mesecih pandemije večinsko uporabljala v kontekstu »različic koronavirusa« (angl. *coronavirus strain*), v naslednjih mesecih pandemije pridobi novo večinsko rabo v kontekstu »obremenitve zdravstvenega sistema« (angl. *strain on the health system*).

Raziskave sprememb v rabi besed v slovenščini so redke. Med tistimi, ki na splošno analizirajo in kategorizirajo različne pomene in pomenske premike, se v slovenščini pojavljajo tako teoretski kot empirični pristopi, predvsem z vidika leksikologije in leksikografije. Med prvimi lahko omenimo dela Ade Vidovič Muha in Jerice Snoj,[30] ki preučujeta večpomenskost leksemov. Med tipi večpomenskosti ločujeta pomensko vsebovanost (pod- in nadpomenskost) ter pomenske prenose, ki vključujejo tri vrste: metaforo, metonimijo in sinekdoho. Med raziskavami, ki bodisi zaznavajo in/ali analizirajo pomenske premike na podlagi dejanske rabe, lahko omenimo dve študiji.

23 Matej Martinc, Petra Kralj Novak in Senja Pollak, »Leveraging contextual embeddings for detecting diachronic semantic shift,« v: *Proceedings of the Twelfth Language Resources and Evaluation Conference* (EACL, 2020): 4811–19.

24 Andrey Kutuzov in Mario Giulianelli, »UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection,« v: *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (International Committee for Computational Linguistics, 2020), 126–34.

25 Montariol et al., »Scalable and interpretable.« Matej Martinc, Syrielle Montariol, Elaine Zosa et al., »Capturing evolution in word usage: Just add more clusters?,« v: *Companion Proceedings of the Web Conference 2020* (Association for Computing Machinery, 2020), 343–49, https://doi.org/10.1145/3366424.3382186. Mario Giulianelli, Marco Del Tredici in Raquel Fernández, »Analysing lexical semantic change with contextualised word representation,« v: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (ACL, 2020): 3960–73.

26 Jianhua Lin, »Divergence measures based on the Shannon entropy,« *IEEE Transactions on Information theory* 37, št. 1 (1991): 145–51.

27 Giulianelli et al., »Analysing lexical semantic change.«

28 Martinc et al., »Capturing evolution in word usage.«

29 Montariol et al., »Scalable and interpretable.«

30 Ada Vidovič Muha, *Slovensko leksikalno pomenoslovje: govorica slovarja* (Ljubljana: Znanstveni inštitut Filozofske fakultete, 2000). Jerica Snoj, »Slovarska večpomenskost in *Slovensko leksikalno pomenoslovje*,« *Slavistična Revija* 51, št. 4 (2003): 387–409.

Gantar, Arhar Holdt in Pollak[31] se ukvarjajo z odkrivanjem nove leksike in pomenov predvsem s pomočjo luščenja kolokacij iz korpusa Janes,[32] ki vsebuje računalniško posredovana besedila. Znotraj istega korpusa, vendar z omejitvijo na tvite, raziskavo izvedeta tudi Fišer in Ljubešić.[33] Natančneje, s pomočjo besednih skic analizirata 200 besed, pri katerih so bile zaznane spremembe v vektorski reprezentaciji v primerjavi z referenčnim korpusom standardne slovenščine. Raziskava je narejena s pomočjo statičnih vektorskih vložitev in vsebuje velik delež napak (45 odstotkov), vendar predstavlja zanimivo kategorizacijo sprememb. Poleg novih pomenov so v analizo namreč vključene tudi manj očitne razlike v rabi, analiza pa razlikuje med manjšimi in večjimi premiki. Pri tem naj bi bili manjši premiki vezani na spremembe v distribuciji (že uveljavljenih) pomenov in omejenost na določene vzorce ali pomene, do večjih premikov pa pride zaradi aktualnih dogodkov, razlik v registru ali razlik v mediju. Raziskava se od pričujoče razlikuje v metodologiji in v tem, da primerja žanrsko in jezikovno zelo različna besedila, medtem ko se naš sistem osredotoča na zaznavanje sprememb skozi čas.

Med raziskavami, ki uporabljajo sodobne metode za samodejno zaznavanje sprememb v slovenščini, je relevantna predvsem pred kratkim izvedena študija Pranjića in sod.[34] V raziskavi je bila izdelana prva testna množica za testiranje različnih slovenskih modelov za zaznavanje sprememb v rabi besed. Ročno označevanje je bilo izvedeno na podlagi kvantitativne, stopenjske ocene podobnosti pomenov besede v paru povedi. V študiji je predstavljen tudi nov model za zaznavanje semantičnih premikov s pomočjo optimalnega transporta, med drugim pa so preizkusili tudi metodologijo, ki jo opisujemo v tej študiji. Nazadnje naj omenimo še raziskavo Martinca in sod.[35], kjer je bil sistem za zaznavanje sprememb v rabi uporabljen za analizo gledišč različnih slovenskih medijev. V raziskavi se osredotočajo na razlike v poročanju med osrednjimi in konservativnimi mediji o tematikah, povezanih s skupnostjo LGBTIQ. Glavna ugotovitev raziskave je, da skupini medijev najbolj drugače uporabljata besedo *globok*. Ta se v osrednjih medijih večinoma uporablja v konvencionalnem pomenu, medtem ko se na konservativnih novičarskih portalih pretežno uporablja v kontekstu zveze »globoka država«.

31 Gantar et al., »Leksikalne novosti.«

32 Tomaž Erjavec, Nikola Ljubešić in Darja Fišer, »Korpus slovenskih spletnih uporabniških vsebin Janes,« v: Darja Fišer, ur., *Viri, orodja in metode za analizo spletne slovenščine* (Ljubljana: Znanstvena založba Filozofske fakultete, 2018), 16–43.

33 Darja Fišer in Nikola Ljubešić, »Tviti kot leksikografski vir za analizo pomenskih premikov v slovenščini,« v: Darja Fišer, ur., *Viri, orodja in metode za analizo spletne slovenščine* (Ljubljana: Znanstvena založba Filozofske fakultete, 2018), 198–226.

34 Marko Pranjić, Kaja Dobrovoljc, Senja Pollak et al., »Semantic change detection for Slovene language: a novel dataset and an approach based on optimal transport,« *arXiv:2402.16596* (arXiv preprint, 2024), https://doi.org/10.48550/arXiv.2402.16596.

35 Matej Martinc, Nina Perger in Senja Pollak, »Viewpoint detection on LGBT+ reporting using contextual embeddings and qualitative thematic analysis: The use case on the word deep,« *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* (2025): 07591063251317085.

Opis sistema za zaznavanje sprememb

Podatkovne in računske zahteve

Za predlagani sistem za zaznavanje sprememb v rabi v prvi vrsti potrebujemo korpus, ki vsebuje besedila iz različnih časovnih obdobij in ga je mogoče razdeliti na časovne rezine. Dolžina posameznih časovnih obdobij in razmejitve med obdobji so poljubne, v praksi pa so pogojene z raziskovalnim vprašanjem in količino podatkov, ki je na voljo. V idealnem primeru naj bi vsaka časovna rezina korpusa vsebovala vsaj pet milijonov besed. To omogoča sestavo obsežnega besedišča, ki mu lahko določimo spremembo v rabi skozi čas. Vsaka beseda, za katero želimo izmeriti spremembe v rabi, se mora za veljavnost rezultatov v vsaki časovni rezini korpusa pojaviti vsaj 20-krat, v idealnem primeru vsaj 100-krat. Manj kot 20 pojavitev določene besede namreč ne omogoča izdelave dovolj kakovostne distribucije rab besede za posamezno obdobje.

Eden od pomembnih kriterijev za izbor metode je tudi skalabilnost. Večina metod, ki temeljijo na kontekstualnih vložitvah, je neprimernih zaradi ogromnih potreb po delovnem spominu (RAM), saj je treba v spomin shraniti vektorsko reprezentacijo za vsako pojavitev besede v korpusu. Izbrana metoda po drugi strani s pomočjo posebnega mehanizma predhodne agregacije vektorskih reprezentacij na podlagi kosinusne podobnosti omogoča, da se za vsako besedo v določeni časovni rezini korpusa shrani do največ 200 besednih vložitev, kar omogoča rabo metode na velikih korpusih in na celotnem besedišču korpusa.[36]

Največji korpus, na katerem je bil preizkušen sistem, je vseboval približno 100 milijonov besed na časovno rezino in besedišče, sestavljeno iz približno 8000 lem,[37] a teoretično zgornje meje za velikost korpusa ni. Vendar pa je treba upoštevati nekatere praktične omejitve, saj se z velikostjo besedišča in številom časovnih obdobij povečajo tudi zahteve po diskovnem spominu.

Cevovod za zaznavanje sprememb v rabi

Sistem za zaznavanje sprememb v rabi besed je sestavljen iz več zaporednih korakov, združenih v tako imenovani »cevovod«. Najprej potekajo predprocesiranje korpusa, adaptacija jezikovnega modela na domenski korpus, razdelitev korpusa na časovne rezine in luščenje kontekstualnih vložitev iz jezikovnega modela. Sledijo gručenje kontekstualnih vložitev, izdelava distribucij gruč glede na časovno obdobje in merjenje sprememb v rabi med časovnimi obdobji. Vsakega od teh korakov pojasnimo spodaj.

36 Montariol et al., »Scalable and interpretable.«
37 Ibidem.

- **Predprocesiranje korpusa:** V prvem koraku korpus tokeniziramo (razdelimo na pojavnice) in lematiziramo (spremenimo pojavnice v leme) s pomočjo orodij za predprocesiranje; v našem primeru smo uporabili orodje za jezikovno obdelavo slovenščine CLASSLA-Stanza.[38]
- **Domenska adaptacija modela:** Nevronski jezikovni model prilagodimo preučevani domeni, tako da ga pet epoh učimo na celotnem korpusu. Učenje poteka na nenadzorovan način, tj. na nalogi napovedovanja naključno skritih besed v besedilu.
- **Razdelitev korpusa na časovne rezine:** Korpus razdelimo na časovne rezine, ki se ločeno vnesejo v model v serijah (angl. *batch*) po 32 besedilnih sekvenc naenkrat. Besedilne sekvence omejimo na dolžino 256 žetonov.[39]
- **Ekstrakcija kontekstualnih vložitev:** Za vsako sekvenco oziroma pojavnice v sekvenci ustvarimo reprezentacijo, tako da vzamemo in seštejemo zadnje štiri izhodne plasti kodirnika nevronske mreže. Tako za vsako pojavnico dobimo 768-dimenzionalno kontekstualno vložitev.[40] Za vsako lemo v pomnilniku hranimo seznam kontekstualnih vložitev, ki predstavljajo njene različne rabe v posamičnem obdobju. Da bi izboljšali skalabilnost sistema, število hranjenih vložitev omejimo na 200. Ob izluščenju nove vložitve iz besedilne sekvence se ta bodisi doda na seznam bodisi združi z eno od že pridobljenih vložitev. Slednje se zgodi, če *a)* je nova vložitev preveč podobna eni od hranjenih vložitev (kosinusna podobnost je večja ali enaka 0,99) ali *b)* če seznam že vsebuje vnaprej določeno največje število vložitev (200). Če pride do združitve, se nova vložitev združi z vložitvijo na seznamu, ki je najbližja po kosinusni razdalji. Na ta način za vsako lemo v besedišču pridobimo do 200 kontekstualnih vložitev, ki predstavljajo posamezno (ali združeno) pojavnico s to lemo v kontekstu.
- **Gručenje kontekstualnih vložitev:** Za ugotavljanje različnih rab posamezne leme v določenem časovnem obdobju s pomočjo algoritma *k-means* izvedemo gručenje kontekstualnih vložitev leme, ki naj bi predstavljale specifično rabo. Združevanje v gruče za dano lemo izvedemo na množici vložitev iz vseh časovnih obdobij skupaj. Število gruč, ki jih pridobimo z algoritmom *k-means*, določimo z vrednostjo $k = 5$. Večina besed ima namreč manj kot pet pogostih rab, kar pomeni, da v večini primerov zadostuje pet gruč za identifikacijo vseh pomenov. Če je k večji, so nekatere gruče narejene ne samo na podlagi semantičnih razlik (ki naj bi vodile v največje razlike med besednimi vložitvami), temveč tudi na podlagi oblikoskladenjskih in drugih razlik. Po zgoraj opisanem postopku gručenja zato izvedemo dodatno združevanje ali odstranjevanje. Po dve gruči združimo, če sta

38 Nikola Ljubešić, Luka Terčon in Katja Dobrovoljc, »CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages,« v: Špela Arhar Holdt in Tomaž Erjavec, ur., *Zbornik konference za jezikovne tehnologije in digitalno humanistiko (JT-DH-2024)* (Ljubljana: Inštitut za novejšo zgodovino, 2024), 251–74, https://doi.org/10.5281/zenodo.13936406.

39 Gre za podbesedne enote (angl. *subword token*), ki ne ustrezajo nujno pojavnicam ali besedam, saj je lahko ena pojavnica razdeljena na več žetonov.

40 Kadar pojavnico sestavlja več žetonov, njeno reprezentacijo izračunamo iz povprečja vložitev žetonov, ki jo sestavljajo.

si zelo podobni, odstranimo pa tiste, v katerih je manj kot deset pojavitev leme, saj to kaže na precej obrobno rabo.

– **Izdelava distribucije različnih rab:** Za vsako lemo v vsakem časovnem obdobju iz zgornjega koraka pridobimo množico gruč, ki predstavljajo različne rabe besede. Distribucijo rab v določenem obdobju pridobimo tako, da število pojavitev leme v vsaki gruči delimo s skupnim številom pojavitev leme v danem časovnem obdobju.
– **Merjenje sprememb v rabi:** Distribucije rab, ki jih za določeno lemo pridobimo za vsako časovno obdobje, primerjamo med sabo s pomočjo Jensen-Shannonove divergence (JSD)[41] za merjenje razlik med verjetnostnimi distribucijami. S pomočjo mere JSD lahko vsem besedam v besedišču korpusa izmerimo spremembe v distribuciji rabe med zaporednimi obdobji, jih razporedimo po velikosti izmerjene spremembe in tako poiščemo tiste besede, katerih raba se je med različnimi časovnimi obdobji najbolj spremenila.

Interpretacija rezultatov sistema

Sistem nam obenem omogoča, da s pomočjo metode za interpretacijo hitro razumemo, kako se raba posamezne besede med časovnimi obdobji spreminja. To dosežemo z uporabo mere TF-IDF (angl. *term frequency-inverse document frequency*). Za vsako rabo posamezne leme imamo na voljo kontekst, tj. poved, v kateri se določena lema pojavi v obliki neke pojavnice. Povedi, ki vsebujejo posamezne rabe besede, ki pripadajo isti gruči, najprej združimo v t. i. »dokument«, nato pa za vsak tak dokument izluščimo najbolj razločevalne unigrame, bigrame in trigrame, torej nize ene, dveh ali treh besed, ki dokumente med seboj najbolje razločijo.[42] Te pridobimo s pomočjo algoritma TF-IDF, pri čemer kot korpus obravnavamo skupek vseh »dokumentov«, tj. množico vseh povedi, v katerih se posamezna lema pojavi. Iz korpusa izključimo nepolnopomenske besede (angl. *stopwords*)[43] in besede, ki se pojavljajo v več kot 80 odstotkih gruč. S tem zagotovimo, da so izbrani ključni izrazi za vsako gručo čim bolj specifični in jih tako kar najbolje ločijo. Na koncu dobimo seznam do sedmih ključnih izrazov za vsako gručo, ki nudijo vpogled v posamezno rabo besede.

41 María L. Menéndez, Julio A. Pardo, Leandro Pardo in María C. Pardo, »The Jensen-Shannon divergence,« *Journal of the Franklin Institute* 334, št. 2 (1997): 307–18, https://doi.org/10.1016/S0016-0032(96)00063-4.

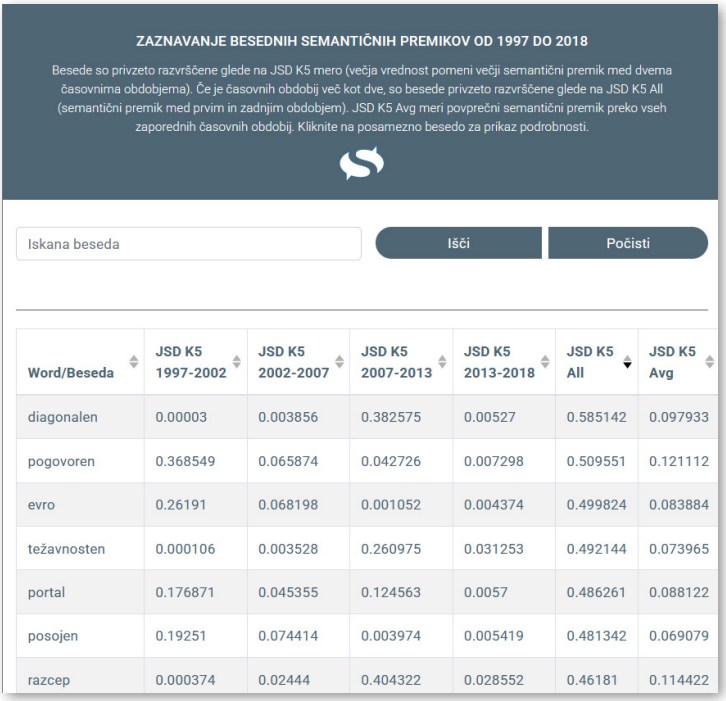
42 Primeri takih razločevalnih nizov so vidni na Sliki 2, prvo gručo tako označujeta mdr. unigram *okno* ter bigram *klikniti jeziček*.

43 Uporabljata se tudi izraza »pomensko prazne« ali »blokirane« besede, ki običajno vključujejo nepolnopomenske besedne vrste in/ali zelo pogoste besede. V predstavljenem eksperimentu smo uporabili seznam 1071 besed, izluščenih iz korpusa Kres, torej korpusa standardne slovenščine. Na seznam so uvrščeni predlogi, vezniki, členki in zaimki. Seznam vsebuje različnice, ne samo lem.

Uporabniški vmesnik

Do rezultatov sistema je mogoče dostopati prek spletnega uporabniškega vmesnika, ki omogoča hitro interpretacijo in analizo sprememb v rabi.[44] Sestavljen je iz dveh ločenih komponent. Prva ponuja globalni pogled na celoten korpus oziroma vsebovana obdobja v obliki tabele (Slika 1), kjer najdemo vse besede, ki se v korpusu pojavijo najmanj 20-krat, skupaj z njihovo izmerjeno spremembo v rabi med dvema obdobjema, skupni seštevek izmerjenih sprememb in število pojavitev v posamičnem obdobju. Besede so privzeto razvrščene glede na skupni seštevek izmerjenih sprememb v rabi med prvim in zadnjim časovnim obdobjem, vendar tabela omogoča razvrščanje po poljubnem stolpcu.

Slika 1: Prva komponenta uporabniškega vmesnika za globalni prikaz in iskanje po korpusu



Word/Beseda	JSD K5 1997-2002	JSD K5 2002-2007	JSD K5 2007-2013	JSD K5 2013-2018	JSD K5 All	JSD K5 Avg
diagonalen	0.00003	0.003856	0.382575	0.00527	0.585142	0.097933
pogovoren	0.368549	0.065874	0.042726	0.007298	0.509551	0.121112
evro	0.26191	0.068198	0.001052	0.004374	0.499824	0.083884
težavnosten	0.000106	0.003528	0.260975	0.031253	0.492144	0.073965
portal	0.176871	0.045355	0.124563	0.0057	0.486261	0.088122
posojen	0.19251	0.074414	0.003974	0.005419	0.481342	0.069079
razcep	0.000374	0.02444	0.404322	0.028552	0.46181	0.114422

Vir: lastno delo

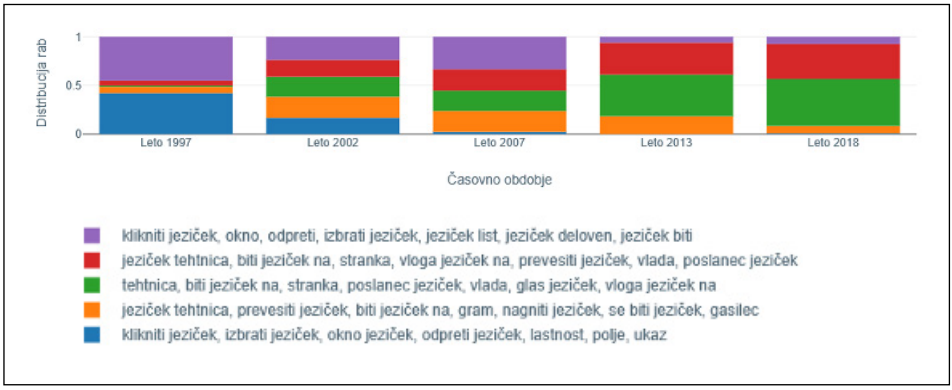
Do druge komponente uporabniškega vmesnika pridemo tako, da kliknemo na posamezno besedo v tabeli. Ta komponenta nudi podrobnejši prikaz in kontekst sprememb v rabi za posamezno besedo po časovnih obdobjih (Slika 2). Komponenta vizualizira posamična časovna obdobja v stolpcih tako, da z različnimi barvami predstavi distribucijo rab besede v posamičnem obdobju. V legendi slike nam vmesnik nudi tudi hitro interpretacijo gruč s ključnimi besedami in besednimi zvezami, specifičnimi

44 Vmesniki so prosto dostopni na naslovu http://kt-nlp-demo.ijs.si:8080.

za posamično gručo (predstavljeno v prejšnjem poglavju *Interpretacija rezultatov sistema*). S klikom na posamezno rabo (tj. barvo, ki predstavlja posamezno gručo) na sliki se nam spodaj izpiše seznam kontekstov (tj. povedi), ki sodijo v to gručo.

Uporabniški vmesnik je zasnovan tako, da lahko uporabnik z bolj splošnih informacij (na korpusni ravni), ki jih prikazuje prva komponenta, hitro (s pomočjo klika na posamezno besedo) prehaja na podrobnejše informacije (na besedni ravni), ki jih prikazuje druga komponenta, kar omogoča hiter vpogled v spremembe v rabi besede in podpira nadaljnjo analizo teh sprememb. V naslednjem poglavju podrobneje prikažemo, kako je sistem mogoče uporabljati v tem sosledju, in evalviramo rezultat sistema na dva načina. Pri prvem sistem uporabimo za odkrivanje in analizo pomenskih premikov, pri drugem pa za sociolingvistično analizo, kjer vzporejamo spremembe v jezikovni rabi s specifičnimi spremembami v družbi.

Slika 2: Primer druge komponente uporabniškega vmesnika, podrobnejši prikaz za besedo *jeziček*



Vir: lastno delo

Implementacija sistema za slovenščino

Za slovenščino smo nevronski model SloBERTa,[45] ki smo ga uporabili za ekstrakcijo kontekstualnih besednih vložitev, naučili na delu korpusa Gigafida 2.0.[46] Gigafida je referenčni korpus standardne pisane slovenščine in vsebuje besedila iz časopisov (47,8 odstotka besedil), revij (16,5 odstotka), internetnih vsebin (28,0 odstotka),[47] stvarnih besedil (3,8 odstotka), leposlovja (3,5 odstotka) in drugih zvrsti.

45 Matej Ulčar in Marko Robnik Šikonja, »SloBERTa: Slovene monolingual large pretrained masked language model,« v: *Zbornik 24. mednarodne multikonference Informacijska družba IS 2021, zvezek C* (Ljubljana: Institut »Jožef Stefan«, 2021), 17–20.

46 Simon Krek, Špela Arhar Holdt, Tomaž Erjavec et al., »Gigafida 2.0: the reference corpus of written standard Slovene,« v: *Proceedings of the 12th Language Resources and Evaluation Conference* (ELRA, 2020): 3340–45.

47 Internetna besedila vsebujejo tudi novice iz novičarskih portalov, ki so po vsebini zelo podobne časopisnim besedilom.

Tabela 1: Število dokumentov, besed in virov po letih v treh korpusih za merjenje sprememb v rabi besed

Obdobje	Št. dokumentov	Št. besed	Št. virov
Korpus za merjenje dolgoročnih sprememb v rabi			
Od 1990 do vključno 1997	6.939	69.794.466	62
2018	870	83.111.440	12
Korpus za merjenje kratkoročnih sprememb v rabi			
2017	1.053	104.031.504	16
2018	870	83.111.440	12
Korpus za merjenje sprememb v rabi med več obdobji			
Od 1990 do vključno 1997	6.939	69.794.466	62
2002	1.809	76.446.366	51
2007	219	113.740.532	73
2013	664	64.232.282	13
2018	870	83.111.440	12

Vir: lastno delo

Da bi lahko analizirali različna obdobja in vrste sprememb v rabi, smo iz besedil, ki jih zajema celotna Gigafida 2.0, sestavili tri korpuse. Prva korpusna različica[48] omogoča merjenje dolgoročnih sprememb v rabi med dvema obdobjema. Tu prvo obdobje pokriva osem let med 1990 in 1997 in vsebuje najstarejša besedila v Gigafidi 2.0. Za nekoliko daljši, osemletni razpon smo se odločili predvsem zato, da smo pridobili dovoljšno količino besedil za učenje modela. Drugo obdobje vsebuje besedila iz leta 2018, kar je zadnje leto, zajeto v Gigafidi 2.0. V tem korpusu nas zanimajo predvsem dolgoročne spremembe v rabi besed, ki so nastale v časovnem obdobju, daljšem od 20 let. Drugo različico korpusa[49] sestavljajo besedila iz zgolj dveh enoletnih obdobij, nastala v letih 2017 in 2018. V tem korpusu želimo meriti kratkoročne spremembe v rabi besed, ki so nastale v časovnem obdobju enega leta. Tretji korpus[50] je za razliko od prvih dveh razdeljen na pet obdobij, in sicer 1990–1997, 2002, 2007, 2013, 2018. S tem korpusom, ki pokriva največ virov in žanrov, želimo meriti spremembe v rabi besed med več zaporednimi obdobji in tako bolje razumeti celotno dinamiko spreminjanja rabe besed, ki ne poteka vedno linearno in v eni smeri. Velikosti posamičnih korpusov glede na število zajetih besedil, besed in virov predstavljamo v Tabeli 1.

48 Sistem na podlagi dveh podkorpusov je na voljo na (E8-NLP) http://kt-nlp-demo.ijs.si:8080/semanticshifttable/2.

49 Sistem na podlagi dveh letnih podkorpusov je dostopen na (E8-NLP) http://kt-nlp-demo.ijs.si:8080/semanticshifttable/3.

50 Sistem na podlagi petih podkorpusov je dostopen na (E8-NLP) http://kt-nlp-demo.ijs.si:8080/semanticshifttable/1.

Uporaba sistema za analizo pomenskih premikov in sprememb v rabi

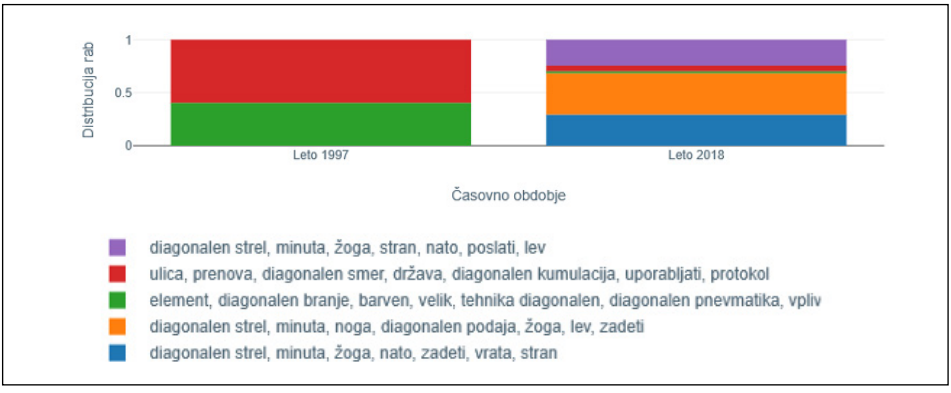
V poglavju analiziramo spremembe v rabi besed v prvi in tretji različici korpusa, tj. korpusa za merjenje dolgoročnih sprememb in korpusa za merjenje sprememb v več zaporednih obdobjih.

Dolgoročne spremembe v rabi

Kot smo že opisali, je korpus za merjenje dolgoročnih sprememb sestavljen na eni strani iz besedil, nastalih v obdobju 1990–1997, in na drugi iz besedil, nastalih v letu 2018. Med prvimi 50 besedami z največ spremembami glede na mero *JSD K5 All* močno prevladujejo pridevniki (29), sledijo samostalniki (16), medtem ko so glagoli (2) in prislovi (2) manj pogosti. V analizi se glede na pogostost posvetimo prvim trem pridevnikom (*diagonalen, stebrn, jonski*) in prvim trem samostalnikom (*podprogram, portal, izbijanje*) na seznamu.

Glede na mero JSD je v drugem obdobju najbolj drugačna raba pridevnika *diagonalen*. V obdobju devetdesetih se pojavljata izključno dve gruči pomenov/rab (Slika 3). Analiza povedi v gručah pokaže, da obe gruči vsebujeta mešane rabe besede, tako dobesedne (»diagonalna razpoka«, »diagonalna črta«), metonimične (»diagonalni korak«, »diagonalni bralec«) kot metaforične (»diagonalno zavezništvo«, »diagonalna kumulacija«). V letu 2018 vse te rabe praktično izginejo, prevladuje raba besede v športnih kontekstih. Ta pomen/raba je v sistemu sicer predstavljena v treh različnih gručah, vendar pa gre tako glede na izredno podobne ključne besede kot tudi glede na povedi v teh gručah za zelo podobno rabo. V povedih se namreč raba manifestira v zgolj nekaj besednih zvezah, in sicer se beseda *diagonalen* pojavlja kot prilastek samostalnikov *strel, udarec, bekhend, forehand, podaja, polvolej, predložek*.

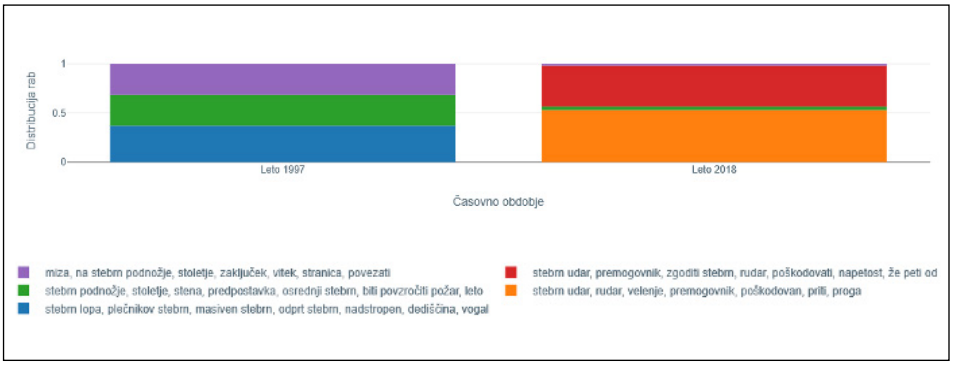
Slika 3: Distribucije rab besede *diagonalen* v obdobju 1990–97 in letu 2018



Vir: lastno delo

Druga najbolj spremenjena beseda je pridevnik *stebrn*. Sistem prikaže, da se je v obdobju do 1997 beseda pojavljala v treh gručah v vijolični, zeleni in modri barvi (Slika 4). Te so okarakterizirane s ključniki, ki med drugim vsebujejo besede *miza, podnožje, stoletje, zaključek, vitek, stranica, povezati* pa *stena, predpostavka, osrednji* ter *lopa, plečnikov, masiven, odprt, dediščina.* Pregled povedi prve in tretje gruče nakazuje rabo besede predvsem v dobesednem pomenu, tj. nanašajoč se na steber kot gradbeni element. Primeri takih sintagem so »stebrno podnožje« (= podnožje iz stebrov), »stebrni okvir« (= okvir iz stebrov), »stebrni obod« (= obod iz stebrov), »stebrna dvorana« (= dvorana s stebri). V drugi gruči rab/pomenov se poleg dobesednih pojavijo tudi metonimične rabe, kot je »stebrni red« (stil stebrov), in metaforične rabe, kot so »stebrna spremljava« (poosebitev), »(tro-)stebrni sistem pokojnin«, »stebrni mit (kulturne industrije)« ali »stebrni plašč«. V letu 2018 se raba popolnoma spremeni. Tu močno prevladujeta gruči, ki se nanašata na pojav t. i. *stebrnega udara,* nesreče v rudniku, pri kateri pride do zrušitve (varnostnega) stebra. Gre za termin, pri katerem lahko prepoznamo metaforično motiviranost, saj ne gre za *steber* kot gradbeni element, temveč za *hribino*, puščeno pri izkopu rudnika, ki je prvemu podobna po svoji podporni funkciji. Glede na kontekste rabe v povedih ugotavljamo, da jih sistem v dve različni gruči najverjetneje razvršča glede na skladenjske lastnosti: medtem ko se v rdeči gruči zveza v veliki večini pojavlja zgolj v imenovalniku, se v oranžni gruči zveza uporablja le v neimenovalniških sklonih.

Slika 4: Distribucije rab besede *stebrn* v obdobju 1990–97 in letu 2018

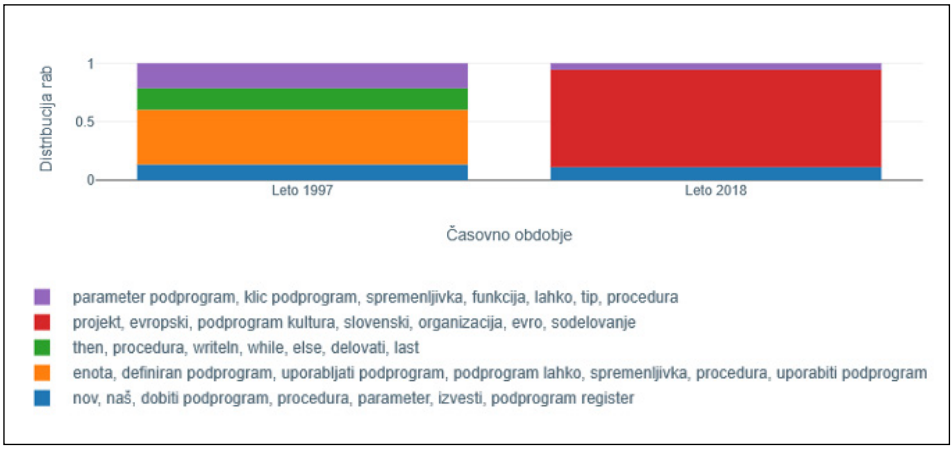


Vir: lastno delo

Tretja beseda po vrsti je pridevnik *jonski*. V obdobju 1990–1997 se pojavlja v mešanih rabah in kontekstih, ki se nanašajo na Jonce. Povečini gre za metonimično rabo (»jonski tempelj«, »jonska mesta«, »jonska šola«), pojavlja se tudi čisto dobesedna raba (»jonski Grki«, »jonski pomorščaki«). V letu 2018 močno prevladuje zgolj ena vrsta rabe/pomena, kjer se pridevnik ne nanaša na Jonce, temveč na geografsko regijo, pokrajino. Tu gre za rabo besede v zvezah »jadransko jonska (makro) regija«, »jadransko jonska pobuda«, »jadransko jonski koridor«, »jadransko jonska

strategija«, ki se v veliki meri pojavljajo v novičarskem žanru in političnem kontekstu. Pojav in porast teh rab je mogoče povezati s specifičnim dogodkom oziroma dogajanjem med obdobjema, in sicer predvsem z oblikovanjem »jadransko-jonske makroregijske strategije« leta 2014 kot združenja držav članic znotraj Evropske unije ter drugih držav v geografski regiji.[51]

Slika 5: Distribucije rab besede *podprogram* v obdobju 1990–97 in letu 2018



Vir: lastno delo

Prvi samostalnik na seznamu je beseda *podprogram*. Beseda je precej pogostejša v obdobju devetdesetih let, kjer naj bi se pojavljala v štirih različnih rabah (Slika 5). Te štiri gruče so opredeljene s podobnimi ključniki, med drugim *parameter, klic, spremenljivka, funkcija, tip, procedura; then, procedura, writeIn, while, else.* Kot dokazujejo tudi konteksti rabe (povedi), se beseda v teh gručah nanaša na računalniški pomen, ki je obeležen v slovarju: 'program v okviru določenega programa, ki se lahko večkrat uporabi v istem ali v drugem programu'. Raba v letu 2018 kaže na pojav in veliko prevlado drugačnega pomena besede, ki ga predstavlja gruča s ključniki *projekt, evropski, podprogram kultura, slovenski, organizacija, evro, sodelovanje*. Pomena ni mogoče najti v slovarju neposredno pod leksemom *podprogram*, temveč pod prvim pomenom pomenskega korena besede oziroma pod leksemom *program*: 'skupek nalog, del, ki se določijo za uresničitev'.[52] Primer kaže v prvem obdobju zožitev pomena na specifični računalniški pomen korena *program*, ki je prav tako edini slovarski pomen, ki sovpada s pojavom interneta, prvim prevodom operacijskega sistema Windows v slovenščino in razvojem drugih informacijsko-komunikacijskih tehnologij v devetdesetih letih.

Zanimivo je, da je povsem nasproten trend viden pri samostalniku *portal* na šestem mestu v tabeli. Tu je v obdobju do 1997 mogoče zaznati rabo besede v treh gručah, opredeljenih med drugim s ključniki *gotski portal, okno, ohranjen, avtocesta biti;*

51 Evropska komisija, »EU Strategy for the Adriatic and Ionian Region,« https://ec.europa.eu/regional_policy/policy/cooperation/macro-regional-strategies/adriatic-ionian_en, dostop 15. 4. 2025.

52 *Slovar slovenskega knjižnega jezika*, druga, dopolnjena in deloma prenovljena izdaja, pridobljeno 1. 2. 2025, www.fran.si.

renesančen portal, kamnit portal, pročelje. Le nekaj primerov rabe je za gručo s ključniki *spleten, portal, portal lahko, podatek, medij, podjetje, slovenski portal, informacija.* Po drugi strani ta in v letu 2018 novonastala gruča s ključniki *poročati portal, spleten portal, hrvaški portal, portal siol, navajati portal, pisati portal, novičarski portal* močno prevladujeta v drugem obdobju, kjer bolj dobesedna raba iz domene arhitekture, gradbeništva praktično izgine. Zanimivo je, da je arhitekturni pomen 'arhitektonsko poudarjen vhod v stavbo'[53] v prvi različici SSKJ še edini pomen, medtem ko se v drugi različici (SSKJ2) že pojavi novi. Pri tem gre za metaforično razširitev etimološko starejšega pomena, ki je v novejši različici slovarja definiran kot 'spletna stran, ki na pregleden način združuje dostop do različnih informacij in storitev'.[54] V novi različici je novi pomen (glede na pogostost rabe, zaznane s tem sistemom, povsem upravičeno) že postavljen na prvo mesto.

Naslednji primer kaže nekoliko manj očitne spremembe v rabi oziroma rabo besede *izbijanje* v zelo podobnih pomenih in kontekstih. Slovar besedo razlaga zgolj z definicijo »glagolnik od izbijati«,[55] medtem ko sistem razločuje štiri gruče rabe. V obdobju 1990–1997 je najpogostejša nevezljiva raba v pomenu 'balinanje', in sicer bodisi samostojno bodisi z levim prilastkom *hitrostno*, *precizno*, *natančno*. V istem obdobju je prisotna, četudi mnogo manj pogosta, raba besede z desnim prilastkom v zvezah »izbijanje žoge«, »izbijanje balina«, »izbijanje ploščka«. V drugem obdobju, tj. v letu 2018, raba v smislu 'balinanja' popolnoma izgine. Poleg že omenjene rabe z desno vezljivostjo sistem v tem obdobju zazna še dve gruči, kjer z analizo primerov ugotovimo, da je beseda *izbijanje* tu večinoma negativno modificirana: »neuspešno izbijanje«, »poskus izbijanja (žoge)«, »po slabem izbijanju«. Sistem v tem primeru rabo razločuje na podlagi resnično subtilnih razlik, ki jih ni mogoče ugotoviti brez vpogleda v kontekst rabe.

Spremembe v zaporednih obdobjih

Primer uporabniškega vmesnika za vhodni korpus, sestavljen iz petih zaporednih časovnih obdobij, smo že prikazali na Sliki 1. Besede so privzeto razvrščene po meri *JSD K5 All*, ki meri razliko med distribucijama v rabi besede med prvim in zadnjim obdobjem v korpusu (angl. beseda »All« označuje, da gre za spremembo v rabi besede od prvega do zadnjega obdobja).[56]

Glede na ta kriterij se je, tako kot v prejšnjem poglavju, najbolj spremenila distribucija rab besede *diagonalen*. S pomočjo vrednosti v drugih stolpcih, ki prikazujejo spremembe med zaporednimi obdobji, opazimo, da je k spremembi na dolgi rok

53 *Slovar slovenskega knjižnega jezika*, pridobljeno 1. 2. 2025, www.fran.si.

54 *Slovar slovenskega knjižnega jezika*, druga, dopolnjena in deloma prenovljena izdaja, pridobljeno 1. 2. 2025, www.fran.si.

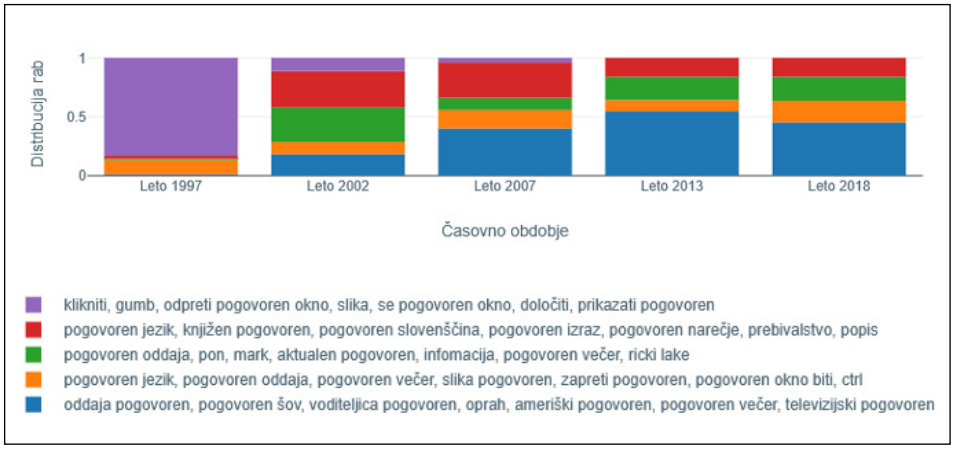
55 Ibidem.

56 Četudi korpus za merjenje sprememb v zaporednih obdobjih zajema isti dve skrajni obdobji in nabor besedil kot korpus za dolgoročno merjenje sprememb, lahko zaradi zajema vseh besedil in obdobij naenkrat pride do drugačnega gručenja primerov in posledično distribucij.

najbolj vplival prehod med obdobjema 2007 in 2013 (vrednost JSD je približno 0,38). Podobno kot v primerjavi rabe v obdobjih 1990–97 in 2018 iz prejšnjega poglavja gre pri tej zaznani spremembi predvsem za zožitev konteksta. Iz splošne rabe v zvezah »diagonalni korak«, »diagonalna črta«, »diagonalna razdalja«, kjer beseda modificira različne samostalnike, se raba v letu 2013 prevesi v praktično izključno (nogometni) športni kontekst, ki ga nakazujejo zveze »diagonalni strel«, »diagonalni predložek«, »diagonalna podaja«.

Drugi pridevnik, ki ga obravnavamo, je beseda *pogovoren*, katere največjo spremembo je sistem zaznal s prehodom med obdobjema 1990–97 in 2002. Podobno kot pri besedi *podprogram* iz prejšnjega poglavja lahko s pomočjo prikaza na Sliki 6 v prvem obdobju opazimo veliko prevlado gruče (več kot 80 odstotkov), ki predstavlja rabo v računalniškem kontekstu s ključnimi izrazi *klikniti, gumb, pogovorno okno, slika*. Po drugi strani se v naslednjem obdobju, tj. v letu 2002, raba besede ponovno posploši, saj je skoraj enakomerno razdeljena med vsemi petimi zaznanimi gručami. Pojavlja se v različnih zvezah, kot so »pogovorni jezik«, »pogovorna oddaja«, »pogovorni šov«, »pogovorna slovenščina«, »pogovorno okno«.

Slika 6: Distribucije rab besede *pogovoren* v petih obdobjih. Največja sprememba je vidna v prvih dveh stolpcih, tj. obdobjih 1997 in 2002.



Vir: lastno delo

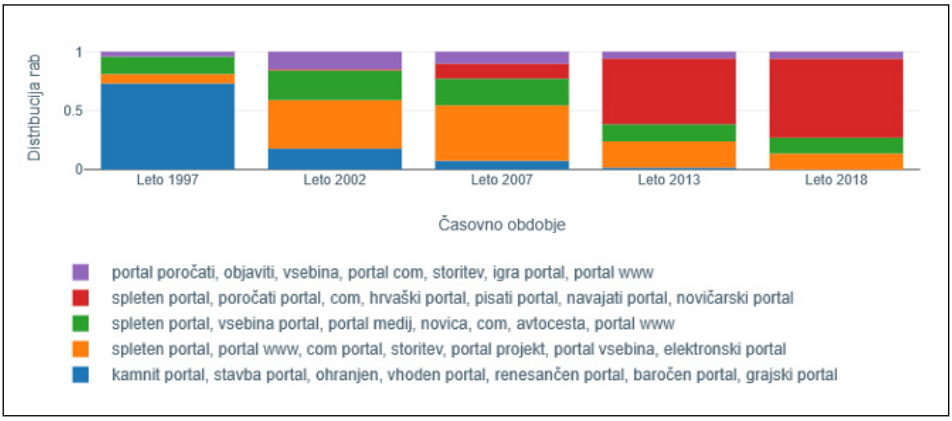
Še en pridevnik na seznamu je beseda *težavnosten.* Sistem največjo spremembo v rabi zazna med obdobjema 2007 in 2013. Pri tem je najvidnejši upad dveh gruč, ki ju zaznamujejo na primer *težavnostna stopnja, godba, vzpon, zahteven, proga* in *težavnostna stopnja, težavnostna skupina, vaja, težavnostni izpit.* Iz primerov rabe ugotovimo, da v obeh prevladuje predvsem zveza »težavnostna stopnja«, pojavi se še ob besedah *skupina, razred, sezona, kategorija, nivo*. Fraze so umeščene v raznovrstne kontekste, denimo športni (»kolesarski izleti različnih težavnostnih stopenj«), umetniški (»godbe v prvi težavnostni stopnji«), zdravstveni (»težavnostna stopnja jecljanja«),

šolski, igričarski idr. V sledečem obdobju pa se poveča raba v gruči, ki jo zaznamujejo *težavnostno plezanje, težavnostni pokal, plezalka, sezona, Janja Garnbret*. Povedi gruče potrjujejo, da se pridevnik tu pojavlja izključno v kontekstu »težavnostnega plezanja«, tj. je prišlo v letu 2018 do izrazite zožitve rabe. Predvidevamo, da je prevlada gruče posledica predvsem medijskega poročanja o uspehih specifične slovenske plezalke, ki je pozornost prvič pritegnila z nastopom na svetovnem prvenstvu leta 2016.[57]

Prvi samostalnik med najbolj spremenjenimi besedami je *evro* na tretjem mestu v tabeli. Zanimivo je, da je največja sprememba po meri JSD zaznana med obdobjema 1990–97 in 2002 (in ne na primer na pragu leta 2007, ko je Slovenija uvedla valuto). V obdobju devetdesetih let se izmenjujeta dve gruči, opredeljeni s ključniki *območje evra, indeks, eurostoxx, uvedba evra, tečaj evra, evropska centralna banka* ter *evropska borza, cena nafte, neenotno, valutni trg.* Ključni izrazi in primeri rabe kažejo, da se beseda *evro* v teh gručah uporablja v bolj generičnem, abstraktnem kontekstu pomena 'denarna enota'. Konteksti vključujejo napovedi vzpostavljanja »evro območja« in načrte vpeljave nove valute. V letu 2002, ko valuta dejansko že zamenja lokalne valute, se pojavi konkretnejša raba besede v bolj specifičnih kontekstih (»500 evrov, »100 evrov«, »milijon evrov«).

Drugo mesto med najbolj spremenjenimi samostalniki, kot pri dolgoročnih spremembah, tudi v tem razseku korpusa zaseda beseda *portal.* Glede na različnost distribucij se je največja sprememba v rabi zgodila med obdobjema 1997 in 2002, kjer je opaziti najvidnejši upad v konkretni rabi, tj. v pomenu gradbenega elementa. V prvem obdobju namreč ta raba predstavlja veliko večino (73 odstotkov) primerov, v letu 2002 pa že pade na manj kot 18 odstotkov. Vse večji upad rabe po posamičnih obdobjih lahko spremljamo na Sliki 7.

Slika 7: Distribucija rab besede *portal* v petih zaporednih obdobjih. Viden je izrazit upad dobesedne rabe (modra gruča) po letu 1997.

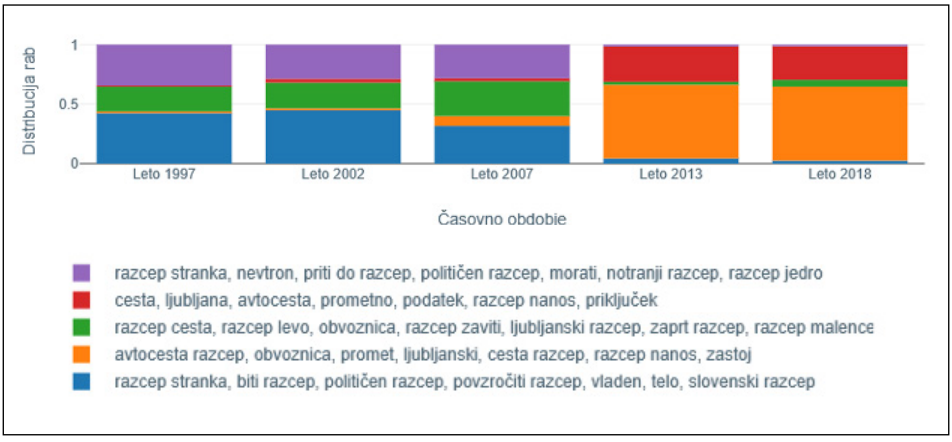


Vir: lastno delo

57 »Janja Garnbret pri 17 splezala na vrh sveta,« *MMC RTV-SLO*, nazadnje spremenjeno 17. 9. 2016, https://www.rtvslo.si/sport/preostali-sporti/janja-garnbret-pri-17-splezala-na-vrh-sveta/403013.

Tretji samostalnik, ki odraža največ sprememb v zaporednih obdobjih glede na skupni seštevek, je *razcep* (Slika 8). Največ prispeva primerjava obdobij 2007 in 2013. V prvih treh obdobjih je raba skoraj enakomerno razporejena med tri prevladujoče gruče, in sicer vijolično s ključnimi izrazi *razcep stranke, nevtron, politični razcep, notranji razcep, razcep jedra,* modro z izrazi *razcep stranke, politični razcep, povzročiti razcep, vladen, telo, slovenski razcep* in zeleno z izrazi *razcep ceste, razcep levo, obvoznica, zaprt razcep.* Četudi ključni izrazi v prvih dveh gručah nakazujejo rabo le v političnem in fizikalnem kontekstu, primeri uporabe pokažejo zelo raznovrstno metaforično rabo: »notranji razcep« (osebe), »razcep na levo ali desno«, »verski razcep«, »razcep med demokrati«, »razcep med človekom in svetom«, »generacijski razcep«, »razcep med umom in telesom«, »razcep na dve identiteti«. Modra gruča vsebuje tudi nekaj primerov, kjer so razvidne bolj fizikalne in konkretne rabe: »razcep jeder«, »jedrni razcep«. Razlika med prvo in drugo gručo je videti zgolj skladenjske narave, v primerih rabe iz modre gruče se *razcep* pojavlja le v imenovalniku. Tretja oziroma zelena gruča zaznamuje rabo v slovarskem pomenu 'vsaka od cest, prog, ki nastane z razcepitvijo ceste, proge'.[58] V zadnjih dveh obdobjih, tj. 2013 in 2018, pa se te rabe skoraj popolnoma umaknejo, v korpusu prevladujeta rdeča in oranžna gruča. Konteksti rabe vsebujejo enake ali vsebinsko zelo podobne izraze *cesta, ljubljana, prometno, priključek, obvoznica, promet, zastoj.* Po ključnih besedah se tematika rabe ujema z zeleno gručo. Iz primerjave povedi teh treh »cestnih« gruč pa ugotavljamo, da sistem ni razločil pomenskih, temveč žanrske in stilistične razlike. Za zeleno je namreč značilna bolj pripovedna, mestoma subjektivna raba, za oranžno in rdečo pa obvestilna raba s suhoparnim, objektivnim slogom.

Slika 8: Distribucije rab besede razcep v petih zaporednih obdobjih. Največja sprememba je vidna pri prehodu iz 2007 v 2013.



Vir: lastno delo

58 *Slovar slovenskega knjižnega jezika*, druga, dopolnjena in deloma prenovljena izdaja, pridobljeno 1. 2. 2025, www.fran.si.

Analiza reprezentacije migracij

V prejšnjih poglavjih smo pokazali, da lahko sistem uporabimo za analizo sprememb v rabi besed v različnih obdobjih. Meje med obdobji so bile določene glede na razpoložljive podatke, korpus Gigafida 2.0 smo razdelili na dve in pet obdobij, da smo preverili, kako uspešen je sistem pri zaznavanju dolgoročnih sprememb in sprememb v več zaporednih obdobjih.

V tem poglavju nas po drugi strani zanima, kako so specifični dogodki, teroristični napad v ZDA 11. septembra 2001 in obdobje »begunske krize« oziroma »dolgega poletja migracij« (2015–2016), vplivali na reprezentacijo fenomena migracij v slovenski družbi. V ta namen smo korpus Gigafida razdelili na pet jasno zamejenih obdobij:[59]

– predobdobje (1995–97);
– čas terorističnega napada (2001–02) v ZDA 11. septembra 2001, ki mu sledi načeloma
– nevtralno obdobje (2010–11);
– obdobje množičnih migracij v Evropi po zahodnobalkanski poti, najpogosteje poimenovan »begunska kriza« (2015–16), in
– poobdobje (2017–18).

Sestava podkorpusov je navedena v Tabeli 2.

Tabela 2: Velikost korpusa za analizo reprezentacije migracij

Obdobje	Št. dokumentov	Št. besed	Št. virov
Pred-obdobje (1995-1997)	4.577	43.300.937	18
9.11 (1.8.2001-31.3.2002)	1.198	47.554.430	22
Nevtralno (1.1.2010-31.12.2011)	1.473	33.707.077	90
Begunska kriza (1.8.2015-31.3.2016)	212	21.123.369	8
Po-obdobje (1.8.2017-31.3.2018)	273	26.775.448	8

Vir: lastno delo

Med besedami, ki so spremenile rabo med temi petimi obdobji, obravnavamo dva specifična primera, *burka* in *pritok*.

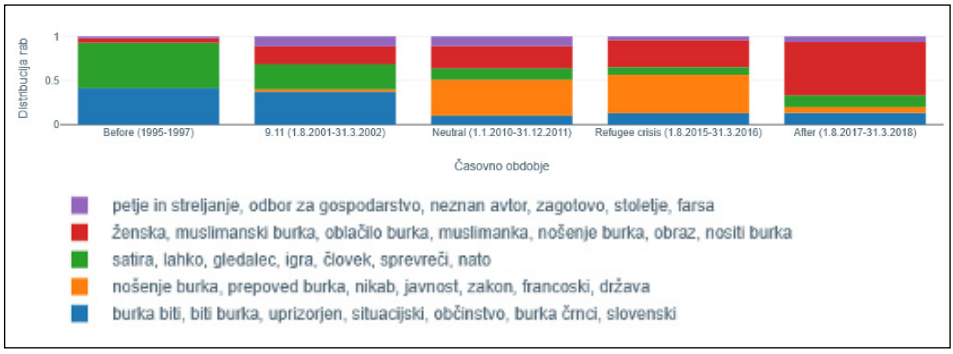
Besedi *burka* in *pritok* smo za analizo izbrali glede na povezanost s tematiko migracij. Med višje uvrščenimi besedami glede na spremembo rabe je bila vrsta besed, ki odražajo splošnejšo spremembo rabe (npr. severnomorski, rafiniran, evro), nas pa je zanimala specifika jezika v zvezi s pojavom migracij. Slika 9 ponazori spremembo v rabi besede *burka*. V obdobju pred napadi v ZDA 11. septembra 2001 je razumevanje besede povezano predvsem s pomenoma 'norčavo vedenje ali govorjenje' ter 'dramsko delo s šaljivo, včasih grobo vsebino, komiko'.[60] Raba besede v pomenu muslimanskega ženskega oblačila v petih obravnavanih obdobjih narašča in je najpogostejša v zadnjem

59 Sistem za analizo sprememb v rabi besed med temi petimi obdobji je dostopen na (*E8-NLP*) http://kt-nlp-demo.ijs.si:8080/semanticshifttable/6.

60 *Slovar slovenskega knjižnega jezika*, druga, dopolnjena in deloma prenovljena izdaja, pridobljeno 1. 2. 2025, www.fran.si.

obdobju (2017–2018). Treba je poudariti, da gre tu v resnici za dve izvorno različni besedi: eno je burka iz družine *burkež, burkati* ipd., druga je *burqa* – žensko muslimansko oblačilo. Tu torej povečana raba besede burka v določenem časovnem obdobju ni odraz pomenskega premika, pač pa posledica prevzema besedne oblike (burqua) iz tujega jezika, ki sovpada (homograf) z v jeziku že obstoječo besedo. Vstop besede *burqa* v prostor prej obstoječe besede *burka* je v tem primeru sociolingvistično pogojen, dejstvo, da sistem zaznava ta prevzem prostora, pa pokaže, da je sistem mogoče uporabiti tudi za sociološko analizo.

Slika 9: Sprememba v rabi besede *burka*



Vir: lastno delo

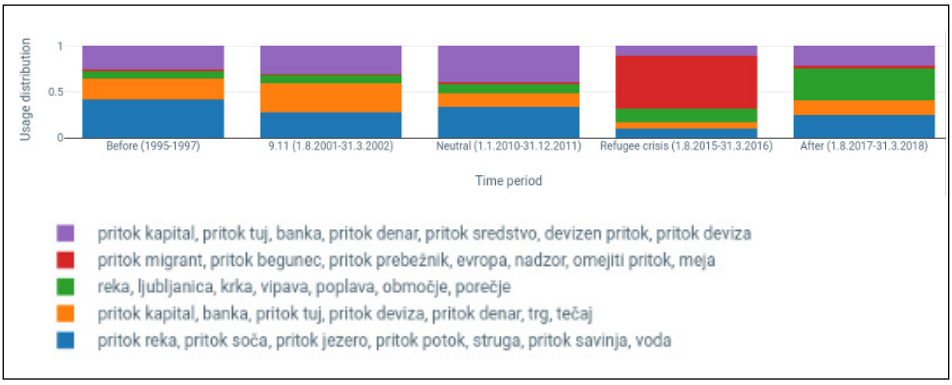
Uporaba besede *burka* v smislu ženskega oblačila začne naraščati takoj po zrušenju dvojčkov WTC v ZDA in postane prevladujoča v času razprave o prepovedi nošenja burke oziroma nikaba v javnosti, ki je tudi v Sloveniji potekala predvsem v smislu, ali naj se na ravni države to zakonsko prepove (kot denimo velja v Franciji vse od leta 2011). Ta vidik je bil pričakovano najbolj izpostavljen v obdobju po napadu na ZDA, ki mu je sledila napoved t. i. vojne proti terorizmu (angl. *the war on terror*), ter v času t. i. begunske krize, ko je ozemlje Slovenije kot ene od držav na zahodnobalkanski migracijski poti v obdobju 2015–2016 prečkalo 400.000 beguncev, za katere se je predvidevalo, da so muslimanske veroizpovedi. Prvotni humanitarni vladni odziv je zamenjala kriminalizacija migracij. Po podatkih Eurobarometra je odstotek anketirancev, ki so navajali priseljevanje kot ključno vprašanje, s katerim se sooča EU, s 25 odstotkov leta 2014 narastel na skoraj 40 odstotkov v letu 2015, priseljevanje ljudi iz držav zunaj EU pa je vzbujalo negativne občutke kar pri 56 odstotkih vprašanih.[61] V Sloveniji se je širil protibegunski in protipriseljenski sovražni govor, v javnem diskurzu pa je tema migracij postajala vse bolj žgoča in polarizirajoča.[62] Najobsežnejša pa je uporaba

61 Evropska komisija, »Standard Eurobarometer 83 – Spring 2015,« pridobljeno 25. 2. 2024, https://europa.eu/eurobarometer/surveys/detail/2099.

62 Za več gl. Veronika Bajt in Ajda Šulc, »Medijsko ustvarjanje protibegunskega sovražnega govora v komentarjih na Facebooku,« *Javnost: The Public* 31, sup 1 (2024): 48–66. Boris Vezjak, »Radical Hate Speech: The Fascination with Hitler and Fascism on the Slovenian Webosphere,« *Šolsko polje* 29, št. 5-6 (2018): 133–51. Maruša Pušnik, »Dinamika novičarskega diskurza populizma in ekstremizma: moralne zgodbe o beguncih,« *Dve domovini* 45 (2017): 137–52.

besede burka v smislu ženskega oblačila v obdobju po ključnih dveh časovnih točkah v poobdobju, kar sovpada z globalnim porastom razprave o migracijah kot problemu, predvsem zaradi domnevne nezdružljivosti islama z zahodno oziroma evropsko (in slovensko) kulturo.[63] V zadnjem obdobju tako pri rabi besede prevladuje vidik spola, razprava pa se osredotoči na muslimansko žensko.[64]

Slika 10: Sprememba v rabi besede *pritok*



Vir: lastno delo

Zanimiva je tudi sprememba v rabi besede *pritok* (Slika 10). Od prevladujoče povezave *pritoka* z vodo (»pritok reke«) v drugi polovici devetdesetih let, ki kaže dobesedno rabo v osnovnem pomenu besede, se v drugem (in tudi tretjem) obdobju kaže metaforični pomen z navezavo na denar, banke in devize (npr. »pritok kapitala«). Očiten porast v rabi v povezavi z migracijami je videti v obdobju »begunske krize« z rabo besede v zvezah »pritok migrantov/beguncev/prebežnikov«. V tem obdobju je sprememba v rabi povezana s političnim dogajanjem v Evropi, kjer v ospredje preide problematika omejevanja in upravljanja migracij ter preprečevanje vstopa beguncem, kar potrjujejo vse obstoječe raziskave medijskega poročanja (gl. npr. Pajnik 2017[65]). Nezaupanje do muslimanskih beguncev, ki naj bi kot neustavljiv »val« ali »reka« (tj. pritok) pritiskali na EU, je razširjeno po vsej Evropi in se povezuje z marginalizacijo muslimanskih priseljencev. Protibegunski diskurz v analiziranem obdobju se torej zaradi prevlade ali domnev o prevladi »izvora« prišlekov iz islamskih držav prepleta s predobstoječimi predsodki do islama in endemičnimi protimuslimanskimi stališči. V poobdobju tega več ni, se pa spet okrepi povezava z vodo in rekami.

63 Arun Kundnani, *The Muslims Are Coming: Islamophobia, Extremism, and the Domestic War on Terror* (Verso, 2015).

64 Sara R. Farris, *In the Name of Women's Rights: The Rise of Femonationalism* (Duke University Press, 2017), http://www.jstor.org/stable/j.ctv11sn2fp.

65 Mojca Pajnik, »Medijsko-politični paralelizem: Legitimizacija migracijske politike na primeru komentarja v časopisu *Delo*,« *Dve domovini* 45 (2017): 169–84.

Zaključek in nadaljnje delo

V članku smo predstavili prvi spletni sistem za zaznavanje sprememb v rabi besed v slovenščini. Pri tem smo podrobneje osvetlili njegovo tehnično zasnovo, metodo za zaznavanje besed in enostavno dostopen uporabniški vmesnik. Ta v enem koraku omogoča hiter pregled največjih sprememb v rabi na ravni celotnega korpusa, v drugem koraku pa podrobnejšo analizo na ravni posamezne besede.

Sistem smo nato uporabili in evalvirali s pomočjo jezikoslovne in sociolingvistične analize. V prvi smo podrobneje interpretirali rezultate sistema z vpogledom v pridevnike in samostalnike, katerih raba naj bi se najbolj spremenila. Pri tem smo gruče analizirali na ravni ključnih izrazov in dejanskih primerov rabe, ki jih prikaže sistem. Tako gruče kot dejanske rabe smo skušali kategorizirati v različne kategorije pomena in pomenskih premikov (dobesedni/osnovni, metaforični, metonimični) in tudi vzporejati s slovarskimi pomeni, obeleženimi v Slovarju slovenskega knjižnega jezika. Analiza je pokazala, da je sistem uporaben za odkrivanje različnih rab v širšem smislu, vendar pa same gruče večinoma ne ustrezajo zgolj semantiki, tj. pomenski plati posamičnih besed. Sistem v veliko primerih prikaže več gruč pogostih rab, kot jih dejansko obstaja, torej več, kot je pomenov v slovarju ali v rabi. Problem izhaja iz narave vektorskih vložitev, ki poleg semantične plati besed ujamejo tudi skladenjske in morfološke lastnosti besed pa tudi druge globalne vzorce, ki jih je mogoče zaznati v širšem kontekstu (v jeziku ponavljajoči se vzorci, kot so na primer stereotipi). Zaradi tega sistem v veliko primerih ustvari več gruč, ki pokrivajo semantično enako rabo oziroma isti leksikalni pomen besede, v različne gruče pa je ta pomensko enaka raba uvrščena zaradi nepomenskih razlik, kot je morfologija, skladnja, slog ali dolžina povedi ipd. Velja tudi obratno, tj. da ena gruča združuje sicer različne pomene s površinsko podobno rabo besede. Večje število gruč, kot je dejanskih rab, izhaja tudi iz metode gručenja, pri kateri je število gruč vnaprej določeno.

Druga omejitev sistema izhaja iz uporabljenih podatkov. O stanju slovenskega (standardnega) jezika in rabi sodimo glede na njegovo reprezentacijo v korpusu Gigafida. Četudi naj bi bil kot referenčni korpus slovenščine karseda reprezentativen in uravnotežen vir, je povsem mogoče, da na (navidezne) spremembe v rabi določenih besed vplivajo predvsem razlike v sestavi virov posameznih časovnih podkorpusov. Pri interpretaciji rezultatov, ki jih poda sistem, velja ohraniti previdnost, saj morda že sam korpus ne prikazuje ustrezne jezikovni realnosti.

V prihodnje načrtujemo uporabo sistema na novejših besedilnih korpusih v slovenščini, ki vsebujejo podatke o rabi besed po letu 2018. V načrtu so tudi raziskave sprememb v rabi besed za specifične primere in dogodke (npr. kako je na evolucijo raznovrstnih konceptov, nova poimenovanja in pomenske prenose vplivala pandemija covida, ki je glede na raziskave imela odločilen vpliv na evolucijo medijskega poročanja[66]). Prav tako bomo preizkusili nove metode za zaznavanje in interpretacijo sprememb v rabi besed in s tem poskušali izboljšati delovanje sistema, na primer z uporabo

66 Montariol et al., »Scalable and interpretable.«

drugega algoritma za gručenje ali bolj informirane metrike za merjenje sprememb v distribuciji rab. Nenazadnje pa se bomo osredotočili tudi na metode za odkrivanje skupine besed in konceptov, ki izražajo podobne spremembe v rabi – denimo iskanje besed, ki kažejo razširitve pomena specifično prek metafor, ali odkrivanje konceptov in semantičnih polj, ki kažejo največjo raznolikost pomenov.

Zahvala

Delo je bilo izvedeno v okviru projekta RSDO (*Razvoj slovenščine v digitalnem okolju*), ki sta ga financirala Ministrstvo za kulturo Republike Slovenije in Evropski sklad za regionalni razvoj, ter v okviru programov in projektov Javne agencije za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS): *Sovražni govor v sodobnih konceptualizacijah nacionalizma, rasizma, spola in migracij* (J5-3102), *Tehnike vektorskih vložitev za medijske aplikacije* (L2-50070), *Veliki jezikovni modeli za digitalno humanistiko* (GC-0002), *Računalniško podprta večjezična analiza novičarskega diskurza s kontekstualnimi besednimi vložitvami* (J6-2581), *Tehnologije znanja* (P2-0103), *Slovenski jezik - bazične, kontrastivne in aplikativne raziskave* (P6-0215) in *Enakost in človekove pravice v dobi globalnega vladovanja* (P5-0413).

Vira in Literatura

Literatura

Aitchison, Jean. *Language Change: Progress or Decay?*. Cambridge University Press, 2001.

Bajt, Veronika in Ajda Šulc. »Medijsko ustvarjanje protibegunskega sovražnega govora v komentarjih na Facebooku.« *Javnost - The Public* 31, sup 1 (2024): 48–66. https://doi.org/10.1080/13183222.2024.2443868.

Computational approaches to semantic Change , uredili Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu in Simon Hengchen. Language Science Press, 2021. https://doi.org/10.5281/zenodo.5040241.

Del Tredici, Marco, Malvina Nissim in Andrea Zaninello. »Tracing metaphors in time through self-distance in vector spaces.« V: *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-It 2016*, 117–22. Accademia University Press, 2016. https://doi.org/10.4000/books.aaccademia.1760.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee in Kristina Toutanova. »BERT: Pre-training of deep bidirectional transformers for language understanding.« V: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers)*. Association for Computational Linguistics, 2019, 4171–86. https://doi.org/10.18653/v1/N19-1423.

Erjavec, Tomaž, Nikola Ljubešić in Darja Fišer. »Korpus slovenskih spletnih uporabniških vsebin Janes.« V: *Viri, orodja in metode za analizo spletne slovenščine*, uredila Darja Fišer, 16–43. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 2018.

Farris, Sara R. *In the Name of Women's Rights: The Rise of Femonationalism.* Duke University Press, 2017. http://www.jstor.org/stable/j.ctv11sn2fp.

Fišer, Darja in Nikola Ljubešić. »Tviti kot leksikografski vir za analizo pomenskih premikov v slovenščini.« V: *Viri, orodja in metode za analizo spletne slovenščine*, uredila Darja Fišer, 198-226.. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 2018.

Gantar, Polona, Špela Arhar Holdt in Senja Pollak. »Leksikalne novosti v besedilih računalniško posredovane komunikacije.« *Slavistična revija* 66, št. 4 (2018): 459–72.

Gillani, Nabeel in Roger Levy. »Simple dynamic word embeddings for mapping perceptions in the public sphere.« V: *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, 2019, 94–99.

Giulianelli, Mario, Marco Del Tredici in Raquel. Fernández. »Analysing lexical semantic change with contextualised word representations.« V: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 3960–73. Association for Computational Linguistics, 2020. https://www.aclweb.org/anthology/2020.acl-main.365.

Gribomont, Isabelle. »From Diachronic to Contextual Lexical Semantic Change: Introducing Semantic Difference Keywords (SDKs) for Discourse Studies.« V: *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, 153–60. Association for Computational Linguistics, 2023.

Hamilton, William L., Jure Leskovec in Dan Jurafsky. »Diachronic word embeddings reveal statistical laws of semantic change.« V: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* 1489–501. Association for computational linguistics, 2016. http://doi.org/10.18653/v1/P16-1141.

Harris, Zellig S. »Distributional Structure.« *WORD* 10, št. 2–3 (1954): 146–62.

Hilpert, Martin in Stefan Th. Gries. »Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition.« *Literary and Linguistic Computing* 24, št. 4 (2008): 385–401.

Juola, Patrick. »The time course of language change.« *Computers and the Humanities* 37, št. 1 (2003): 77–96.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde in Slav Petrov. »Temporal analysis of language through neural language models.« V: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science* (2014): 61–65. http://doi.org/10.3115/v1/W14-2517.

Krek, Simon, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraž Repar, Polona Gantar idr. »Gigafida 2.0: the reference corpus of written standard Slovene.« V: *Proceedings of the 12th Language Resources and Evaluation Conferenc*e, 3340–45. ELRA, 2020.

Kundnani, Arun. *The Muslims are Coming: Islamophobia, Extremism, and the Domestic War on Terror*. Verso, 2015.

Kutuzov, Andrey in Mario Giulianelli. »UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection.« V: *Proceedings of the Fourteenth Workshop on Semantic Evaluation,* 126–34. International Committee for Computational Linguistics, 2020. https://www.aclweb.org/anthology/2020.semeval-1.14.

Lakoff, George in Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

Lin, Jianhua. »Divergence measures based on the Shannon entropy.« *IEEE Transactions on Information Theory* 37, št. 1 (1991): 145–51.

Ljubešić, Nikola, Luka Terčon in Kaja Dobrovoljc. »CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.« V: *Zbornik konference za jezikovne tehnologije in digitalno humanistiko (JT-DH-2024)*, uredila Špela Arhar Holdt in Tomaž Erjavec. 251–74. Ljubljana: Inštitut za novejšo zgodovino, 2024. https://doi.org/10.5281/zenodo.13936406.

Martinc, Matej, Veronika Bajt, Špela Rot in Senja Pollak. »Sistem za zaznavanje sprememb v rabi besed in njegova uporaba za sociolingvistično analizo.« V: *Zbornik konference Jezikovne tehnologije*

in digitalna humanistika 2024, 298–318. Ljubljana: Inštitut za novejšo zgodovino, 2024. https://doi.org/10.5281/zenodo.13936410.

Martinc, Matej, Petra Kralj Novak in Senja Pollak. »Leveraging contextual embeddings for detecting diachronic semantic shift.« V: *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 4811–19. ELRA, 2020. https://aclanthology.org/2020.lrec-1.592.

Martinc, Matej, Syrielle Montariol, Elaine Zosa in Lidia Pivovarova. »Capturing evolution in word usage: Just add more clusters?.« V: *Companion Proceedings of the Web Conference 2020,* 343–49. Association for Computing Machinery, 2020. https://doi.org/10.1145/3366424.3382186.

Martinc, Matej, Nina Perger, Andraž Pelicon, Matej Ulčar, Andreja Vezovnik in Senja Pollak. »EMBEDDIA hackathon report: Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+.« V: *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation,* 121–26. 2021.

Martinc, Matej, Nina Perger in Senja. Pollak. »Viewpoint detection on LGBT+ reporting using contextual embeddings and qualitative thematic analysis: The use case on the word *deep*.« *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 165–166, št. 1–2 (2025): 154–85. https://doi.org/10.1177/07591063251317085.

Menéndez, María L., Julio A. Pardo, Leandro Pardo in María C. Pardo. »The Jensen-Shannon divergence.« *Journal of the Franklin Institute* 334, št. 2 (1997): 307–18, https://doi.org/10.1016/S0016-0032(96)00063-4.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado in Jeff Dean. »Distributed representations of words and phrases and their compositionality.« *Advances in Neural Information Processing Systems* 26 (2013).

Montariol, Syrielle, Matej Martinc in Lidia Pivovarova. »Scalable and interpretable semantic change detection.« V: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies,* 4642–52. ACL, 2021.

Pajnik, Mojca. »Medijsko-politični paralelizem. legitimizacija migracijske politike na primeru komentarja v časopisu Delo.« *Dve domovini / Two Homelands* 45 (2017): 169–84.

Pranjić, Marko, Kaja Dobrovoljc, Senja Pollak in Matej Martinc. »Semantic change detection for slovene language: a novel dataset and an approach based on optimal transport.« *arXiv:2402.16596* (arXiv preprint, 2024). https://doi.org/10.48550/arXiv.2402.16596.

Pušnik, Maruša. »Dinamika novičarskega diskurza populizma in ekstremizma: moralne zgodbe o beguncih.« *Dve domovini / Two Homelands* 45 (2017): 137–52.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky in Nina Tahmasebi. »SemEval-2020 task 1: Unsupervised lexical semantic change detection.« V: *Proceedings of the Fourteenth Workshop on Semantic Evaluation.* International Committee for Computational Linguistics, 2020, 1–23. https://www.aclweb.org/anthology/2020.semeval-1.1.

Snoj, Jerica. »Slovarska večpomenskost in Slovensko leksikalno pomenoslovje.« *Slavistična Revija* 51, št. 4 (2003): 387–409.

Sweetser, Eve. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure.* Cambridge University Press, 1990.

Tahmasebi, Nina, Lars Borin in Adam Jatowt. »Survey of computational approaches to lexical semantic change detection.« V: Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu in Simon Hengchen, ur. *Computational Approaches to Semantic Change,* uredili Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu in Simon Hengchen. Language Science Press, 2021, 1–91. https://doi.org/10.5281/zenodo.5040302.

Tang, Xuri. »A state-of-the-art of semantic change computation,« *Natural Language Engineering* 24, št. 5 (2018): 649–76.

Ulčar, Matej in Marko Robnik Šikonja. »SloBERTa: Slovene monolingual large pretrained masked language model.« V: *Zbornik 24. mednarodne multikonference Informacijska družba 2021, zvezek C,* 17-20. Ljubljana: Institut »Jožef Stefan«, 2021.

Vezjak, Boris. »Radical Hate Speech: The Fascination with Hitler and Fascism on the Slovenian Webosphere.« *Šolsko polje* 29, št. 5–6 (2018): 133–51.

Wei, Yuting, Meiling Li, Yangfu Zhu, Yuanxing Xu, Yuqing Li in Bin Wu. »A diachronic language model for long-time span classical Chinese.« *Information Processing & Management* 62, št. 1 (2025), 103925. https://doi.org/10.1016/j.ipm.2024.103925.

Vidovič Muha, Ada. *Slovensko leksikalno pomenoslovje: govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske fakultete, 2000.

Würschinger, Quirin in Barbara McGillivray. »Semantic change and socio-semantic variation: the case of COVID-related neologisms on Reddit.« *Linguistics Vanguard*, 2024. https://doi.org/10.1515/lingvan-2023-0106.

Zamora-Reina, F. D., F. Bravo-Marquez in D. Schlechtweg. »LSCDiscovery: A shared task on semantic change discovery and detection in Spanish.« V: *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change,* 149–64. Association for Computational Linguistics, 2022.

Spletni viri

Evropska komisija. »Standard Eurobarometer 83 - Spring 2015.« Pridobljeno 24. 2. 2024. ttps://europa.eu/eurobarometer/surveys/detail/2099.

Evropska komisija. »EU Strategy for the Adriatic and Ionian Region.« Pridobljeno 15. 4. 2025. https://ec.europa.eu/regional_policy/policy/cooperation/macro-regional-strategies/adriatic-ionian_en.

MMC RTV-SLO. »Janja Garnbret pri 17 splezala na vrh sveta.« Nazadnje spremenjeno 17. september 2016. https://www.rtvslo.si/sport/preostali-sporti/janja-garnbret-pri-17-splezala-na-vrh-sveta/403013.

Slovar slovenskega knjižnega jezika. Druga, dopolnjena in deloma prenovljena izdaja. Pridobljeno 1. 2. 2025. www.fran.si.

Slovar slovenskega knjižnega jezika. Pridobljeno 1. 2. 2025. www.fran.si.

Mojca Brglez, Veronika Bajt, Senja Pollak, Špela Rot, Matej Martinc

A SYSTEM FOR WORD USAGE CHANGE DETECTION: ITS USE IN LINGUISTIC AND SOCIOLINGUISTIC STUDIES

SUMMARY

In this article, we present the first online system for detecting changes in Slovene word usage. We provide an in-depth overview of its technical design, the method for detecting words, and its user-friendly interface. The system provides a quick and concise general overview of the most significant usage changes across the entire corpus, while also allowing for a more detailed analysis at the level of individual words.

We demonstrate the application of the system on a Slovene reference corpus, delimited into different combinations of temporal slices, and evaluate the system

through its use for linguistic and sociolinguistic analysis. In the linguistic analysis, we closely examine the results of the system, focusing on the most altered adjectives and nouns. We analyse clusters at the level of key terms and real usage examples. Both the clusters and actual usage patterns are categorised into various semantic and usage-shift categories (basic/literal/ordinary, metaphorical, metonymic, broadening, narrowing) and compared with dictionary definitions. Our analysis concludes that the system is effective in detecting various usage patterns in a broad sense. However, the clusters generated do not always correspond strictly to semantic aspects, i.e., the senses of individual words. In many cases, the system identifies more clusters than actually exist in real use – more than the number of meanings recorded in dictionaries or observable in discourse.

On the one hand, this issue arises from the nature of vector embeddings, which capture not only the semantic aspects of words but also their syntactic and morphological properties, as well as other global patterns detectable in a broader linguistic context (e.g., recurring patterns in language, such as stereotypes). As a result, the system often generates multiple clusters that, in fact, represent the same semantic usage or lexical meaning. Conversely, some clusters combine distinct meanings due to their surface-level similarity in usage. Furthermore, the system sometimes classifies meaning-equivalent usages into different clusters based on non-semantic factors, such as morphology, syntax, style, or simply sentence length. On the other hand, the tendency to generate more clusters than would be observed in actual usage is also influenced by the clustering method itself, as the number of clusters is predetermined. A second limitation of the system stems from the dataset itself. Our insights into the state of the Slovenian (standard) language and its usage are based on its representation in the Gigafida corpus. Although this corpus is designed to be as representative and balanced a resource as possible for Slovenian, it is entirely possible that (apparent) changes in word usage are primarily influenced by differences in the composition of sources across different time-based subcorpora. Therefore, when interpreting the system's results, caution is advised, as the corpus itself may not accurately reflect linguistic reality.

1.01 DOI: https://doi.org/10.51663/pnz.65.3.08

Špela Arhar Holdt,* Magdalena Gapsa,♦
Polona Gantar,◦ Iztok Kosem•

Potencial ChatGPT pri razvoju Slovarja sopomenk sodobne slovenščine

IZVLEČEK

V raziskavi preverjamo, kako dobro se ChatGPT-4 odreže pri dveh slovaropisnih nalogah: (a) čiščenju seznama strojno pridobljenih sopomenskih kandidatov in umeščanju sopomenskega gradiva pod besedne pomene ter (b) izdelavi slovarskega gesla, vključno s pomensko členitvijo, definicijami in zgledi, na podlagi različnih vhodnih podatkov. Kot zlati standard upoštevamo slovaropisne odločitve, vključene v Digitalno slovarsko bazo za slovenščino. V prvem preizkusu analiziramo rezultate za 246 slovarskih iztočnic in ugotavljamo, da je ChatGPT podatke uredil povsem enako kot slovaropisci pri 41,9 odstotka iztočnic, pri 58,1 odstotka pa se je v odločitvi razlikoval. Pri presojanju relevantnosti sopomenskih kandidatov je bil ChatGPT popustljivejši od zlatega standarda. Razlike v razvrščanju sopomenk (umestitev pod drug pomen pri 14,6 odstotka iztočnic, manjkajoča umestitev pri 19,9 odstotka) deloma pripisujemo značilnostim vhodnih podatkov, kot sta kompleksnost naloge in kratkost pomenskih indikatorjev. V drugem preizkusu preverjamo zmožnost ChatGPT za samostojno izdelavo slovarskih gesel za 116 iztočnic. Analiza kakovosti generiranih pomenskih členitev in definicij kaže, da sistem deluje zmerno dobro: v 57 odstotkih primerov je zaznal vse pomene,

* **Dr., znan. sod., Univerza v Ljubljani, Filozofska fakulteta, Aškerčeva cesta 2, Ljubljana; Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana, spela.arharholdt@ff.uni-lj.si; ORCID: 0000-0003-0565-0531**

♦ **Inform. spec., Centralna tehniška knjižnica Univerze v Ljubljani, Trg republike 3, 1000 Ljubljana, magdalena.gapsa@ctk.uni-lj.si; ORCID: 0000-0003-2763-4495**

◦ **Dr., znan. sod., Univerza v Ljubljani, Filozofska fakulteta, Aškerčeva cesta 2, Ljubljana, apolonija.gantar@ff.uni-lj.si; ORCID: 0000-0001-5822-6414**

• **Dr., viš. znan. sod., Univerza v Ljubljani, Filozofska fakulteta, Aškerčeva cesta 2, Ljubljana; Institut »Jožef Stefan«, Jamova cesta 39, Ljubljana, iztok.kosem@ijs.si; ORCID: 0000-0002-4282-9031**

skoraj 80 odstotkov generiranih gesel je doseglo povprečno oceno 3,5 ali več, 19 odstotkov pa najvišjo oceno obeh ocenjevalcev. Glavni izzivi so pretirano drobljenje pomenov, neprepoznane prenesene rabe in manjša predvidljivost rezultatov. Sklenemo lahko, da ima ChatGPT potencial za pohitritev ročnega slovaropisnega dela, če se njegovi rezultati ustrezno preverjajo in nadgrajujejo.

Ključne besede: digitalno slovaropisje, ChatGPT, sopomenke, besedni pomen, slovenščina

ABSTRACT

THE POTENTIAL OF CHATGPT IN THE DEVELOPMENT OF THE THESAURUS OF MODERN SLOVENE

In this study, we examine how well ChatGPT-4 performs in two lexicographic tasks: (a) cleaning the list of automatically retrieved synonym candidates and assigning synonymic material to lexical senses, and (b) generating dictionary entries, including sense division, definitions, and examples, based on different input data. As a gold standard, we consider the lexicographic decisions recorded in the Digital Dictionary Database for Slovene. In the first experiment, we analyse the results for 246 dictionary entries and find that ChatGPT processed the data identically to lexicographers in 41.9 % of cases, while in 58.1 % of cases, it made different decisions. When assessing the relevance of synonym candidates, ChatGPT was more permissive than the gold standard. Differences in synonym placement (assignment to a different sense in 14.6 % of entries, missing placement in 19.9 %) can be partly attributed to input data characteristics, such as task complexity and the brevity of semantic indicators. In the second experiment, we test ChatGPT's ability to autonomously generate dictionary entries for 116 headwords. The analysis of generated sense divisions and definitions reveals that the system performs moderately well: in 57 % of cases, it identified all senses, almost 80 % of generated entries received an average score of 3.5 or higher, and 19 % received the highest score from both evaluators. The main challenges include excessive splitting of senses, failure to recognise figurative meanings, and reduced predictability of results. We conclude that ChatGPT has potential for speeding up manual lexicographic work if its results are properly monitored and refined.

Keywords: digital lexicography, ChatGPT, synonyms, word senses, Slovenian language

Uvod

Generativna umetna inteligenca, ki temelji na velikih jezikovnih modelih, je prek klepetalnih vmesnikov, kakršen je ChatGPT, postala široko dostopna za številne z jezikom povezane naloge.[1] Med področji, ki zadnja leta preizkušajo moč in omejitve novih tehnologij, je tudi slovaropisje.

Kot pričajo Rundell,[2] Lew,[3] Bartosz et al.,[4] McKean in Fitzgerald[5] ter Tiberius et al.,[6] se dosedanji preizkusi rabe ChatGPT za slovaropisne namene osredotočajo na generiranje bolj ali manj celostnih slovarskih gesel za (pogosto dokaj priložnostno) izbran nabor iztočnic. De Schryver v svojem kritičnem pregledu prvih prispevkov na temo z umetno inteligenco podprtega slovaropisja poroča, da je trenutno največ pozornosti posvečene definicijam in primerom rabe.[7] Skoraj vse študije oziroma preizkusi pa so bili izvedeni v angleščini in za angleščino, čeprav Jakubíček in Rundell naslavljata tudi problem večjezičnosti.[8]

Obstoječim raziskavam dodajamo dva preizkusa za slovenščino: (a) preizkus, kako dobro se ChatGPT-4 odreže pri čiščenju seznama strojno pridobljenih sopomenskih kandidatov in umeščanju sopomenskega gradiva pod besedne pomene, ter (b) preizkus izdelave slovarskega gesla (s pomensko členitvijo, definicijami in zgledi) na podlagi različnih vhodnih podatkov. Delo se povezuje z nadgrajevanjem Slovarja sopomenk sodobne slovenščine, velike zbirke slovenskih sopomenk, ki je bila v prvem koraku pripravljena povsem strojno iz podatkov Velikega angleško-slovenskega slovarja Oxford®-DZS in referenčnega korpusa Gigafida, kot opisujejo Krek, Laskowski in Robnik-Šikonja.[9] Od objave leta 2018 se slovar postopoma ročno pregleduje in

1 »ChatGPT (veliki jezikovni model),« OpenAI, pridobljeno 31. 5. 2024, https://chatgpt.com.

2 Michael Rundell, »Automating the Creation of Dictionaries: Are We Nearly There?,« v: *Proceedings of the 16th International Conference of the Asian Association for Lexicography* (Yonsei University, 2023), 9–17, pridobljeno 20. 5. 2025, https://www.asialex.org/pdf/Asialex-Proceedings-2023.pdf.

3 Robert Lew, »ChatGPT as a COBUILD Lexicographer,« *Humanities and Social Sciences Communications* 10 (2023), pridobljeno 20. 5. 2025, https://doi.org/10.1057/s41599-023-02119-6.

4 Bartosz Ptasznik, Sascha Wolfer in Robert Lew, »A Learners' Dictionary versus ChatGPT in Receptive and Productive Lexical Tasks,« *International Journal of Lexicography* 37, št. 3 (2024): 322–36, pridobljeno 20. 5. 2025, https://doi.org/10.1093/ijl/ecae011.

5 Erin McKean in Will Fitzgerald, »The ROI of AI in Lexicography,« *Lexicography* 11, št. 1 (2024): 7–27, pridobljeno 20. 5. 2025, https://utppublishing.com/doi/abs/10.1558/lexi.27569.

6 Carole Tiberius et al., »LLMs and Evidence-based Lexicography,« v: Simon Krek, ur., *Large Language Models and Lexicography*, 2024, 44–48, pridobljeno 25. 1. 2025, https://www.cjvt.si/wp-content/uploads/2024/10/LLM-Lex_2024_Book-of-Abstracts.pdf.

7 Gilles-Maurice de Schryver, »Generative AI and Lexicography: The Current State of the Art Using ChatGPT,« *International Journal of Lexicography* 36, št. 4 (2023): 355–87, pridobljeno 20. 5. 2025, https://doi.org/10.1093/ijl/ecad021.

8 Miloš Jakubíček in Michael Rundell, »The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography?,« v: *Electronic Lexicography in the 21st Century (eLex 2023): Proceedings of the eLex 2023 Conference*, ur. Marek Medveď et al. (Lexical Computing CZ, 2023), 522–23, pridobljeno 20. 5. 2025, https://elex.link/elex2023/wp-content/uploads/102.pdf.

9 Simon Krek, Cyprian Laskowski in Marko Robnik-Šikonja, »From Translation Equivalents to Synonyms: Creation of a Slovene Thesaurus Using Word Co-occurrence Network Analysis,« v: Iztok Kosem et al., ur., *Electronic Lexicography in the 21st Century* (Leiden: Dutch Language Institute, Lexical Computing CZ s.r.o., Trojina, 2017), 93–109, pridobljeno 20. 5. 2025, https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf.

čisti v sodelovanju med strokovnjaki za slovaropisje ter zainteresirano uporabniško javnostjo. Različico 1.0 predstavljajo Arhar Holdt et al.,[10] različico 2.0 pa Arhar Holdt et al.[11] in Gantar et al.[12]

Ideja pričujočega prispevka temelji na realnih potrebah nadaljnje slovarske gradnje. V prihodnje bi bilo v slovaropisne postopke mogoče vključiti dodatno strojno predprocesiranje podatkov s pomočjo programa ChatGPT. Ta bi podatke uredil na način, primerljiv slovaropisnemu, čemur bi sledil končni ročni pregled. Uspešna integracija strojne podpore bi lahko pomembno pohitrila nadgrajevanje slovarja, s tem pa pripravo odprto dostopnega sopomenskega gradiva, ki je dragoceno tudi za razvoj številnih nadaljnjih jezikovih virov in tehnologij za sodobno slovenščino. Da bi lahko izbrali ustrezno metodologijo tovrstne strojne podpore, je v prvem koraku treba ugotoviti, kakšne rezultate daje ChatGPT v primerjavi s slovaropisci za različne avtentične slovaropisne naloge.

Prispevek je razširjena različica konferenčnega prispevka, v katerem je bil predstavljen prvi zgoraj navedeni preizkus.[13] Za razširjeno različico smo dodali še drugi preizkus in članek ustrezno posodobili in nadgradili. V nadaljevanju zaporedno predstavimo metodologijo in rezultate obeh preizkusov, strnemo ugotovitve in napovemo nadaljnje delo na obravnavanem področju.

Prvi preizkus: selekcioniranje sopomenk in razvrščanje pod pomene

Metodologija

Preizkus temelji na delu podatkovnega vzorca za doktorsko raziskavo Sopomenskost v Slovarju sopomenk sodobne slovenščine in izbranih različicah Wordneta, tj. seznamu 546 samostalnikov, ki se kot iztočnice pojavijo v podatkovni bazi Slovarja sopomenk sodobne slovenščine 1.0[14] (SSSS 1.0) in drugih prosto dosto-

10 Špela Arhar Holdt et al., »Thesaurus of Modern Slovene: By the Community for the Community,« v: Jaka Čibej et al., ur., *Proceedings of the XVIII EURALEX International Congress, Lexicography in Global Contexts* (Ljubljana: Znanstvena založba Filozofske fakultete, 2018), 401–10, pridobljeno 20. 5. 2025, https://doi.org/10.4312/9789610600961.

11 Špela Arhar Holdt et al., »Thesaurus of Modern Slovene 2.0,« v: Marek Medveď et al., ur., *Electronic Lexicography in the 21st Century (eLex 2023)* (Brno: Lexical Computing CZ, 2023), 366–81, pridobljeno 20. 5. 2025, https://elex.link/elex2023/wp-content/uploads/82.pdf.

12 Polona Gantar et al., »Sopomenke 2.0 in Kolokacije 2.0: Novi koraki za slovenske odzivne slovarje,« *Jezik in slovstvo* 68, št. 4 (2023): 157–75, pridobljeno 20. 5. 2025, https://doi.org/10.4312/jis.68.4.157-175.

13 Magdalena Gapsa, Špela Arhar Holdt in Iztok Kosem, »Kako dober je ChatGPT pri umeščanju sopomenk pod besedne pomene,« v: Špela Arhar Holdt in Tomaž Erjavec, ur., *Jezikovne tehnologije in digitalna humanistika: Zbornik konference* (Ljubljana: Inštitut za novejšo zgodovino, 2024), 144–62, pridobljeno 20. 5. 2025, https://zenodo.org/records/13912515.

14 Simon Krek et al., *Thesaurus of Modern Slovene 1.0* (Repozitorij raziskovalne strukture CLARIN.SI, 2018), pridobljeno 20. 5. 2025, http://hdl.handle.net/11356/1166.

pnih leksikalnih virih, kot opisuje Gapsa.[15] Ta nabor je bil omejen na 266 iztočnic, ki so bile ob posodobitvi SSSS 1.0 v verzijo 2.0 slovaropisno urejene, kar pomeni, da imajo v verziji 2.0 pripisano pomensko členitev, strojno pridobljeni sopomenski kandidati iz verzije 1.0 pa so bili ročno pregledani, potrjeni (oziroma odstranjeni) in razvrščeni pod identificirane pomene.

Za izbranih 266 iztočnic je bilo v prvem koraku iz baze SSSS 1.0 izluščenih skupno 1049 sopomenskih kandidatov (z morebitnimi področnimi slovarskimi oznakami). V drugem koraku so bile iz Digitalne slovarske baze za slovenščino[16] izvožene pomenske členitve s pomenskimi indikatorji (tj. kratkimi opisi za ločevanje pomenov, kot pojasni Gantar[17]) za izbrane iztočnice. Podatki so bili pretvorjeni v tabelo, kjer je posamezna vrstica vsebovala izvožene podatke po vzoru: iztočnica – pomenska členitev – sopomenski kandidati. Tabela je služila kot nabor vhodnih podatkov za preizkus s sistemom ChatGPT. Za preverbo uspešnosti naloge smo iz baze Slovarja sopomenk sodobne slovenščine 2.0[18] (SSSS 2.0) pridobili slovaropisno pripravljene pomensko členjene iztočnice z razvrščenimi sopomenkami.

V prvem koraku analize je bilo med 266 iztočnicami prepoznanih 20 iztočnic, kjer se pomenska členitev iz DSBS ne ujema s SSSS 2.0 (npr. iztočnica *bonbon* ima v DSBS en pomen, v SSSS 2.0 sta dva). Ti primeri so posledica dejstva, da se DSB dinamično razvija s podatki iz različnih virov, in so bili za ohranitev koherentnega zlatega standarda odstranjeni iz nadaljnje analize.

Struktura poziva za ChatGPT

Za izbrane iztočnice smo pripravili poziv za ChatGPT (Priloga 1), pri čemer smo uporabili API model GPT-4. Poziv je bil pripravljen v angleščini in je bil med razvojem postopka večkrat testiran z uporabo brezplačne verzije sistema.

Med testiranjem se je izkazalo, da ChatGPT vrne boljše rezultate, če je v poziv vključen primer želenega rezultata. Posledično smo v poziv dodali primer vhodnih podatkov, tj. večpomensko iztočnico s sopomenskimi kandidati, in želene izhodne podatke, tj. pravilno razporejene sopomenske kandidate po pomenih.

Odgovori so bili vrnjeni v formatu YAML, sledila je pretvorba v format JSON. Na podlagi teh podatkov smo za raziskovalne analize in evalvacijo ustvarili še povzemalno datoteko CSV in Excelovo datoteko z vsemi zbranimi podatki.

15 Magdalena Gapsa, »But why?? Evaluation of User-Suggested Synonyms in the Thesaurus of Modern Slovene,« *Lang Resources & Evaluation* (2025), pridobljeno 20. 5. 2025, https://doi.org/10.1007/s10579-025-09821-8.

16 Iztok Kosem, Simon Krek in Polona Gantar, »Semantic Data Should No Longer Exist in Isolation: The Digital Dictionary Database of Slovenian,« v: Zoe Gavriilidou et al., ur., *EURALEX XIX: Congress of the European Association for Lexicography* (Democritus University of Thrace, 2021), 81–83, pridobljeno 20. 5. 2025, https://euralex.org/wp-content/uploads/2022/04/ABS2020.pdf.

17 Polona Gantar, *Leksikografski opis slovenščine v digitalnem okolju* (Ljubljana: Znanstvena založba Filozofske fakultete, 2015), pridobljeno 20. 5. 2025, https://doi.org/10.4312/9789612377922.

18 Simon Krek et al., *Thesaurus of Modern Slovene 2.0* (Repozitorij raziskovalne strukture CLARIN.SI, 2023), pridobljeno 20. 5. 2025, http://hdl.handle.net/11356/1916.

Slovaropisna ekipa je določala sopomenskost na podlagi korpusne analize možnosti zamenjave sopomenskih besed v sobesedilu. V poziv nismo vključili celotnih smernic, ki jim je sledila slovaropisna ekipa, saj bi s tem v postopek vnesli preveč informacij in spremenljivk, kar bi privedlo do neuporabnih in težje razložljivih rezultatov. Prav tako v poziv nismo vključili možnosti dodajanja ali spreminjanja besednih pomenov, ki jih je imela slovaropisna ekipa, saj smo želeli, da pomenska členitev ostane metodološko transparentna, rezultati pa dovolj enoznačni za analizo. Testiranja so pokazala optimalno delovanje poziva, ki je izvleček najpomembnejših navodil. Navodila, ki jih nismo vključili v poziv, navajamo ob analizi rezultatov, kadar olajšajo interpretacijo razlik med ročnim in strojnim delom.

Postopek analize gradiva

Pridobljeni podatki so bili organizirani v preglednice. Strojno pripravljene rezultate smo primerjali s slovaropisnimi rešitvami in najprej ugotovili, katere iztočnice so obravnavane povsem enako in katere vsebujejo razlike. Razlike smo nato natančneje analizirali v dveh korakih: (a) katere vrste odstopanja se pojavljajo pri odstranjevanju neustreznih sopomenskih kandidatov in kako pogosto in (b) katere vrste odstopanja se pojavljajo pri umeščanju neodstranjenega gradiva pod besedne pomene in kako pogosto.

V raziskavi rešitve slovaropisne ekipe obravnavamo kot zlati standard, kar pomeni, da odstope načeloma razumemo kot neželene. Vendar pa rezultati nakažejo, da je v določenih primerih rešitev, ki jo ponudi ChatGPT, drugačna od slovaropisne, vendar kljub temu sprejemljiva. Če bodo s ChatGPT pripravljeni podatki vključeni v slovaropisne delotoke, bo v prihodnje treba presoditi, kako v praksi obravnavati take primere skladno z izbranim slovaropisnim konceptom.

Splošna uspešnost

Pri analiziranih 246 iztočnicah je ChatGPT v 103 primerih (41,9 odstotka) podatke uredil povsem enako kot slovaropisci, v 143 primerih (58,1 odstotka) pa se je v odločitvi tako ali drugače razlikoval.

Podatke s primeri iztočnic prikazuje Tabela 1, v kateri podajamo tudi povprečno število kandidatov in slovarskih pomenov v posamezni skupini. V skupini ustrezno urejenih sopomenskih podatkov sta obe povprečji nižji, kar je skladno s pričakovanji, saj se s številom sopomenk za razvrstitev in številom besednih pomenov viša možnost za razlike v odločitvah. Povezava ni povsem enoznačna, saj se ChatGPT (lahko) razlikuje tudi pri iztočnicah z malo pomeni in sopomenkami ter uspešno uredi kompleksnejše iztočnice.

Tabela 1: Ujemanje med slovaropisci in ChatGPT s številom iztočnic, primeri in povprečnim številom sopomenskih kandidatov ter besednih pomenov na skupino

Vrsta rezultata	**Primeri**	**Št. iztočnic**	Povpr. št. kandidatov	Povpr. št. pomenov
Strojni rezultat enak ročnemu	adolescenca, aerodinamika, agonija, alkohol, ambicija, anatomija	103	2,2	1,7
Strojni rezultat drugačen od ročnega	adaptacija, anonimnost, aplikacija, arbiter, arhitektura, arhiv	143	5,1	2,4
Skupaj analiziranih		246	3,9 (vseh kandidatov: 951)	2,1 (vseh pomenov: 516)

Vir: lastno delo

Natančnejša analiza je pokazala, da se med 143 iztočnicami pojavlja 107 takih, ki kažejo razlike na ravni odstranjevanja neustreznih sopomenskih kandidatov (43,5 odstotka analiziranih iztočnic), 71 takih, ki kažejo razlike na ravni razvrščanja pod pomene (28,9 odstotka), od tega pa je 35 primerov (14,2 odstotka), kjer se pojavljajo tako razlike prvega kot drugega tipa.

Razlike v odstranjevanju neustreznih sopomenskih kandidatov

Prva naloga za ChatGPT je bila odstraniti sopomenske kandidate, ki ne sodijo pod nobenega od pomenov izbrane iztočnice. V zlatem standardu je bilo na ta način odstranjenih 249 (26,2 odstotka) od 951 kandidatov. ChatGPT je odstranil le 110 kandidatov (11,6 odstotka). Rezultati so prikazani v Tabeli 2, kjer so navedeni primeri, ki jih je ChatGTP glede na zlati standard ustrezno obdržal (true negatives, TN), ustrezno odstranil (true positives, TP), neustrezno obdržal (false negatives, FN) ali neustrezno odstranil (false positives, FP). V tabeli je najprej navedena iztočnica, nato pa sopomenski kandidat, o katerem je ChatGPT presojal.

Tabela 2: Primeri in število pravilnih in napačnih odločitev pri presojanju ChatGPT, ali je sopomenski kandidat ustrezen za dano iztočnico ter pomen ali ne

	Primeri	Vsota
Ustrezno obdržanih (TN)	adaptacija – preureditev, adolescenca – odraščanje, aerodinamika – aerodinamičnost, agonija – trpljenje, ambicija – želja po uspehu, anatomija – telesna zgradba	674
Ustrezno odstranjenih (TP)	arbiter – posrednik, argument – razlaga, avto – vagon, birokrat – velika živina, čajnik – kavnik, cedilo – posodica za kuhinjske odpadke	82
Neustrezno obdržanih (FN)	arbiter – gospodar, arhiv – arhivi, avtoriteta – premoč, dedek – babica, dražba – razpis del, električar – vzdrževalec telefonskega omrežja	167
Neustrezno odstranjenih (FP)	adaptacija – predelava, anonimnost – nepoznanost, aplikacija – prekritje, atentat – umor, bife – prehranjevalnica, cenzura – predelava [tiskarstvo]	28
Skupaj		951

Vir: lastno delo

Tabela 3: Natančnost (kolikšen delež odstranjenih primerov je dejansko neustreznih sopomenskih kandidatov), priklic (kolikšen delež vseh neustreznih kandidatov je bil identificiran) in F1 (harmonična sredina obeh vrednosti)

Natančnost	Priklic	F1
0,7455	0,3293	0,4568

Vir: lastno delo

Iz rezultatov je razvidno, da je ChatGPT pri presojanju relevantnosti sopomenskih kandidatov opazno popustljivejši od zlatega standarda, čeprav so uredniška načela SSSS že izhodiščno naravnana k širšemu razumevanju sopomenskosti in odločitvi za karseda široko vključevanje kandidatov.[19] Kot smo zapisali v razdelku Struktura poziva za ChatGPT, poziv za strojno obdelavo ni vseboval celotnih slovaropisnih smernic, po katerih velja, da se moške in ženske slovnične oblike ne obravnavajo kot neposredne sopomenke, ampak se uvrščajo pod spolsko ustrezne iztočnice (npr. *dedek – stari oče, babica – stara mama,* ne pa **dedek – babica*), da se množinske oblike ne upoštevajo kot sopomenke, razen če so za to v rabi utemeljeni razlogi (**arhiv – arhivi*), in da se opisne, definicijam podobne zveze obdržijo le, če se kot take pogosto pojavljajo v rabi (**dražba – razpis del*). Razlike v navodilih pojasnijo del razlik v rezultatih. Pri morebitni uporabi ChatGPT za pohitritev ročnega dela bi bila ta odstopanja predvidljiva, hitro opazna in enostavno rešljiva.

19 Gantar et al., »Sopomenke 2.0 in Kolokacije 2.0: Novi koraki za slovenske odzivne slovarje,« 161.

V naboru neustrezno prepoznanih so tudi mejni primeri, ki so bili zahtevni že za slovaropisno odločitev. Pri teh bi raba ChatGPT za pohitritev ročnega dela lahko doprinesla k lažjim, morda še širše vključujočim odločitvam. Na drugi strani so problematične neprepoznane sopomenske besede, kot denimo *atentat – umor*, *debelost – obilnost*, *kaos – razdejanje*. Pri takšnih primerih bi bila pri morebitni rabi postopka potrebna pozornost.

Napake v razvrščanju sopomenk

Pri analizi razvrščanja sopomenk pod pomene smo ločili dve vrsti razlik: (a) ChatGPT je sopomenko umestil pod neustrezen besedni pomen in (b) ChatGPT sopomenke ni umestil pod ustrezen pomen oziroma vse ustrezne pomene glede na zlati standard. Umestitev pod neustrezen pomen smo prepoznali pri 36 iztočnicah (14,6 odstotka analiziranih iztočnic), manjkajočo umestitev pri 49 iztočnicah (19,9 odstotka), od tega je 14 (5,7 odstotka) takih, kjer se pojavljata obe vrsti problema, tj. umestitev pod neustrezen pomen in manjkajoča umestitev. V Tabeli 4 so prikazani primeri, število razlik in iztočnic ter povprečno število kandidatov in slovarskih pomenov v posamezni od skupin. Pri primerih je najprej navedena iztočnica, sledi sopomenka, o kateri je ChatGPT presojal, in pomen, pod katerega jo je ali je ni umestil. Kot smo opozorili v razdelku Postopek analize gradiva, ustreznost oziroma neustreznost razumemo v razmerju do zlatega standarda, vendar se med rezultati pojavljajo tudi mejni primeri, kjer je lahko poleg slovaropisne odločitve sprejemljiva tudi odločitev ChatGPT.

Tabela 4: Primeri v iztočnicah, kjer je ChatGPT umestil sopomenko pod napačen pomen ali je ni umestil pod vse pomene. V stolpcih 3–6 je navedeno število napak, število iztočnic, povprečno število sopomenskih kandidatov in besednih pomenov za obe skupini.

Vrsta rezultata	Primeri	Št. napak	Št. iztočnic	Povpr. št. kandidatov	Povpr. št. pomenov
Umeščeno pod neustrezen pomen	bazar – sejem [ekonomija]: pod 'orientalska tržnica' namesto 'prireditev'; hazarder – igralec na srečo: pod 'kdor rad veliko tvega' namesto 'kdor rad stavi'	55	36	6,7	2,8
Neumeščeno pod pomen	bolnik – pacient: ustrezno pod 'kdor je bolan', manjka pri 'kdor je neprijeten ali krut [izraža negativen odnos]; gneča – množica: ustrezno pri 'o ljudeh', manjka pri 'o stvareh'	78	49	5,3	2,9
Skupaj		133	71	5,3	2,8

Vir: lastno delo

Podatki v Tabeli 4 kažejo, da se razlike pri razvrščanju pojavljajo pri iztočnicah, ki so v povprečju kompleksnejše glede števila sopomenk za razvrstitev ter števila besednih pomenov. Sklepati je mogoče, da na razlike vpliva tudi abstraktnost pomenskih indikatorjev, ki so človeku morda laže razumljivi, vendar se to ne kaže enoznačno: kot primer uspešno urejenega kompleksnega gesla lahko navedemo iztočnico *jagoda* s štirimi kratkimi in medsebojno podobnimi indikatorji (1. rastlina, 2. plod, 3. okrogel plod, 4. droben predmet), h katerim je sistem ustrezno razvrstil sopomenske kandidate *biser, koralda, kroglica.* Podobno je bilo tudi pri iztočnici *telefon* s šestimi indikatorji (1. naprava, 2. telekomunikacijska povezava, 3. napeljava, 4. o storitvi, 5. telefonska številka [neformalno], 6. družabna igra), kamor sta bila ustrezno razvrščena kandidata *telefonska slušalka* in *telefonski klic*.

Najtežji primeri

Med vsemi analiziranimi iztočnicami je šest takšnih, pri katerih so se pojavile težave tako na ravni odstranjevanja neustreznih kandidatov kot umeščanja sopomenskih kandidatov pod vse ustrezne pomene. Te obravnavamo kot težke primere in jih v nadaljevanju ponazarjamo v celoti, tako v slovaropisni različici kot po presoji ChatGPT. Razlike so prikazane z ležečim tiskom.

Tabela 5: Prikaz obravnave najtežjih primerov z opredelitvami razlik med odločitvami

Iztočnica	Slovaropisno	ChatGPT
adaptacija	1. prenova stavbe: prilagoditev, prenova, preureditev	1. prenova stavbe: preureditev, prenova
	2. priredba [v umetnosti]: *predelava [glasba]*, priredba [glasba], *preureditev [glasba]*, aranžma [glasba]	2. priredba [v umetnosti]: priredba [glasba], aranžma [glasba]
	3. prilagoditev [ponavadi v športu]: PRAZNO	3. prilagoditev [ponavadi v športu]: *prilagoditev*
	Odstranjeni kandidati: /	Odstranjeni kandidati: *predelava [glasba]*
čarovnica	1. pravljično bitje: čarodejka, coprnica	1. pravljično bitje: čarodejka, coprnica, *vračarica, vešča*
	2. ženska, obtožena čarovništva: *vračarica*	2. ženska, obtožena čarovništva: *vešča, hudičevka*
	3. zlobna, starejša ženska [izraža negativen odnos]: *vešča*, babura, *coprnica, hudičevka*, stara baba, zlobnica	3. zlobna, starejša ženska [izraža negativen odnos]: babura, stara baba, zlobnica
	4. ženska z nadnaravnimi sposobnostmi [lahko izraža negativen odnos]: *čarodejka, coprnica*	4. ženska z nadnaravnimi sposobnostmi [lahko izraža negativen odnos]: *iluzionistka*
	Odstranjeni kandidati: *iluzionistka*	Odstranjeni kandidati: /

Iztočnica	Slovaropisno	ChatGPT
gurman	1. ljubitelj hrane in pijače: sladokusec, uživač, gastronom, poznavalec hrane, epikurejec	1. ljubitelj hrane in pijače: sladokusec, uživač, *bonvivan*, poznavalec hrane, gastronom, epikurejec
	2. poznavalec: *sladokusec*	2. poznavalec: *bonvivan, poznavalec hrane, gastronom*
	Odstranjeni kandidati: *bonvivan*	Odstranjeni kandidati: /
kandidat	1. za službo ali funkcijo: prosilec, aspirant, prijavljenec, *interesent [ekonomija]*, prijavljeni	1. za službo ali funkcijo: prosilec, aspirant, prijavljenec, *potencialna kandidatka, možna kandidatka, predlaganec*, prijavljeni, *vložnik*
	2. na izpitu: *prijavljenec, prijavljeni, udeleženec*	2. na izpitu: *izpitna kandidatka, izprašanec*
	3. za nagrado: *predlaganec*, nominiranec	3. za nagrado: nominiranec, *volilna agitatorka, volilni agitator*
	Odstranjeni kandidati: *potencialna kandidatka, možna kandidatka, izpitna kandidatka, vložnik*, potencialni kupec [ekonomija], možni kupec [ekonomija], potencialna stranka [ekonomija], možna stranka [ekonomija], *izprašanec, volilna agitatorka, volilni agitator*, tekmovalec [šport], anketiranec, intervjuvanec	Odstranjeni kandidati: *interesent [ekonomija]*, potencialni kupec [ekonomija], možni kupec [ekonomija], potencialna stranka [ekonomija], možna stranka [ekonomija], udeleženec, tekmovalec [šport], anketiranec, intervjuvanec
ljubezen	1. močna naklonjenost: naklonjenost, nagnjenje, posebna naklonjenost, posebno nagnjenje, eros	1. močna naklonjenost: naklonjenost, *strast*, nagnjenje, posebna naklonjenost, *posebna nagnjenost*, posebno nagnjenje, *oboževanje, občudovanje, čustva*
	2. o intimnem odnosu: romanca, ljubezensko razmerje	2. o intimnem odnosu: ljubezensko razmerje, romanca
	3. ljubljena oseba: draga, ljuba, dragi, ljubi, fant, punca	3. ljubljena oseba: draga, ljubi, dragi, ljuba, fant, punca
	4. o dejavnosti: *naklonjenost, strast, nagnjenje, posebna naklonjenost, posebno nagnjenje*	4. o dejavnosti: PRAZNO
	5. seks: *strast*	5. seks: *eros*
	Odstranjeni kandidati: nežnost, toplina, *posebna nagnjenost, oboževanje, občudovanje, čustva*	Odstranjeni kandidati: nežnost, toplina

Iztočnica	Slovaropisno	ChatGPT
urok	1. v magiji: čarovnija, prekletstvo, zakletev, zaklinjanje, *zarotitev*, magične besede, coprnija, *zarekanje*, čarovniški izrek, čarobni napev	1. v magiji: čarovnija, prekletstvo, zakletev, zarotitev, magične besede, coprnija, zaklinjanje, *abrakadabra, vudu, čaranje, čarobna formula*, čarovniški izrek, čarobni napev, *čarobni simbol*
	2. o neuspehu [pogosto v športnem kontekstu]: *prekletstvo*	2. o neuspehu [pogosto v športnem kontekstu]: *nesreča, smola, zarekanje*
	Odstranjeni kandidati: *nesreča, smola, abrakadabra, vudu, čaranje, čarobna formula, čarobni simbol*, molitveni obrazec [religija]	Odstranjeni kandidati: molitveni obrazec [religija]

Vir: lastno delo

Primeri pregledno prikazujejo različne težave. Kot je razvidno, so določena razhajanja lahko posledica korpusnega gradiva, na osnovi katerega se oblikujejo pomenske členitve in potrjuje sopomenska raba (npr. *bonvivan,* ki se v rabi najbrž pojavlja preredko, da bi ga obdržali, ali *vešča* v pomenu 'pravljično bitje'). Prav tako so lahko mestoma zavajajoče ubeseditve v pomenskih indikatorjih, ki jih slovaropisna ekipa lahko interpretira na podlagi preostalih podatkov v DSBS, kot so na primer kolokacije, v nalogi za ChatGPT pa so bili predstavljeni brez dodatnega konteksta (denimo indikator poznavalec pri iztočnici *gurman*, ki je v opoziciji do 1. pomena /'ljubitelj hrane in pijače'/ in se v prenesenem pomenu ne navezuje več na hrano/pijačo, česar ChatGPT ne razbere). Nekaj je primerov, pri katerih slovaropisci upoštevajo smernice, ki ChatGPT niso bile podane v razdelku Struktura poziva za ChatGPT, na primer pri (ne)vključevanju moško-ženskih parov (*kandidat – izpitna kandidatka*). Najti pa je tudi razlike, kjer so odločitve ChatGPT težko razložljive, denimo *kandidat* v pomenu 'za nagrado' – *volilna agitatorka, volilni agitator*.

Drugi preizkus: izdelava novih pomenskih členitev

Ker številne iztočnice v Slovarju sopomenk sodobne slovenščine še nimajo izdelane pomenske členitve in pomenskih opisov, smo se odločili ChatGPT preizkusiti še pri nalogi pomenskega členjenja, ki je vključevala tudi oblikovanje definicij[20] in ne zgolj pomenskih indikatorjev.

20 V tem prispevku uporabljamo termin slovarska *definicija* (v pozivu *definition*) tudi za definicije cobuildskega tipa, čeprav bi zanje po obliki in vsebini ustrezal tudi termin *razlaga*. S tem sledimo predhodnim študijam, ki so bile zasnovane za podobne namene (gl. Razdelek 1). Sprememba poziva z navodilom za *explanation* bi v konkretni raziskavi uvedla novo spremenljivko in drugačne rezultate, zanimivo pa bi jo bilo preizkusiti pri nadaljnjem delu.

Metodologija

Za preizkus smo iz DSBS izbrali 116 ročno izdelanih iztočnic (63 samostalnikov, 32 pridevnikov in 21 glagolov), pri čemer je bil glavni pogoj, da so vključene tudi v Kolokacijski slovar sodobne slovenščine,[21] saj je to pomenilo, da so vsebovale zadostno količino kontekstualnih podatkov (kolokacij in zgledov). Čeprav smo zaradi poudarka na pomenski členitvi in razlikovanju med pomeni v nabor vključili predvsem večpomenske iztočnice (55 z dvema pomenoma, 34 s tremi, 12 s štirimi, po dve s petimi in šestimi ter eno s sedmimi pomeni), smo dodali tudi deset enopomenskih.

- Za vsako iztočnico smo iz baze izvozili ročno pregledane kolokacije in avtomatsko izluščene zglede (po en zgled na kolokacijo):
- Za vsak pomen smo izvozili do 20 kolokacij in zgledov.
- Pri izbiri kolokacij smo upoštevali podatek o jakosti logDice.
- Pri izbiri zgleda za kolokacijo smo izbrali tistega z najvišjo oceno kakovosti dobrega zgleda v orodju GDEX, ki ga opisujejo Kosem, Husak in McCarthy.[22]
- Glede na izsledke predhodnih kolokacijskih raziskav smo dali poudarek na izvozu kolokacij za pomensko bolj obvestilne skladenjske strukture:
  - Za samostalnike smo za strukture glagol + samostalnik v tožilniku, pridevnik + samostalnik in samostalnik + samostalnik v rodilniku izvozili po pet kolokacij in zgledov, preostanek smo zapolnili s kolokacijami in zgledi iz preostalih struktur.
  - Za pridevnike smo za strukturo pridevnik + samostalnik izvozili po deset kolokacij in zgledov, preostale iz drugih struktur.
  - Za glagole smo za strukturo glagol + samostalnik izvozili po sedem kolokacij in zgledov, za strukturo prislov + glagol po pet kolokacij in zgledov, preostale iz drugih struktur.

V primerih, ko prioritetne strukture niso vsebovale dovolj kolokacij, smo jih nadomestili s kolokacijami iz drugih struktur.

Druga informacija, ki smo jo pripravili, so bile slovarske definicije, ki smo jih pridobili iz dveh virov: semantičnega slovenskega leksikona Open Slovene Wordnet 1.0[23] in Angleško-slovenskega slovarja Bridge.[24] V obeh primerih smo pridobljene definicije še dodatno prilagodili oziroma izboljšali:

- V slovenskem Wordnetu so slovenske definicije zgolj avtomatski prevod angleških definicij in so v številnih primerih kratke in slabo obvestilne, na primer

21 Iztok Kosem et al., *Kolokacijski slovar sodobne slovenščine* (Ljubljana: Znanstvena založba Filozofske fakultete, 2018–), pridobljeno 20. 5. 2025, https://viri.cjvt.si/kolokacije/slv/#.

22 Iztok Kosem, Miloš Husak in Diana McCarthy, »GDEX for Slovene,« v: Iztok Kosem in Karmen Kosem, ur., *Electronic Lexicography in the 21st Century: New Applications for New Users* (Ljubljana: Trojina, Institute for Applied Slovene Studies, 2011), 150–59, pridobljeno 20. 5. 2025, http://www.trojina.si/elex2011/elex2011_proceedings.pdf.

23 Jaka Čibej et al., *Open Slovene WordNet OSWN 1.0* (Slovenian language resource repository CLARIN.SI, 2023), pridobljeno 20. 5. 2025, http://hdl.handle.net/11356/1888.

24 Angleško-slovenski slovar Bridge (Ljubljana: Državna založba Slovenije, 2000).

humanost → kakovost človeškosti; forma → določen način, na katerega se nekaj izrazi.
Pri obdelavi s ChatGPT-4 smo iskali daljše, celostavčne definicije. Primer pretvorbe za *prevajati*:
  - Izvorna definicija: *restate (words) from one language into another language.*
  - Avtomatski slovenski prevod: *ponovno izraziti (besede) iz enega jezika v drugem jeziku.*
  - Izboljšana definicija: *Prevajati pomeni izražati ali podajati pomen besedil ali izrazov iz enega jezika v drugega, tako da ohranjamo njihov pomen.*
- V Angleško-slovenskem slovarju Bridge so definicije na voljo v celostavčni obliki, vendar pa vsebujejo angleške iztočnice. V tem primeru smo do definicij za naš preizkus prišli po sledečem postopku:
  - Najprej smo angleške iztočnice avtomatsko zamenjali s slovenskimi prevodi, npr. *Kar je huge, je izjemno veliko po obsegu, količini ali stopnji. → Kar je velikanski, je izjemno po obsegu, količini ali stopnji.* in *Kadar nekaj browns ali is browned, postane temnejše barve. → Kadar nekaj porjaveti ali porjaveti, postane temnejše barve.*
  - Nato smo v analizi najprej izločili neproblematične definicije, pri preostalih pa smo prepoznali pet vzorcev težav, od takih, ki so zahtevale samo popravek sklona, do takih, kjer smo morali odpraviti podvajanje iztočnice v definiciji ali celo daljši del ubeseditve. Na podlagi tega smo za odpravo napak prilagodili sistemske pozive za ChatGPT-4, tako da smo dobili izboljšane definicije, na primer *Kar je velikansko, je izjemno veliko po obsegu, količini ali stopnji.* in *Kadar nekaj porjavi, postane temnejše barve.*

Pri združitvi definicij iz dveh virov smo opazili, da definicije niso deloma prekrivne zgolj med viroma, temveč tudi znotraj posameznega vira, kot kaže primer za glagol *degradirati:*

SLOVENSKI WORDNET

- ***Degradirati*** *pomeni opraviti dejanje, s katerim zmanjšamo stopnjo, rang ali vrednost nečesa, zaradi česar je to nekaj manj cenjeno ali spoštovano.*
- ***Degradirati*** *pomeni povzročiti zmanjšanje nivoja zemlje, na primer zaradi erozije.*
- ***Degradirati*** *pomeni uradno ali neformalno znižati nekoga ali nekaj v oceni, vrednosti ali ugledu, zaradi dejanskega dejanja, situacije ali presoje.*
- ***Degradirati*** *pomeni uradno prenesti nekoga na nižjo pozicijo ali mu uradno zmanjšati čin.*
- ***Degradirati*** *pomeni zmanjšati nečiji ali nečesa stopnjo ali rang, ali povzročiti, da se nekdo znajde v neprijetni ali nedostojni situaciji.*

SLOVAR BRIDGE

– *Če človek, ki ima oblast,* ***degradira*** *nekoga, mu podeli nižji položaj, pogosto kot znamenje kazni.*
– ***Degradirati*** *pomeni dati osebi ali stvari manj pomemben položaj ali veljavo.*

Hkrati je analiza pokazala, da tudi izboljšane definicije lahko vsebujejo slovnične ali druge napake, zato smo v poziv za ChatGPT dodali navodilo, naj podane definicije po potrebi združuje in izboljša.

Za deset iztočnic definicij nismo imeli na voljo, kar se je izkazalo za koristno, saj smo tako preverili tudi delovanje ChatGPT samo s podanimi kolokacijami in zgledi.

Struktura poziva za ChatGPT

Pri pripravi poziva smo najprej opravili obsežno testiranje, pri čemer smo prišli do podobnih ugotovitev kot pri razvrščanju sopomenk pod pomene, da bolje deluje poziv v angleščini in s primerom vhodnih in želenih podatkov. Pri tem poizkusu smo že uporabili novejši model GPT-4o.

Poziv je bil razdeljen na sistemsko navodilo, ki je bilo vedno enako, je bilo pa nekoliko drugačno za vsako besedno vrsto zaradi vključenih vzorcev definicij. Primer sistemskega navodila za pridevniške iztočnice predstavlja Priloga 2.

V glavni del poziva smo potem vključili definicije, kolokacije in zglede:

> Here are definitions, and collocations and their examples for the Slovenian word <*iztočnica*>. Collocations are numbered. Definitions come from various sources, and need to be improved, merged, and even omitted if they are referring to the same sense. Using all this data, create senses with definitions, distributing collocations and examples under senses. Provide only numbers of collocations, do not repeat the entire text of collocations and examples.
> DEFINITIONS:
> COLLOCATIONS AND EXAMPLES:

Odgovori so bili vrnjeni v formatu YAML, sledila je pretvorba v format JSON. Na podlagi teh podatkov smo za raziskovalne analize in evalvacijo ustvarili še povzemalno datoteko CSV in Excelovo datoteko z vsemi zbranimi podatki.

Analiza

Pri analizi smo preverili tri vidike podatkov, pridobljenih s ChatGPT: pokritost pomenov v Digitalni slovarski bazi za slovenščino, splošno ustreznost generiranih gesel in splošno ustreznost generiranih definicij.

Analizo pokritosti pomenov je opravil en slovaropisec, pri čemer je uporabil lestvico od 0 do 5 (0 – ni bil zaznan noben pomen, neuporabni podatki; 1 – zaznani redki pomeni; 2 – zaznana približno polovica pomenov; 3 – zaznana več kot polovica pomenov; 4 – zaznani skoraj vsi pomeni ali pa vsi pomeni, a nekateri le delno; 5 – zaznani vsi pomeni). Morebitne pomanjkljivosti, kot je pretirano drobljenje pomenov, prekrivnost definicij in podobno, niso bile upoštevane, zanimalo nas je samo, ali so bili vsi ročno zaznani pomeni tudi avtomatsko identificirani.

Splošno ustreznost generiranih gesel in ustreznost definicij sta ocenjevala dva slovaropisca. Pri splošni ustreznosti gesel so bila gesla ocenjena z uporabniškega vidika, torej smiselnosti, razumljivosti in dodelanosti. Pri ocenjevanju nismo upoštevali primerjav s pomensko členitvijo v DSBS, saj ChatGPT ni dobil podatkov o načelih in pravilih, ki jim pri izdelavi pomenov sledijo slovaropisci. Uporabljena je bila ocenjevalna lestvica od 0 do 5:

– 5 – V celoti je geslo zelo informativno, definicije dobre, pomenska členitev ustrezna; možne so manjše pomanjkljivosti, npr. napačno razporejen zgled, slovnična napaka v definiciji ipd.
– 4 – V celoti je geslo dobro izdelano, je pa pomanjkljivo v enem ali dveh elementih, npr. neustrezne definicije pri določenih pomenih, več zgledov napačno razvrščenih, preveč pomenov.
– 3 – Geslo je dokaj informativno, pomenska členitev deloma neustrezna, a posreduje relevantne informacije; pomeni se delno prekrivajo, določene definicije so lahko problematične za razumevanje.
– 2 – Posamezni pomeni so ustrezni in smiselni, pomenska členitev je pretežno neustrezna, definicije so prekrivne, možna je neustrezna razdelitev zgledov ali ubeseditve definicij ipd.
– 1 – Pomenska členitev je nejasna oz. nelogična, med pomeni je težko ali nemogoče razlikovati, razporeditev zgledov je neustrezna, ubeseditve definicij so pretežno neustrezne.
– 0 – Geslo je povsem neustrezno, npr. ne pojasnjuje besede v iztočnici, prevladuje tuj jezik, definicije so povsem neustrezne.

Tudi pri ocenjevanju ustreznosti definicij smo uporabili lestvico od 0 do 5, pri čemer nismo upoštevali morebitnih nezaznanih pomenov, napačno umeščenih zgledov ali soodvisnosti z indikatorji. Uporabljena lestvica:

– 5 – Definicije so dokaj dobro ubesedene in pomeni jasno razločeni (manjše napake toleriramo).
– 4 – Ubeseditve definicij so lahko problematične ali pa so pomensko prekrivne.
– 3 – Nekatere definicije so slabo ubesedene, prihaja tudi do pomenske prekrivnosti.
– 2 – Večina ali vse definicije so slabo ubesedene, nekateri ali vsi pomeni so prekrivni in slabo razlikovalni.
– 1 – Večina ali vse definicije so slabo ubesedene, niso razlagalne, pa tudi med njimi je slaba razlikovalnost.
– 0 – Definicije pojasnjujejo napačne pomene ali iztočnice.

Rezultati

Analiza pokritosti generiranih pomenov v Digitalni slovarski bazi za slovenščino (Tabela 6) je pokazala zmerno dobre rezultate, pri čemer je bila pri več kot 93 odstotkih iztočnic zaznana polovica ali več pomenov, od tega so bili pri 57 odstotkih iztočnic zaznani vsi pomeni. V primeru dveh iztočnic (*bakren, padalski*) z oceno 0 je šlo za jasno napako modela, ki je ponudil podatke za povsem napačno iztočnico.

Tabela 6: Pokritost pomenov v Digitalni slovarski bazi za slovenščino

Ocena	Število iztočnic	Odstotek
5 – zaznani vsi pomeni	66	57
4	21	18
3	21	18
2	5	4
1	1	1
0 – zaznan ni noben pomen	2	2

Vir: lastno delo

Rezultati ocenjevanja splošne ustreznosti gesel, generiranih s ChatGPT, ki so prikazani v Tabeli 7, kažejo precejšnjo uporabnost modela, saj je skoraj 80 odstotkov gesel dobilo povprečno oceno 3,5 ali več in 19 odstotkov najvišjo oceno obeh ocenjevalcev. Izračun strinjanja ocenjevalcev sicer kaže na relativno nizko ujemanje (Krippendorff's Alpha[25] (Ordinal Scale): 0.497, Bootstrap Confidence Interval (95% CI): [0.328, 0.631]), vendar pa je pri šeststopenjski lestvici in številnih elementih ocenjevanja to pričakovano.

Tabela 7: Splošna ustreznost gesel, generiranih s ChatGPT

Povprečna ocena	Število iztočnic	Odstotek
5	22	19
4,5	24	20,7
4	24	20,7
3,5	22	19
3	6	5,2
2,5	7	6
2	8	6,9
1	1	0,8
0	2	1,7

Vir: lastno delo

25 Gabriele Marzi, Marco Balzano in Davide Marchiori, »K-Alpha Calculator – Krippendorff's Alpha Calculator: A User-Friendly Tool for Computing Krippendorff's Alpha Inter-Rater Reliability Coefficient,« *MethodsX* 12 (2024), 102545, pridobljeno 20. 5. 2025, https://doi.org/10.1016/j.mex.2023.102545.

Primeri gesla s povprečno oceno 5:

bičati
- o fizičnem udarjanju z bičem; Kadar kdo biča nekoga, ga večkrat udari z bičem, šibo ali palico, da mu povzroči bolečino ali poškodbo.
- o kritiziranju ali obsodbi; Kadar kdo biča napake, grehe ali oblast, ostro kritizira in obsoja ti dve ali osebe.
- o udarjanju vremena ali naravnih elementov; Ko naravni elementi, kot sta veter in dež, bičajo nekaj, to pomeni, da silovito udarjajo nanj.

Primer gesla s povprečno oceno 3:

vešč
- Oseba ali način izvedbe, ki obvlada znanje ali spretnosti; Kdor je vešč česa, ima znanje in sposobnosti, ki mu omogočajo, da nekaj dela dobro.
- Oseba, ki je seznanjena s čim ali ima potrebna sredstva. Če si vešč nečesa, si s tem seznanjen ali imaš potrebna sredstva, znanje ali pooblastila za uspešno izvedbo določenega dejanja.

Primer gesla s povprečno oceno 2:

zatrjevati
- dati vtis ali trditi brez dejanske osnove; Ko kdo zatrjuje nekaj, daje vtis, da ima določene lastnosti ali trdi nekaj, kar je pogosto na neprepričljiv ali neresničen način.
- formalno predstaviti argumente ali trditve v pravnem kontekstu; Ko kdo v pravnem kontekstu zatrjuje nekaj, formalno predstavi svoje argumente ali trditve pred sodiščem ali v uradnih dokumentih.
- jasno trditi ali potrjevati nekaj; Ko kdo zatrjuje, potrjuje ali jasno izraža neko trditev ali prepričanje, pogosto z namenom, da prepriča druge, da je to resnica.
- potrditi brez dvoma; Ko nekdo zatrjuje nekaj, to pomeni, da to izjavi z gotovostjo in brez dvoma, pogosto kot dejstvo.
- zagotoviti ali izjaviti nasprotje nečesa drugega; Ko kdo zatrjuje nasprotno, trdi nekaj, kar je v nasprotju z obstoječimi informacijami ali prepričanji.

Uporabnost modela pri generiranju definicij potrjujejo tudi rezultati analize splošne ustreznosti generiranih definicij, ki je podana v Tabeli 8. Kot lahko vidimo, je skoraj 90 odstotkov gesel za ustreznost definicij dobilo povprečno oceno 3 ali več. Strinjanje ocenjevalcev je bilo tokrat še nekoliko nižje kot pri ocenjevanju splošne ustreznosti gesel (Krippendorff's Alpha (Ordinal Scale): 0.395, Bootstrap Confidence Interval (95 % CI): [0.226, 0.547]).

Tabela 8: Splošna ustreznost definicij, generiranih s ChatGPT

Povprečna ocena	Število iztočnic	Odstotek
5	12	10,4
4,5	28	24,1
4	26	22,4
3,5	19	16,4
3	18	15,5
2,5	5	4,3
2	3	2,6
1,5	3	2,6
0	2	1,7

Vir: lastno delo

Problemi generiranih podatkov

Pri analizi smo opazili nekatere ponavljajoče probleme v generiranih podatkih. Med najbolj opaznimi je težnja ChatGPT k pretiranemu drobljenju pomenov. Vendar pa statistični podatki kažejo nekoliko drugačno sliko, saj je samo pri 25 odstotkih iztočnic (29) ChatGPT zgeneriral več pomenov, kot jih najdemo v DSBS (pri tem je zanimiv podatek, da je med njimi kar osem od desetih iztočnic v DSB enopomenskih). Podrobnejša analiza je namreč pokazala, da so bili pri številnih iztočnicah, tudi ko je ChatGPT generiral enako število ali manj pomenov, kot jih je v DSB, generirani pomeni prekrivni ali skoraj identični (pri čemer so lahko nekateri drugi pomeni manjkali), kot denimo pri iztočnici *civilist*, kjer je ChatGPT zgeneriral sledeča pomena:

– človek, ki ni vojak; Civilist je človek, ki ni član oboroženih sil.
– oseba zunaj vojske; Civilist je oseba, ki ne pripada vojaškim silam in ne sodeluje v vojaških dejavnostih.

V drugih primerih je bila pomenska členitev prepodrobna, na primer definicije so pojasnjevale posamezno kolokacijo oziroma njihovo rabo. Dober primer tega je pridevnik *kamelji*, ki ima v DSB štiri pomene, ChatGPT pa jih je zgeneriral kar deset (Tabela 9).

Tabela 9: Pomenska členitev za *kamelji,* kot jo je generiral ChatGPT

Št. pomena	Generirani pomenski indikator	Generirana slovarska definicija
1	o barvi ali odtenku;	Barva ali odtenek, ki je kamelji, je podobna barvi kamelje dlake.
2	o dlaki ali volni;	Kar je kamelje, je narejeno iz dlake kamele ali jih kako drugače vključuje.
3	o festivalih ali dirkah;	Dogodek, ki je kamelji, vključuje kamelje dirke ali festivale, povezane s kamelami.
4	o glavi ali kosteh;	Predmeti, ki so kamelji, so narejeni iz delov telesa kamele, kot so glave ali kosti.
5	o hrani in pijači;	Kadar je hrana ali pijača kamelja, je narejena iz kameljih sestavin ali se nanaša na kamele.
6	o hrbtih;	Kar je kamelje, se nanaša na hrbte kamel in njihov specifičen videz.
7	o mladih ali drugih lastnikih;	Kar je kamelje, se nanaša na mlade kamele ali ljudi, ki skrbijo za kamele.
8	o prevozu in potovanjih;	Prevoz, ki je kamelji, vključuje uporabo kamel kot prevozno sredstvo ali se nanaša na relacije povezane z njimi.
9	o trgovini ali sejmu;	Kadar je nekaj kamelje, je povezano z dejavnostmi, povezanimi s prodajo kamel.
10	o usnju ali koži;	Kar je kamelje, je narejeno iz usnja ali kože kamele.

Vir: lastno delo

Druga pogosta težava so bili neprepoznani pomeni, med katerimi gre pogosto za preneseno rabo. Na primer pri *deževati* je ChatGPT zaznal pomen 'vremenski pojav' in pomen 'padanje predmetov', ne pa tudi pomena 'nenadna pojavitev velike količine česa' (na primer *Pritožbe in grožnje zdaj* ***dežujejo*** *z vseh strani.*).

Kot pogosta težava se je izkazala tudi umestitev zgledov pod pomene, kar je bilo v številnih primerih posledica (hkratne) neustreznosti pomenske členitve oziroma prekrivnosti generiranih pomenov. Tu smo prepoznali tako umestitev zgledov pod napačne pomene kot tudi podvajanje pri umeščanju, tj. umestitev istega zgleda pod več kot en pomen.

Čeprav je bila ubeseditev definicij glede na navodila[26] celostno gledano precej ustrezna, pa smo vseeno zaznali kar nekaj primerov problematične ubeseditve. Po eni

26 Gantar, *Leksikografski opis slovenščine v digitalnem okolju.* Ustreznost definicij smo ocenjevali skladno s slovarskimi navodili, ki smo jih oblikovali pri izdelavi LBS, in sicer smo definicije opredelili ločeno za posamezno besedno vrsto, pri čemer smo za glagolske pomene preferirali navedbo stavčne definicije, ki naj vključuje vse ključne skladenjsko-pomenske elemente posameznega pomena, tj. udeležence in okoliščine kot tudi konotativne in pragmatične pomenske elemente.

strani take definicije niso sledile vzorcem iz sistemskih navodil, bile so tudi predolge ali zelo kratke. Nekatere so vsebovale slovnične ali skladenjske napake. Naleteli smo tudi na nekaj primerov neustreznih oziroma napačnih definicij, na primer za pomen samostalnika *kajak*: *Ko govorimo o izposoji ali najemu* ***kajakov****, mislimo na možnost, da plovilo najdemo za določeno obdobje proti plačilu.*

Med redkejšimi težavami smo zaznali uporabo angleških indikatorjev in generiranje povsem napačne definicije (za neko drugo iztočnico).

Problemi vhodnih podatkov

Analiza rezultatov je pokazala tudi nekatere pomanjkljivosti vhodnih podatkov, ki so lahko privedli do slabših rezultatov pri pomenskem členjenju (oblikovanju definicij, razporeditvi kolokacij in zgledov ipd.). Predvsem gre tu za pomensko ustreznost in kakovost avtomatsko pridobljenih zgledov. Na primer, pri večpomenskih kolokacijah, ki se potrjujejo z zgledi, se lahko zgodi, da jih večina potrjuje samo določen pomen. Za drugi pomen, ki mu pripada enaka kolokacija, pa zgledov ni ali pa jih je malo. Primer tega je samostalnik *agonija* in večpomenska raba številnih kolokacij (*huda agonija, mučna agonija, podaljšati agonijo* ipd.), pri čemer smo za kar 33 zgledov kolokacij pri pomenu 'umiranja' ugotovili, da spadajo v pomen 'težavnega obdobja' (na primer *Obljube se niso izpolnile, letališče pa je zapadlo v še hujšo agonijo.*). To je pomenilo, da je imel model za pomen 'umiranja' na voljo zelo malo zgledov. Povezana težava je slaba kakovost nekaterih zgledov, predvsem smo zaznali težavo pomanjkljivega konteksta ali referenta, na katerega se definicija nanaša. Nekaj primerov:

– 1. Sodeč po fotografijah, ki jih prejemamo v uredništvu, so že lepo **košati**.
– 2. Nekatere so čisto **benigne**, nekatere pa ogrožajo celo človeška življenja.
– 3. Označujeta jo izviren izraz in bogata **barvitost**.
– 4. So tudi raj za prave **gurmane**.

Ena od pomanjkljivosti, ki je vplivala na število zaznanih pomenov pri sicer le nekaj iztočnicah, je bila zastopanost pomenov v vhodnih podatkih. V nekaterih primerih namreč pomen v DSBS še ni imel pripisanih kolokacij (mogoče so bili v bazi samo zgledi), zato ga tudi ni bilo mogoče vključiti v vhodne podatke za ChatGPT.

Glede definicij, ki smo jih kot vhodne podatke vzeli iz slovenskega Wordneta in slovarja Bridge, lahko rečemo, da so bile ne glede na morebitno prekrivnost in deloma slabšo kakovost pogosto v pomoč, saj jih je v številnih primerih ChatGPT uporabil dobesedno. Pri iztočnicah, za katere nismo imeli na voljo definicij, nismo opazili izstopajočih značilnosti, saj so bila generirana gesla različnih ocen splošne ustreznosti, je pa mogoče pomenljiv podatek, da sta bila med njimi obe problematični iztočnici s popolnoma napačnimi podatki (*bakren* in *padalski*).

Sklep in nadaljnje delo

V raziskavi smo preverili, kako uspešen je ChatGPT pri umeščanju sopomenskega gradiva pod besedne pomene in pri generiranju slovarskih gesel. Analizirali smo rezultate razvrščanja 951 sopomenskih kandidatov za 246 slovarskih iztočnic ter kakovost generiranih pomenskih členitev in definicij za 116 iztočnic.

Pri prvem poizkusu je strojni postopek v 41,9 odstotka primerov vrnil rezultate, povsem skladne s slovaropisnimi. Pri drugih iztočnicah, ki so v povprečju kompleksnejše (prinašajo več sopomenskih kandidatov za razvrstitev in več slovarskih pomenov), se pojavljajo odstopanja različnih vrst. Ob odstranjevanju neustreznih sopomenskih kandidatov se sistem razlikuje v 43,5 odstotka analiziranih iztočnic. Večina odstopanj je posledica popustljivosti sistema do sopomenskih kandidatov, ki jih je slovaropisna ekipa odstranila. Ker koncept SSSS načelno teži k širokemu vključevanju gradiva, slovarski vmesnik pa omogoča odziv uporabniške skupnosti na neustrezne kandidate, so ti odstopi manj problematični. V 28,9 odstotka analiziranih iztočnic se pojavijo napačne razporeditve sopomenk pod pomene ali neumestitve sopomenk pod vse ustrezajoče pomene. Ti odstopi so pogostejši pri kompleksnejših geslih, predvidevamo pa, da so vsaj delno (lahko) posledica kratkosti oziroma specifične vloge indikatorjev znotraj DSBS, pa tudi specifik korpusnega gradiva, ki v slovaropisnih delotokih pogojuje pomensko členjenje in preverbo sopomenskosti. Natančnejši pregled primerov, v katerih se pojavljajo različna odstopanja, pokaže, da se ChatGPT tudi pri najtežjih primerih ne razlikuje radikalno od slovaropisne presoje, razlike pa so lahko za slovaropisno delo tudi uporabne, saj omogočajo dodatne razmisleke, zlasti pri mejnih primerih. Skleniti je mogoče, da postopek deluje dokaj dobro in ima uporabno vrednost za pohitritev ročnega slovaropisnega dela.

V drugem preizkusu smo testirali zmožnost ChatGPT za samostojno izdelavo slovarskih gesel. Analiza kakovosti generiranih gesel kaže, da je sistem zaznal vse pomene v 57 odstotkih primerov, skoraj 80 odstotkov generiranih gesel je doseglo povprečno oceno 3,5 ali več, 19 odstotkov pa najvišjo oceno obeh ocenjevalcev. Pri generiranju pomenov se je kot težava izkazala pretirana granularnost, zlasti kot posledica ponovljenih ali pretirano podrobnih pomenov. Med problematične vidike spadajo tudi neprepoznane prenesene rabe ter težave pri razvrščanju zgledov pod ustrezne pomene. Pri ocenjevanju definicij smo ugotovili, da so bile nekatere neustrezno oblikovane ali premalo informativne, druge so vsebovale slovnične napake, v redkih primerih pa so bile generirane definicije povsem napačne. Kljub temu so v večini primerov definicije sledile pričakovanim smernicam in so bile ocenjene kot uporabne. Podatki kažejo, da so bili rezultati zanesljivejši, kadar so bili vhodni podatki bogatejši, še posebej v primerih, kjer so bile na voljo kakovostne kolokacije in zgledi.

Eden izmed ključnih izzivov obeh preizkusov je nepredvidljivost postopka. ChatGPT kot generativni model ne deluje po strogo določenih pravilih strojnega procesiranja podatkov, kar pomeni, da rezultati niso nujno ponovljivi ali povsem razložljivi. Ta značilnost pomembno omejuje domet evalvacijskih raziskav, kot je

naša, ne more pa biti razlog, da generativnih tehnologij v slovaropisju ne bi uporabljali in/ali ocenjevali.

Prvi korak za nadaljnje delo je s preizkušeno metodologijo pripraviti nove rezultate in testirati, ali delo s strojno predpripravo slovaropisne odločitve dejansko pohitri ali ne. Ker je strojni postopek, ki ga preizkušamo, odvisen od izbranega poziva, vhodnih podatkov in različice uporabljenega sistema, je raziskavo mogoče ponoviti na zmogljivejših različicah ChatGPT ali drugih podobnih sistemih, z nadgrajenimi pozivi in na novem gradivu (na primer za razvrščanje uporabniško dodanih sopomenk ali protipomenk). Jezikoslovno preglednejše in jasnejše rezultate bi lahko dobili, če bi se omejili na homogeno gradivo, denimo celoten razred vrstnih pridevnikov tipa *kamelji* (tudi *slonji, krokodilji* itd.) ali glagolov s primerljivimi vezljivostnimi značilnostmi. Preizkusiti pa je mogoče tudi druge naloge v podporo slovaropisnemu delu, tako za urejanje gradiva posameznega slovarja kot povezovanje leksikalnih podatkov iz različnih virov. Z ustreznimi metodološkimi premisleki je mogoče preveriti in vključiti tudi ustvarjalne generativne naloge, kot je denimo predlaganje novih sopomenk in protipomenk za podane iztočnice. V širšem smislu bi bilo zanimivo raziskavam, ki preverjajo razumevanje koncepta sopomenskosti med različnimi uporabniškimi skupinami slovarja,[27] dodati še »razumevanje« pri rabi ChatGPT oziroma umetne inteligence.

Zahvala

Raziskovalna programa št. P6-0411 (*Jezikovni viri in tehnologije za slovenski jezik*) in št. P6-0215 (*Slovenski jezik – bazične, kontrastivne in aplikativne raziskave*) ter raziskovalni projekt *Veliki jezikovni modeli za digitalno humanistiko* (GC-0002) sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna.

Viri in literatura

Literatura

Angleško-slovenski slovar Bridge. 2000. Ljubljana: Državna založba Slovenije.

Arhar Holdt, Špela, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Bojan Klemenc, Iztok Kosem, Simon Krek, Cyprian Laskowski in Marko Robnik-Šikonja. »Thesaurus of Modern Slovene: By the Community for the Community.« V: *Proceedings of the XVIII EURALEX International Congress, Lexicography in Global Contexts, 17–21 July 2018, Ljubljana*, uredili Jaka Čibej, Vojko Gorjanc, Iztok Kosem in Simon Krek, 401–10. Ljubljana: Znanstvena založba Filozofske fakultete, 2018. Pridobljeno 20. 5. 2025. https://doi.org/10.4312/9789610600961.

27 Gapsa, »But why?? Evaluation of User-Suggested Synonyms in the Thesaurus of Modern Slovene.«

Arhar Holdt, Špela, Polona Gantar, Iztok Kosem, Eva Pori, Marko Robnik Šikonja in Simon Krek. »Thesaurus of Modern Slovene 2.0.« V: *Electronic Lexicography in the 21st Century (eLex 2023), Proceedings of the eLex 2023 Conference, 27–29 June 2023*, uredili Marek Medveď, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas, Miloš Jakubíček in Simon Krek, 366–81. Brno: Lexical Computing CZ, 2023. Pridobljeno 20. 5. 2025. https://elex.link/elex2023/wp-content/uploads/82.pdf.

de Schryver, Gilles-Maurice. »Generative AI and Lexicography: The Current State of the Art Using ChatGPT.« *International Journal of Lexicography* 36, št. 4 (2023): 355–87. Pridobljeno 20. 5. 2025. https://doi.org/10.1093/ijl/ecad021.

Gantar, Polona. *Leksikografski opis slovenščine v digitalnem okolju.* 1. izd., elektronska izd. Ljubljana: Znanstvena založba Filozofske fakultete, 2015. Zbirka Sporazumevanje. Pridobljeno 20. 5. 2025. https://doi.org/10.4312/9789612377922.

Gantar, Polona, Špela Arhar Holdt, Iztok Kosem in Simon Krek. »Sopomenke 2.0 in Kolokacije 2.0: Novi koraki za slovenske odzivne slovarje.« *Jezik in slovstvo* 68, št. 4 (2023): 157–75. Pridobljeno 20. 5. 2025. https://doi.org/10.4312/jis.68.4.157-175.

Gapsa, Magdalena, Špela Arhar Holdt in Iztok Kosem. »Kako dober je ChatGPT pri umeščanju sopomenk pod besedne pomene.« V: *Jezikovne tehnologije in digitalna humanistika: Zbornik konference, 19.–20. september 2024, Ljubljana, Slovenija*, uredila Špela Arhar Holdt in Tomaž Erjavec, 144–62. Ljubljana: Inštitut za novejšo zgodovino, 2024. Pridobljeno 20. 5. 2025. https://zenodo.org/records/13912515.

Gapsa, Magdalena. »But why?? Evaluation of User-Suggested Synonyms in the Thesaurus of Modern Slovene.« *Lang Resources & Evaluation* (2025). Pridobljeno 20. 5. 2025. https://doi.org/10.1007/s10579-025-09821-8.

Jakubíček, Miloš in Michael Rundell. »The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography?« V: *Electronic Lexicography in the 21st Century (eLex 2023): Proceedings of the eLex 2023 Conference*, uredili Marek Medveď, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas, Miloš Jakubíček in Simon Krek, 518–33. Lexical Computing CZ, 2023. Pridobljeno 20. 5. 2025. https://elex.link/elex2023/wp-content/uploads/102.pdf.

Kosem, Iztok, Simon Krek in Polona Gantar. »Semantic Data Should No Longer Exist in Isolation: The Digital Dictionary Database of Slovenian.« V: *EURALEX XIX: Congress of the European Association for Lexicography, Lexicography for Inclusion, 7–9 September 2021, Virtual, Book of Abstracts*, uredili Zoe Gavriilidou, Lydia Mitits in Spyros Kiosses, 81–83. Democritus University of Thrace, 2021. Pridobljeno 20. 5. 2025. https://euralex.org/wp-content/uploads/2022/04/ABS2020.pdf.

Kosem, Iztok, Husak, Miloš in McCarthy, Diana. »GDEX for Slovene.« V: *Electronic Lexicography in the 21st Century: New Applications for New Users: Proceedings of eLex 2011, 10–12 November 2011, Bled, Slovenia,* uredila Iztok Kosem in Karmen Kosem, 150–159. Ljubljana: Trojina, Institute for Applied Slovene Studies, 2011. Pridobljeno 20. 5. 2025. http://www.trojina.si/elex2011/elex2011_proceedings.pdf.

Krek, Simon, Cyprian Laskowski in Marko Robnik-Šikonja. »From Translation Equivalents to Synonyms: Creation of a Slovene Thesaurus Using Word Co-occurrence Network Analysis.« V: *Electronic Lexicography in the 21st Century. Proceedings of eLex 2017 Conference: Lexicography from Scratch,* uredili Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek in Vít Baisa, 93–109. Leiden: Dutch Language Institute, Lexical Computing CZ s.r.o., Trojina, 2017. Pridobljeno 20. 5. 2025. https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf.

McKean, Erin in Will Fitzgerald. »The ROI of AI in Lexicography.« *Lexicography* 11, št. 1 (2024): 7–27. Pridobljeno 20. 5. 2025. https://utppublishing.com/doi/abs/10.1558/lexi.27569.

Lew, Robert. »ChatGPT as a COBUILD Lexicographer.« *Humanities and Social Sciences Communications* 10 (2023), Article 704. Pridobljeno 20. 5. 2025. https://doi.org/10.1057/s41599-023-02119-6.

Marzi, Gabriele, Marco Balzano in Davide Marchiori. »K-Alpha Calculator—Krippendorff's Alpha Calculator: A User-Friendly Tool for Computing Krippendorff's Alpha Inter-Rater Reliability Coefficient.« *MethodsX* 12 (2024), 102545. Pridobljeno 20. 5. 2025. https://doi.org/10.1016/j.mex.2023.102545.

Ptasznik, Bartosz, Sascha Wolfer in Robert Lew. »A Learners' Dictionary versus ChatGPT in Receptive and Productive Lexical Tasks.« *International Journal of Lexicography* 37, št. 3 (2024): 322–36. Pridobljeno 20. 5. 2025. https://doi.org/10.1093/ijl/ecae011.

Rundell, Michael. »Automating the Creation of Dictionaries: Are We Nearly There?.« V: *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users,"* 22–24 June 2023, Seoul, South Korea, 9–17. Yonsei University, 2023. Pridobljeno 20. 5. 2025. https://www.asialex.org/pdf/Asialex-Proceedings-2023.pdf.

Tiberius, Carole, Kris Heylen, Jesse de Does, Bram Vanroy, Vincent Vandeghinste in Job van Doeselaar. »LLMs and Evidence-based Lexicography.« V: *Large Language Models and Lexicography, Book of Abstracts, 8th October 2024, Cavtat, Croatia*, uredil Simon Krek, 44–48. 2024. Pridobljeno 25. 1. 2025. https://www.cjvt.si/wp-content/uploads/2024/10/LLM-Lex_2024_Book-of-Abstracts.pdf.

Spletni viri

Čibej, Jaka, Luka Terčon, Simon Krek, Andraž Repar, Erik Novak, Polona Gantar, Iztok Kosem, Špela Arhar Holdt, Kaja Dobrovoljc, Amadea Berginc, Irena Hvala, Damijan Klement, Manja Kolenc, Ana Močnik, Tina Munda, David Pavlas, Anamari Pečan, Aleksandra Poljak, Davorin Sečnik, Jure Šešet, Jan Štumberger, Tina Toličič in Laura Trpin. *Open Slovene WordNet OSWN 1.0.* Slovenian language resource repository CLARIN.SI, 2023. Pridobljeno 20. 5. 2025. http://hdl.handle.net/11356/1888.

Kosem, Iztok, Špela Arhar Holdt, Simon Krek, Polona Gantar, Eva Pori, Urška Kamenšek, Primož Ponikvar, Rebeka Roblek, Jure Šešet, Petra Zaranšek, Karolina Zgaga, Jaka Čibej, Bojan Klemenc, Cyprian Laskowski, Kaja Dobrovoljc, Vojko Gorjanc in Nikola Ljubešić. *Kolokacijski slovar sodobne slovenščine.* Ljubljana: Znanstvena založba Filozofske fakultete, 2018–. Pridobljeno 20. 5. 2025. https://viri.cjvt.si/kolokacije/slv/#.

Krek, Simon, Cyprian Laskowski, Marko Robnik-Šikonja, Iztok Kosem, Špela Arhar Holdt, Polona Gantar, Jaka Čibej, Vojko Gorjanc, Bojan Klemenc in Kaja Dobrovoljc. *Thesaurus of Modern Slovene 1.0.* Repozitorij raziskovalne strukture CLARIN.SI, 2018. Pridobljeno 20. 5. 2025. http://hdl.handle.net/11356/1166.

Krek, Simon, Cyprian Laskowski, Marko Robnik-Šikonja, Iztok Kosem, Špela Arhar Holdt, Polona Gantar, Jaka Čibej, Vojko Gorjanc, Bojan Klemenc, Kaja Dobrovoljc, Eva Pori, Rok Roblek in Klemen Zgaga. *Thesaurus of Modern Slovene 2.0.* Repozitorij raziskovalne strukture CLARIN.SI, 2023. Pridobljeno 20. 5. 2025. http://hdl.handle.net/11356/1916.

OpenAI. »ChatGPT (veliki jezikovni model).« Pridobljeno 31. 5. 2024. https://chatgpt.com.

Špela Arhar Holdt, Magdalena Gapsa, Polona Gantar, Iztok Kosem

THE POTENTIAL OF CHATGPT IN THE DEVELOPMENT OF THE THESAURUS OF MODERN SLOVENE

SUMMARY

This study examines the potential of ChatGPT-4 to support lexicographic work by evaluating its performance in two tasks: filtering and assigning synonym candidates to their corresponding lexical senses, and generating complete dictionary entries, including sense distinctions, definitions, and usage examples. The evaluation is based on a comparison with expert lexicographic decisions recorded in the Digital Dictionary Database for Slovene. The goal is to determine how closely ChatGPT's outputs align with established lexicographic practices and to explore whether the model can reliably contribute to streamlining dictionary compilation. By assessing the accuracy and utility of the generated content, the research aims to clarify the practical role large language models might play in digital lexicography.

In the first experiment, ChatGPT processed 951 synonym candidates across 246 dictionary entries. The model's decisions fully matched those of the lexicographers in 41.9 % of the cases, while in 58.1 % of the cases, it made different choices. A key finding was that ChatGPT was more permissive in retaining synonym candidates that experts had excluded. In 14.6 % of the entries, synonyms were assigned to different senses than in the gold standard, and in 19.9 %, expected synonym placements were missing. These differences often stemmed from the complexity of the entries and the brevity or ambiguity of semantic indicators. Despite these issues, the system's performance suggests that it could serve as a valuable tool for the preliminary classification of synonyms, supporting rather than replacing human judgment.

The second experiment assessed ChatGPT's ability to generate complete dictionary entries for 116 headwords without human input. The model correctly identified all lexical senses in 57 % of cases. Nearly 80 % of the entries received an average quality rating of 3.5 or above, while 19 % were given the highest score by both evaluators. However, several challenges were noted, including excessive granularity in sense division, a tendency to overlook figurative meanings, and occasional mismatches between definitions and examples. Some definitions lacked precision or included minor grammatical errors, though most adhered to conventional lexicographic norms. The quality of the outputs was notably higher when the input data included clear collocations and illustrative examples, confirming the importance of structured input for effective generative processing.

A central challenge across both tasks is the unpredictability inherent in generative models such as ChatGPT. Because the model's outputs are not deterministic, results are not always repeatable or easily interpretable, complicating evaluation and

integration into structured editorial workflows. Nevertheless, the findings demonstrate that with proper monitoring and refinement, ChatGPT has real potential to accelerate routine lexicographic tasks. Future work should explore more advanced model versions, improved prompt engineering, and broader applications such as sorting user-submitted content or generating lexical suggestions. With appropriate methodology, ChatGPT could become a valuable tool in lexicography, complementing expert work with increased speed and additional insights.

PRILOGA 1: Poziv za selekcioniranje sopomenk in razvrščanje pod pomene

You are a lexicographer preparing a comprehensive language resource. You work in the Slovenian language. You should respond in the Slovenian language and only provide output that is relevant and valid for Slovenian.

You are given a word with its various meanings and its synonyms. Assign given synonyms to suitable meaning from one of the provided ones.

Not all meanings have synonyms. One synonym can suit multiple meanings, you can assign it to more than one meaning. You can discard synonyms that are not suitable for any of the meanings.

Each prompt represents a word and its synonyms and should be treated as unit. You will provide a response for each unit.

Unit "argument" looks like this:

word: argument

meanings:

1. utemeljen razlog
2. neodvisna spremenljivka

synonyms: dokazni razlog, neodvisna spremenljivka [matematika], razlaga, utemeljitev, razmišljanje, smerni kot, udeležena vloga [jezikoslovje]

Your response should look like this in a valid YAML format:

```
word: argument
meanings:
- id: 1
   meaning: utemeljen razlog
   synonyms:
   - dokazni razlog
   - utemeljitev
   - razlog
- id: 2
   meaning: neodvisna spremenljivka
   synonyms:
     - neodvisna spremenljivka [matematika]
   discarded_synonyms:
```

- razlaga
- razmišljanje
- smerni kot
- udeleženska vloga [jezikoslovje]

Only answer in YAML format. Write nothing else.

Izbrani primer za poziv je samostalnik *adaptacija*, ki prinaša več pomenov, slovarske oznake in primer odstranjenih (ang. *discarded*) sopomenskih kandidatov.

Here is unit "adaptacija":
word: adaptacija
meanings:
1. prenova stavbe
2. priredba [v umetnosti]
3. prilagoditev [ponavadi v športu]

synonyms: predelava [glasba], priredba [glasba], preureditev, prilagoditev, prenova, aranžma [glasba]

```
word: adaptacija
meanings:
  - id: 1
    meaning: prenova stavbe
    synonyms:
    - preureditev
    - prenova
  - id: 2
    meaning: priredba [v umetnosti]
    synonyms:
    - priredba [glasba]
    - aranžma [glasba]
  - id: 3
    meaning: prilagoditev [ponavadi v športu]
    synonyms:
    - prilagoditev
  discarded_synonyms:
  - predelava [glasba]
```

PRILOGA 2: Poziv za generiranje slovarskega gesla

You are a lexicographer. The definitions should be written in Slovene, in full-sentence form, as in Cobuild-style.

Formats used can be \"Kdor je <adjective>, je...\" or \"Kar je <adjective>, je...\" or \"Kadar je kdo <adjective>, je...\" but other formats can be used. Examples of good definitions:

Žival, ki je <HH>amfibijska</HH>, lahko živi tako na kopnem kot v vodi.
Kar je <HH>krvavo</HH>, je prekrito s krvjo.
Kadar je človek <HH>zaskrbljen</HH>, je zaradi nečesa živčen ali v skrbeh.
Z besedo <HH>oren</HH> opisujemo tisto, kar je povezano s pridelovanjem poljščin.
Kdor je <HH>strasten</HH> do nečesa, je za to zelo navdušen ali vnet.
Vozilo, ki je <HH>blindirano</HH>, ima trd oklep, ki potnike varuje pred morebitnimi streli in izstrelki.
Podjetje, ki je <HH>multinacionalno</HH>, ima podružnice v številnih državah.
Za človeka rečemo, da je <HH>odbijajoč</HH>, kadar se nam zdi neprijeten in ga ne želimo bolje spoznati.
Kar je <HH>oljnato</HH>, je prekrito z oljem ali ga vsebuje.

The output should follow this format:

1. short indicator
Full sentence-definition
Numbers of collocations + examples.

Example for the adjective \"prostaški\":

1. o komunikaciji
Govorica, ki je prostaška, vsebuje kletvice ali neotesane besede.
(5), (8), (14)

1.01 DOI: https://doi.org/10.51663/pnz.65.3.09

Ivana Filipović Petrović,* Slobodan Beliga**

Can AI Understand Croatian Idioms? Assessing Large Language Models in Lexicographic Tasks

IZVLEČEK

ALI LAHKO UMETNA INTELIGENCA RAZUME HRVAŠKE IDOME? OCENA VELIKIH JEZIKOVNIH MODELOV PRI LEKSIKOGRAFSKIH NALOGAH

Ta članek preučuje potencial ChatGPT pri avtomatizaciji dveh leksikografskih nalog v Spletnem slovarju hrvaških frazemov (ODCI): (1) prepoznavanje semantičnih ekvivalentov med frazemi in (2) generiranje semantičnih polj za frazeološke enote. Cilj raziskave je oceniti, kako učinkovito lahko umetna inteligenca avtomatizira proces razlikovanja in razvrščanja frazemov glede na pomen ter s tem zmanjša obseg ročnega leksikografskega dela. Ker se metodologije za uporabo jezikovnih tehnologij še vedno razvijajo vzporedno s tehnološkimi inovacijami, ta študija prispeva k boljšemu razumevanju delovanja orodij umetne inteligence ter njihove sposobnosti ustvarjanja kakovostnih in uporabnih jezikovnih podatkov. Rezultati kažejo, da ChatGPT izkazuje velik potencial za konceptualno organizacijo v leksikografiji. Kljub temu ostajajo izzivi, predvsem v zvezi z nedeterminističnо naravo odgovorov, generiranih z UI, in potrebo po ročnem urejanju avtomatskih podatkov.

Ključne besede: umetna inteligenca, veliki jezikovni modeli, leksikografija, frazemi, konceptualna organizacija

* **PhD, Senior Research Associate, Linguistic Research Institute, Croatian Academy of Sciences and Arts, Ante Kovačića 5, 10000 Zagreb, Croatia, ifilipovic@hazu.hr; ORCID: 0000-0001-8952-0202**

** **PhD, Assistant Professor, University of Rijeka, Faculty of Informatics and Digital Technologies, Radmile Matejčić 2, 51000 Rijeka, Croatia; University of Rijeka, Center for Artificial Intelligence and Cybersecurity, Trg braće Mažuranića 10, 51000 Rijeka, Croatia, sbeliga@inf.uniri.hr; ORCID: 0000-0003-1407-6156**

ABSTRACT

This paper explores the potential of ChatGPT in automating two lexicographic tasks within the Online Dictionary of Croatian Idioms (ODCI): (1) identifying semantic equivalents among idioms and (2) generating semantic fields for idiomatic expressions. The study evaluates how effectively AI can automate the process of distinguishing and grouping idioms by meaning, with the aim of reducing manual lexicographic work. As contemporary methodologies for employing language technologies continue to develop alongside technological progress, this research enhances understanding of AI capabilities in linguistic analysis. The findings suggest that ChatGPT shows considerable potential for conceptual organisation in lexicography. Nevertheless, challenges persist, especially regarding the unpredictable nature of AI-generated responses and the necessity for human post-editing.

Keywords: artificial intelligence, large language models, lexicography, idioms, conceptual organisation

Introduction

The rapid advancement of artificial intelligence (AI) is transforming nearly all areas of knowledge and society, including lexicography. The reflection of social and technological changes in dictionaries is not a new phenomenon – lexicographers have long embraced technological innovations such as corpora, dictionary writing systems, and user interfaces. Before advanced AI tools like ChatGPT,[1] semi-automatic dictionary creation based on post-editing lexicography gradually became the preferred method,[2] combining automatic data generation with human post-editing, where lexicographers assess, refine, and finalise entries.

The rise of AI, especially large language models (LLMs), has prompted many questions about its effect on lexicography. These questions range from whether traditional methods can be abandoned to identifying which lexicographical tasks could benefit from AI. For instance, there has been an ongoing debate about whether generative AI chatbots like ChatGPT could replace corpus-based technologies such as concordances and keyword analysis, providing greater efficiency, lower costs, and access to data that is otherwise difficult to obtain.[3]

1 ChatGPT is a chatbot and virtual assistant developed by OpenAI (launched on November 30, 2022).

2 Vít Baisa et al., "Automating Dictionary Production: A Tagalog-English-Korean Dictionary from Scratch," 805–18 (2019). Mark Davies, "AI/LLM Integration with the Corpora from English-Corpora.org," English-Corpora.org, 2025, accessed on 9 April 2025, https://www.english-corpora.org/ai-llms/corpora-vs-llms.html. Miloš Jakubíček et al., "Million-Click Dictionary: Tools and Methods for Automatic Dictionary Drafting and Post-Editing," in *Book of Abstracts of the 19th EURALEX International Congress*, 65–67 (2021). Iztok Kosem et al., "Automation of Lexicographic Work Using General and Specialized Corpora: Two Case Studies," in Andrea Abel et al., eds., *Proceedings of the 16th EURALEX International Congress* (Bolzano, Italy: EURAC Research, 2014), 355–64.

3 Gilles-Maurice de Schryver, "Generative AI and Lexicography: The Current State of the Art Using ChatGPT," *International Journal of Lexicography* 36, No. 4 (2023): 355–87, https://doi.org/10.1093/ijl/ecad021. Robert Lew,

Simultaneously, scepticism persists regarding the quality and reliability of AI-generated content.[4] Critics point out risks such as hallucinations, data inaccuracies, and the erosion of user trust in AI-generated content. The latter is particularly critical in lexicography, as dictionaries have long served as trusted sources of information,[5] a legacy rooted in the Enlightenment's prescriptive tradition.[6] Consequently, it is evident that if modern lexicography aims to incorporate advanced technologies, more empirical evidence is required to evaluate their effectiveness and the quality of the outcomes they produce.

In this paper, we examine the potential of LLMs, particularly ChatGPT, to automate two tasks within the *Online Dictionary of Croatian Idioms*[7] (ODCI): (1) identifying semantic equivalents among idioms and (2) generating semantic fields or conceptual categories for idioms. The lexicographers working on this dictionary aim not only to present dictionary content in the traditional format – entries featuring idioms – but also to develop a specialised resource: a thematic index containing semantic fields (concepts) in which idioms are grouped according to meaning and linked to their corresponding dictionary entries. Although this conceptual organisation was initially created manually, making it a highly time-consuming process, it offers a solid foundation, as human-annotated subsets of idioms can serve as benchmarks for evaluating the accuracy and quality of AI-generated results.

Testing LLMs in this context is especially important because of the complexity of idiomatic expressions, which continue to challenge current language technologies. For example, machine translation tools often translate Croatian idioms literally, ignoring their true meanings. Previous research[8] has made significant progress in connecting idioms across languages based on semantic similarity. However, these studies relied on relatively small datasets.

"ChatGPT as a COBUILD Lexicographer," *Humanities and Social Sciences Communications* 10, No. 704 (2023): doi:10.1057/s41599-023-02119-6. Hanh Thi Hong Tran et al., "Definition Extraction for Slovene: Patterns, Transformer Classifiers, and ChatGPT," in Marek Medveď et al., eds., *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century* (Brno: Lexical Computing, 2023), 19–38. Pedro A. Fuertes-Olivera, "Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes," *Lexikos* 34, No. 1 (2024): 123–40, https://lexikos.journals.ac.za/pub/article/view/1883.

4 Piek Vossen, "ChatGPT Is a Waste of Time," VU Magazine (2022), https://vumagazine.nl/professor-piek-vossen-chatgpt-is-a-waste-of-time?lang=en.

5 Michael Rundell, "Automating the Creation of Dictionaries: Are We Nearly There?," in *Proceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography (ASIALEX 2023 Proceedings)*, 1–9 (Seoul, Korea: Yonsei University, 2023).

6 Ivana Filipović Petrović, *Kada se sretnu leksikografija i frazeologija: O statusu frazema u rječniku* (Zagreb: Srednja Europa, 2018).

7 *Lexonomy*, https://lexonomy.elex.is/#/frazeoloskirjecnikhr.

8 Diego Moussallem et al., "LIdioms: A Multilingual Linked Idioms Data Set," in Nicoletta Calzolari et al., eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan: European Language Resources Association (ELRA), 2018), https://aclanthology.org/L18-1392. Ivana Filipović Petrović, Miguel López Otal, and Slobodan Beliga, "Croatian Idioms Integration: Enhancing the LIdioms Multilingual Linked Idioms Dataset," in Nicoletta Calzolari et al., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024) (Torino, Italy: ELRA and ICCL, 2024), 4106–12, https://aclanthology.org/2024.lrec-main.366.

This paper aims to evaluate how effectively AI tools can automate the lexicographic task of distinguishing and grouping the meanings of idioms to reduce the post-lexicographic workload to a manageable level. Additionally, as the methodology for integrating language technologies continues to develop and is influenced by new technological advancements, this paper also seeks to deepen the understanding of AI tools' language processing capabilities and determine how they can produce high-quality, useful linguistic data for the scientific community.

This paper builds on our previous study[9] by expanding its scope and improving its methodology. The original research examined large language models (LLMs) for automating lexicographic tasks in the *Online Dictionary of Croatian Idioms* (ODCI). With the release of GPT-4o, a key motivation for this extension was to assess its improvements over the previously tested GPT-3.5-turbo in terms of accuracy, consistency, and usability. The main enhancements introduced in this paper include: (1) evaluation of a new LLM model – a systematic comparison of GPT-4o and GPT-3.5-turbo to identify improvements in recognising semantic equivalents and generating conceptual categories for idioms; (2) refined methodology – the initial test, which compared LLMs in ranking idioms by semantic field, was restructured as a selection step, while the follow-up categorisation task with a limited idiom set was omitted due to limited additional insight; (3) optimised prompt engineering – the second lexicographic task was re-examined using refined prompt formulations to improve LLM performance and reliability. By adopting these modifications, this paper not only reinforces the empirical basis of our initial study but also provides insights into the advancing capabilities of cutting-edge LLMs for lexicographic applications.

The paper is organised as follows: the next section describes the linguistic resource and the theoretical framework for conceptual organisation in lexicography. This is followed by an outline of the tasks and findings, while the final section presents the conclusion and future directions.

9 Slobodan Beliga and Ivana Filipović Petrović, "Large Language Models Supporting Lexicography: Conceptual Organization of Croatian Idioms," in Špela Arhar Holdt and Tomaž Erjavec, eds., *Proceedings of the Conference on Language Technologies and Digital Humanities* (Ljubljana: Institute of Contemporary History, 2024), 23–46.

The Online Dictionary of Croatian Idioms: Technology and Post-editing

Lexicography

Despite promising advances in applying language technologies to dictionary compilation over the last decade, many European languages remain under-resourced, including Croatian, particularly in terms of freely available e-dictionaries and resources.[10] The project[11] to create the *Online Dictionary of Croatian Idioms* was launched in 2019 at the Croatian Academy of Sciences. It aimed to develop an open-access, born-digital dictionary based on a corpus, built with freely available lexicographic tools and the expertise of linguistically trained lexicographers.

The project introduced a post-editing lexicography model, where lexicographers evaluate and refine automatically generated data. Although this model has not been fully implemented in this dictionary, several automated processes have been utilised. For corpus searches, we used the Sketch Engine,[12] which was freely accessible to academic members through the ELEXIS project (2018–2022). Sketch Engine provided concordances from hrWaC, the largest Croatian corpus at the time.[13] Additionally, Lexonomy,[14] a platform for creating and publishing dictionaries, served as both the dictionary writing system and publishing platform.

Lexicographic processing combines manual and automated approaches. Concordances were manually examined and analysed, while multi-word expressions were extracted using the Word Sketch feature. Frequency and usage statistics, including the LogDice metric, which evaluates the strength of word associations in collocations, helped identify commonly co-occurring terms. The GDEX (Good Dictionary Example) algorithm was employed to generate a list of candidate examples, which lexicographers reviewed manually to ensure they were typical and illustrative for dictionary entries. Entries in Lexonomy were compiled manually. Version 2, released in 2023, contains 563 entries and 1,165 idioms.[15]

10 Georg Rehm and Andy Way, *European Language Equality: A Strategic Agenda for Digital Language Equality* (Springer Nature, 2023), https://doi.org/10.1007/978-3-031-28819-7.

11 *Frazeološki rječnik*, https://frazeoloski-rjecnik.eu/en/

12 Adam Kilgarriff et al., "The Sketch Engine: Ten Years On," Lexicography 1, No. 1 (2014): 7–36.

13 Nikola Ljubešić and Filip Klubička, *Croatian Web Corpus hrWaC 2.1* (Slovenian language resource repository CLARIN.SI, 2016), http://hdl.handle.net/11356/1064.

14 *Lexonomy*, https://lexonomy.elex.is/

15 Ivana Filipović Petrović and Jelena Parizoska, *Frazeološki rječnik hrvatskoga jezika v2* (Zagreb: Hrvatska akademija znanosti i umjetnosti, 2023), https://lexonomy.elex.is/#/frazeoloskirjecnikhr.

Conceptual organisation

The advancement of technology has created many opportunities for presenting dictionary content digitally. The introduction of hyperlinks connecting entries that are far apart alphabetically – such as multi-word expressions with similar meanings but different structures – would likely have impressed lexicographers from the pre-digital era. They often compiled extensive lists to categorise expressions by domains of knowledge, attempting to make them searchable within the linear format of printed media. For phraseological dictionaries in particular, the ability to link idioms with very different expressions is revolutionary. For example, idioms like *fali komu daska u glavi* (lit. someone is missing a plank in the head) and *nisu komu sve koze na broju* (lit. someone's not keeping a tab on all their goats) both mean 'to be crazy or insane.' In a printed dictionary, these idioms might only appear under the first noun in the construction (such as plank or goat), potentially missing the connection to the related idiom. As a result, lexicographers have long sought ways to show semantically related words, though alphabetical order still dominates most dictionaries. Advocates of conceptual organisation believe it better reflects how the human mind categorises ideas and words, arguing that lexicography should assist users in finding both words and meanings, starting from ideas or concepts.[16]

In line with this, the conceptual organisation for the *Online Dictionary of Croatian Idioms* (ODCI) was manually compiled. Sixty-four concepts were established, into which 430 idioms have been categorised so far. The lexicographers responsible for this process based their work on best practices from notable lexicographical works, such as the *Collins COBUILD Idioms Dictionary* (2002) and the *Cambridge Idioms Dictionary* (2006). These dictionaries organise idioms alphabetically and also include sections where they are grouped by themes such as love, honesty, deception, disagreement, success, failure, happiness, and sadness. This idea of organisational grouping can be found in renowned thesauruses, such as *Roget's Thesaurus of English Words and Phrases* (1852), and has been adapted over time to suit the constraints of the medium and the nature of the dictionary. A word or idiom may be categorised under multiple concepts, with decisions guided by human knowledge, beliefs, and instincts.

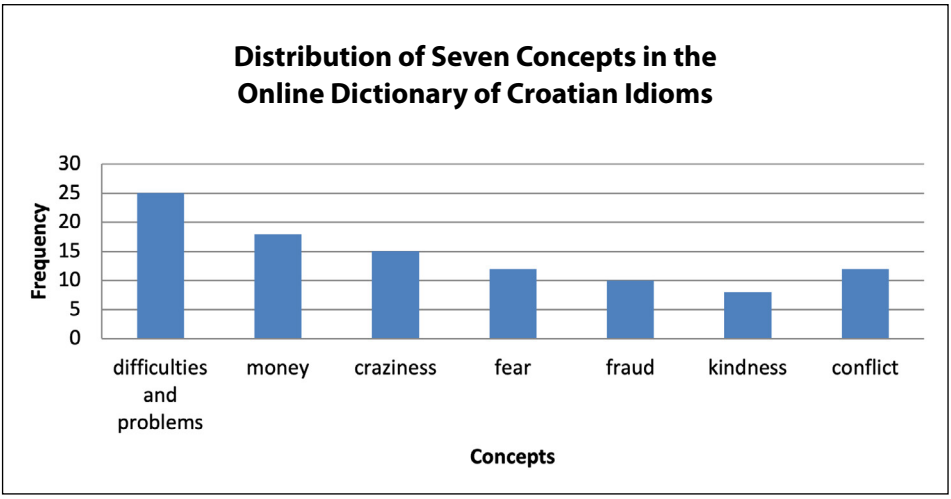
In this paper, we focus on comparing the outputs of artificial intelligence and human intelligence in conceptual organisation. The criterion for linking semantically similar idioms in the ODCI involves identifying common semantic and structural elements.[17] As an example, we randomly selected seven concepts to demonstrate the manually crafted conceptual organisation for the ODCI, accompanied by a diagram (Chart 1) showing the frequency distribution of idioms across these concepts.

16 Dirk Geeraerts, "Principles of Monolingual Lexicography," in Franz J. Hausmann, ed., *Wörterbücher. Ein Internationales Handbuch Zur Lexikographie*, Vol. 1 (Berlin: Walter de Gruyter, 1989), 287–96. Tom McArthur, *Worlds of Reference: Lexicography, Learning, and Language from the Clay Tablet to the Computer* (Cambridge: Cambridge University Press, 1986).

17 Ivana Filipović Petrović and Jelena Parizoska, "Konceptualna organizacija frazeoloških rječnika u leksikografiji," *Filologija* 73 (2019): 27–45.

Concepts such as difficulties/problems, money, and conflict stand out as rich sources of idiomatic expressions. Entries in the ODCI include links to semantically related idioms. Additionally, a separate conceptual index was created, listing concepts and corresponding idioms as links to corresponding dictionary entries, allowing users to search by ideas rather than solely by words. Although a valuable resource, this conceptual index was labour-intensive to produce and would benefit from further refinement and expansion.

Chart 1: Distribution of seven concepts in the Online Dictionary of Croatian Idioms



Source: Own work, based on data from the Online Dictionary of Croatian Idioms

As corpus research and the automatic identification of idioms in corpora continue to advance, the ODCI will be expanded with new entries. To facilitate this, further technological improvements are being sought to automate the process of conceptual organisation. The aim is to implement this process at three levels.

– On the existing material: The objective is to categorise the remaining uncategorised idioms by determining whether they fit into current concepts or by proposing new ones. Each idiom should be assigned to a specific concept based on its meaning, even if it currently stands alone. This method will enable future idioms to be grouped under the same concept, allowing users to search the dictionary by ideas and meanings. Over time, as new idioms are added, these concepts will evolve and expand.
– On new material: As new entries are added to the dictionary, new meanings will emerge. Corresponding concepts will be identified, and additional idioms will be linked to them.
– For new idioms not fitting into existing concepts: New concepts will be proposed for idioms that do not fit into any established category, thus expanding the list of entries in the conceptual index.

In this research, we conducted a pilot study involving several large language models (LLMs) and a selection of manually created concepts from the ODCI. We performed one experiment, followed by two tasks assigned to the AI system, which completed both tasks. The next section outlines the procedures used in this study.

Lexicography and Large Language Models: Let's Give It a Try

In the initial phase of this research, we aimed to test LLMs on a trial example to identify which one produces the best results and which model we will continue using for the two planned tasks. First, we tested three readily available, open-source large language models (LLMs) designed for the academic community and trained on some Croatian texts, to evaluate their performance and potential usefulness for other studies.

We used the Cro-CoV-cseBERT model,[18] a fine-tuned version of the CroSloEngualBERT (cseBERT) model.[19] CroSloEngualBERT is a trilingual BERT-based language model pre-trained on a large corpus of online news articles in Croatian, Slovenian, and English (5.9 billion tokens, comprising 31% Croatian, 23% Slovenian, and the rest in English). Cro-CoV-cseBERT was specifically fine-tuned on Croatian language corpora related to COVID-19, including 186,738 news articles, 500,504 user comments from Croatian online news portals, and 28,208 COVID-19-related tweets.[20] Cro-CoV-cseBERT is fine-tuned for masked language modelling.

The second model employed was the bcms-bertic (BERTić),[21] a transformer model pre-trained on 8 billion tokens of crawled text from Croatian, Bosnian, Serbian, and Montenegrin web domains. BERTić was trained using the ELECTRA transformer architecture. Both BERTić and cseBERT are base-sized models.

In addition to these BERT and ELECTRA architectures, we examined the effectiveness of a Generative Pre-trained Transformer (GPT) model, specifically gpt2-vrabac,[22] a smaller generative model for the Serbian language. Considering the linguistic proximity of Croatian and Serbian, we hypothesised that gpt2-vrabac might provide useful insights. This model, based on the GPT2-small architecture, contains 136 million parameters and was trained on approximately 4 billion tokens derived

18 Karlo Babić et al., "Characterisation of COVID-19-Related Tweets in the Croatian Language: Framework Based on the Cro-CoV-cseBERT Model," *Applied Sciences* 11, No. 21 (2021), https://www.mdpi.com/2076-3417/11/21/10442, doi:10.3390/app112110442.

19 Matej Ulčar and Marko Robnik-Šikonja, "FinEst BERT and CroSloEngual BERT: Less Is More in Multilingual Models," in *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings* (Berlin, Heidelberg: Springer-Verlag, 2020), 104–11, https://doi.org/10.1007/978-3-030-58323-1_11.

20 Karlo Babić et al., "Characterisation of COVID-19-Related Tweets," 2021.

21 Nikola Ljubešić and Davor Lauc, "BERTić – The Transformer Language Model for Bosnian, Croatian, Montenegrin, and Serbian," in *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing* (Kiyv, Ukraine: Association for Computational Linguistics, 2021), 37–42.

22 Mihailo Škorić, "Novi jezički modeli za Srpski jezik," *Infoteka* 24 (2024).

from doctoral dissertations, a corpus of Serbian public discourse, web-crawled texts, and the Society for Language Resources and Technologies corpus.

These LLMs were used to determine the semantic similarity between a specified semantic field (e.g., kindness) and a comprehensive corpus of Croatian idiomatic expressions. For each idiom and semantic field, the lexical units were tokenised, and the resulting tokens were embedded using the relevant language model. The token vectors were then combined and normalised by token count to obtain an averaged vector representation (centroid-averaged token vectors). This process yielded a unique vector for each idiom and a distinct vector for the semantic field. Afterwards, cosine similarity was calculated to quantify the semantic correspondence between the semantic field and each idiom, with higher scores indicating stronger semantic similarity.

The selection of specific transformer architectures was crucial for effectively measuring semantic similarity. While the Cro-CoV-cseBERT (BERT-based) and bcms-bertic (ELECTRA-based) models, as encoder-focused transformers, are inherently designed for comprehending bidirectional context and generating robust semantic vector representations of text, the inclusion of gpt2-vrabac (GPT-based) model, which relies on a decoder-only unidirectional architecture, allowed for comparative analysis. This architectural diversity provided comprehensive insights into their respective strengths and limitations when calculating semantic correspondence between semantic fields and Croatian idiomatic expressions, ensuring a thorough evaluation of each model's suitability for lexicographical tasks.

Beyond open-source models, we also evaluated the performance of the commercially developed GPT-3.5-turbo and GPT-4o models (from OpenAI)[23] for idiom-to-semantic-field matching using prompt engineering. GPT-3.5-turbo, a transformer-based model with 175 billion parameters, features 96 transformer layers, 12,288-dimensional hidden states, and 96 attention heads per layer. This architecture offers significant advantages in pattern recognition and generation compared to base-sized models like BERTić and cseBERT (12 hidden layers, 768 hidden states). Notably, despite not being trained on extensive Croatian corpora, recent research[24] has demonstrated the efficacy of GPT models for Croatian causal commonsense reasoning, including dialectal variations (DIALECT-COPA). Although the exact number of parameters in GPT-4o has not been officially disclosed, it is presumed to be substantially larger than the 175 billion parameters of GPT-3.5-turbo, enhancing its ability to generate more complex and accurate responses. GPT-4o supports a contextual window of 128,000 tokens, a considerable increase compared to the 4,096 tokens in GPT-3.5-turbo. This enhancement enables the model to better comprehend and generate longer and more intricate texts.

23 Tom B. Brown et al., "Language Models Are Few-Shot Learners," in Hugo Larochelle et al., eds., *Advances in Neural Information Processing Systems*, Vol. 33 (Curran Associates, Inc., 2020), 1877–1901.

24 Benedikt Perak, Slobodan Beliga, and Ana Meštrović, "Incorporating Dialect Understanding into LLM Using RAG and Prompt Engineering Techniques for Causal Common-Sense Reasoning," in Yves Scherrer et al., eds., *Proceedings of the 11th Workshop on NLP for Similar Languages, Varieties, and Dialects* (VarDial 2024) (Mexico City, Mexico: ACL, 2024), 220–29, https://aclanthology.org/2024.vardial-1.19.

GPT-4o shows significant progress in language understanding and handling complex tasks, making it more reliable and accurate than earlier versions like GPT-3.5-turbo. In terms of linguistic comprehension, GPT-4o demonstrates improved skill in correctly interpreting sentence meanings, contextual nuances, and deeper logical relationships.

The initial experiment utilising the GPT-3.5-turbo model was carried out in April 2024, while the follow-up experiment employing the more advanced GPT-4o model took place in February 2025.

Selection of the Best-performing LLM for Subsequent Tasks

From the manually created conceptual organisation in the ODCI, a sample of 150 idioms was selected, distributed across 27 concepts. For the testing of LLMs, three concepts were chosen from this conceptual index: KINDNESS, MADNESS, and CONFLICT. Table 1 presents the selected concepts along with their corresponding idioms.

Table 1: Concepts and corresponding idioms in the selection experiment for the best-performing LLM in subsequent tasks

Concept	Idioms
kindness	dobar kao kruh (lit. as good as bread) 'very good, hearted', duša od čovjeka (lit. soul of a person) 'a kind person', ne bi ni mrava zgazio 'wouldn't hurt a fly'
madness	fali daska u glavi komu (lit. someone is missing a plank in the head) 'not normal', lud kao šiba 'crazy like a hatter', lud sto gradi 'crazy like a hundred', nisu sve koze na broju komu (lit. not all the goats are in the pen) 'crazy, not normal', nisu svi doma komu (lit. not everyone is at home) 'crazy, not normal', posvađao se s mozgom (lit. quarreled with the brain) 'lost one's mind', zreo za ludnicu (lit. ripe for the madhouse), puknuti kao kokica (lit. to pop like a popcorn) 'go crazy', najesti se ludih gljiva (lit. to eat mad mushrooms) 'go crazy'
conflict	dolijevati ulje na vatru (lit. to pour oil on the fire) 'further inflame a conflict or disagreement', izvrijeđati na pasja kola koga 'to verbally abuse someone thoroughly', lome se koplja (lit. spears are breaking) 'there's a fierce conflict', posijati sjeme razdora (lit. to sow the seeds of discord), posvađati se na mrtvo ime 'to fight bitterly', posvađati se na pasja kola 'to fight fiercely', stvarati zlu krv (lit. to create bad blood), svađati se kao pas i mačka (lit. to fight like cats and dogs), prosipati žuč (lit. to spill bile) 'to express bitterness', spaliti mostove (lit. to burn bridges), ukrstiti koplja (lit. to cross swords) 'to engage in a conflict'

Source: Own work

In the experiment, LLMs were used to calculate the semantic similarity between idioms and their respective semantic fields. The task was structured as follows: from a list of 150 idioms, the algorithm identified those belonging to the following semantic fields: 1) kindness, 2) madness, and 3) conflict. LLMs such as Cro-CoV-cseBERT, bcms-bertic, and gpt2-vrabac ranked idioms related to kindness between 47th and 65th place on average. The highest ranking was achieved by gpt2-vrabac, which placed the idiom *duša od čovjeka* (lit. a soul of a person) – i.e., 'a kind person' – in fifth place. The idioms *zlatna koka* (lit. golden goose) or 'cash cow', *mala beba* (lit. little baby) or 'something easy to use, harmless', and *malo sutra* or 'no way, no chance' were ranked first.

For the concept of MADNESS, the Cro-CoV-cseBERT model ranked *zreo za ludnicu* (lit. 'ripe for the madhouse') in first place, *lud kao šiba* ('crazy as a hatter') in 5th, and *lud sto gradi* ('one hundred per cent crazy') in 10th. The gpt2-vrabac model placed *lud sto gradi* in the 5th, *zreo za ludnicu* in 6th, and *lud kao šiba* in 13th, while bcms-bertic ranked *lud sto gradi* in 22nd, with all other idioms ranked further down.

For the concept of CONFLICT, the bcms-bertic model ranked *stvarati zlu krv* (lit. 'to create bad blood') in 8th place, while gpt2-vrabac placed *lome se koplja* (lit. 'spears are breaking') or 'there's a fierce conflict' in first position, and *ukrstiti koplja* (lit. 'to cross swords') or 'to engage in a conflict' in 6th. Meanwhile, Cro-CoV-cseBERT ranked *prosipati* žuč (lit. 'to spill bile') or 'to express bitterness' highest, assigning it 24th place. In this ranking, a lower number indicates a better result. For instance, being ranked first indicates the system considers the idiom the best match for the given concept of KINDNESS. Conversely, rankings of 47th and 65th suggest those idioms are considered poor matches for the concept.

Although several idioms paired with predefined concepts were successfully ranked, the overall results for all idioms listed in Table 1 are not adequate for lexicographic use. The examined LLMs for Croatian do not produce high-quality results for figurative language, possibly because of the varied types of texts used in model training. For instance, BERTić was trained on a large corpus containing diverse content, including web pages, literary works, and newspaper articles.[25] While this corpus is not specifically tailored for idioms, it naturally includes many idiomatic expressions found in everyday language. However, the number of idioms present appears insufficient for the LLM to be effective in our lexicographic task, indicating there is significant room for improvement in this area. A more idiom-rich corpus, combined with techniques such as fine-tuning, transfer learning, or other model enhancement methods, could produce better results. Moreover, Croatian currently lacks extensive corpora rich in idiomatic expressions, which are crucial for training language models to improve their performance on our lexicographic tasks.

Furthermore, the difficulty of multi-word constructions not reflecting the sum of their parts is well known in natural language processing. Even human learners struggle

25 Ljubešić and Lauc, "BERTić," 2021.

to master idiomatic expressions when learning a foreign language.[26] The choice of idioms such as *mala beba* (lit. little baby) and *zlatna koka* (lit. golden goose) for the concept of kindness suggests that the literal meanings of the components were considered, with words like 'baby' and 'golden' being associated with the notion of goodness.

Table 2: Ranking results of idioms by their semantic proximity to the concepts of kindness, madness and conflict, as produced by transformer-based language models. Rankings closer to the top are considered more successful.

Model	KINDNESS	MADNESS	CONFLICT
	best ranked idiom		
gpt2-vrabac	duša od čovjeka 'a very kind, good-hearted person' (5th place)	lud sto gradi 'completely mad, insane' (5th place)	lome se koplja 'there's a fierce argument or conflict' (1st place)
Cro-CoV-cseBERT	ne bi ni mrava zgazio 'extremely gentle, harmless' (47th place)	zreo za ludnicu 'ready for the asylum; mentally unstable' (1st place)	prosipati žuč 'to express bitterness or strong anger' (24th place)
bcms-bertic	duša od čovjeka 'a very kind, good-hearted person' (65th place)	lud sto gradi 'completely mad, insane' (22nd place)	stvarati zlu krv 'to cause hostility or resentment' (8th place)

Source: Own work

The query was then repeated with ChatGPT, asking it to, in the role of a lexicographer and linguist, identify the ten most relevant idioms from the list of 150 Croatian idioms that belong to the semantic fields of MADNESS, CONFLICT, and KINDNESS – that is, those that are semantically closest to these concepts. When role-play prompting is used in a prompt, this technique provides the model with a contextual instruction to adjust its reasoning, response style, and task approach according to the assigned role. Existing research[27] shows that this method improves the reasoning skills of LLMs, even in scenarios where the model has no prior examples (zero-shot settings). Our findings align with the manual organisation of concepts and idioms in 98% of cases. Three idioms related to kindness ranked in the top three positions, and nine idioms related to madness appeared among the top nine. For the concept of conflict, six idioms matched, but ChatGPT did not include *dolijevati ulje na vatru* ('further inflame a conflict'), *prosipati žuč* ('to express bitterness'), *ukrstiti koplja* ('to cross swords'),

26 Julia Miller, "Research in the Pipeline: Where Lexicography and Phraseology Meet," *Lexicography ASIALEX* 5, No. 1 (2018): 23–33, doi:10.1007/s40607-018-0044-z.

27 Aobo Kong et al., "Better Zero-Shot Reasoning with Role-Play Prompting," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (Mexico City: Association for Computational Linguistics, 2024), 4099–113, https://aclanthology.org/2024.naacl-long.228/.

or *stvarati zlu krv* ('to create bad blood'). Instead, it added idioms such as *braniti se rukama i nogama* ('to defend oneself tooth and nail'), *digla se kuka i motika* ('to rebel'), and *dignuti se na zadnje noge* ('to stand up on hind legs'). In manual classification, the first idiom is categorised as avoidance, while the latter two are classified as rebellion. ChatGPT's classification is not necessarily wrong, as categorisation depends on interpretation and idiom usage is heavily context-dependent. Conflict generally refers to disagreement, opposition, or tension, while avoidance involves deliberately steering clear of conflict, which can be implied in some contexts. Rebellion indicates resistance or opposition to authority or established norms, which can sometimes lead to conflict. In this sense, ChatGPT performed well in this experiment.

Task One

We conducted a preliminary test experiment to gain insights into the data provided by LLMs. The aim was to identify their strengths and weaknesses. Based on our findings, we decided to focus our research on OpenAI's GPT model, as it demonstrated superior results compared to other models. Therefore, the following steps involve utilising AI to generate a dataset that lexicographers can use for dictionary creation. As mentioned, there are currently 1,165 idiomatic expressions in the ODCI. Thematic fields were manually identified for 430 entries to establish a dictionary feature that allows users to easily find expressions related to their desired topic or idea. To ensure accuracy, we wanted to verify if the remaining idiomatic expressions can be classified into one of the already manually defined semantic fields.

The experiment utilised a role-playing prompt designed for zero-shot settings (prompts were in Croatian):

model="gpt-3.5-turbo", messages=[

{"role": "system", "content": "Take on the role of a lexicographer creating a new conceptually organized phraseological dictionary of Croatian. Please respond in Croatian."}

{"role": "user", "content": "A list of pre-defined semantic fields is provided."}

{"Link the idiom to the most appropriate semantic field from the provided list. Respond by choosing only one of the offered semantic fields."

}]

To demonstrate the results, we will use examples of two concepts: communication and knowledge. Using manual classification, we sorted 19 idioms into the category of communication. In Table 3, we show how these idioms relate to the results obtained from ChatGPT, which also identified 13 of them as being associated with communication.

Table 3: Results of task 1 inquiry using the example of the COMMUNICATION category

Idioms manually classified into the concept communication	**ChatGPT-3.5-turbo responses**
baciti bubu u uho komu (lit. to plant a bug in someone's ear) 'to make someone suspicious or curious'	communication
bacati drvlje i kamenje na koga, što (lit. to throw sticks and stones at someone/something) 'to criticize harshly'	conflict
čašica razgovora 'a friendly chat'	communication
čupati kliještima iz koga što (lit. to extract something from someone with pliers) 'to forcefully extract information'	fighting
pričati Markove konake 'to tell long and boring stories'	communication
pričati kao navijen 'to talk incessantly, like a broken record'	communication
razgovarati na ravnoj nozi 'to talk on equal terms'	communication
reći komu što ga ide 'to tell someone off'	communication
reći popu pop, a bobu bob 'to call a spade a spade'	communication
reći u lice 'to say to someone's face'	communication
šutjeti kao pizda 'to keep silent' (vulgar, lit. to be silent like a cunt)	communication
šutjeti kao zaliven 'to be silent as the grave'	communication
zatvoriti se u ljušturu 'to withdraw into one's shell'	unknown
prosipati pamet 'to dispense wisdom, to pretend to be wise'	communication
srati kvake 'to talk nonsense' (vulgar, lit. to shit handles)	communication
prenositi se od usta do usta 'to spread by word of mouth'	communication
umotati u celofan 'to sugarcoat'	ingratiation
obilaziti kao mačak oko vruće kaše 'to beat around the bush'	avoidance
lagati u oči komu 'to lie to someone's face'	fraud

Source: Own work

Furthermore, under the concept of KNOWLEDGE, we manually classified the following idioms: *znati* što *kao vodu piti* ('to know something like the back of your hand'), *imati u malom prstu* što ('to have something at your fingertips'), and *isisati iz malog prsta* što ('to pull something out of thin air, to come up with something effortlessly'). GPT-3.5-turbo classified the idiom *znati* što *kao vodu piti* ('to know something like the back of your hand') under knowledge, while it associated *imati u malom prstu* ('to have something at your fingertips') with the concept of control, and *isisati iz malog prsta* što ('to pull something out of thin air, to come up with something effortlessly') with the concept of ease or difficulty. However, GPT-3.5-turbo also classified the idiom *imati dobar nos* (lit. 'to have a good nose'), which was previously unclassified, under the concept of KNOWLEDGE, as it means to have the ability or instinct for something (which can include KNOWLEDGE).

Additionally, GPT-3.5-turbo included two uncategorised idioms: *gurati pod nos komu* što, meaning 'to shove something in someone's face' (literally 'nose') or 'to impose something on someone'; and *objaviti na sva zvona*, meaning 'to shout it from the rooftops' or 'to announce something to everyone'. Examples of usage for the idiom *to shove something in someone's face* (1 and 2) and for the idiom *to shout it from the rooftops* (3 and 4) found in the ODCI show the context of COMMUNICATION:

If you push your views and principles under his nose on the first date and show him your great intelligence, he will get the impression that you're lecturing him.
In every argument, he brings up the issues that have been resolved, re-analyzes them, and puts them under the nose.
After deciding to get engaged, many couples in love don't want to shout it from the rooftops to everyone right away but will keep their sweet secret for some time.
Don't shout it from the rooftops that you've just received your paycheck, bought new household appliances, or saved a large sum of money, as some of the useful tips the police have given to citizens.

Overall, the results offered by GPT-3.5-turbo for Task 1 proved helpful in further lexicographical considerations. In other words, while these results cannot be considered a finished dataset, they can help by providing a comprehensive overview and potential ideas for different categorisations. To enhance efficiency in dictionary creation, a model should perform better and make fewer errors, such as merging *krenuti čijim stopama* ('to follow in someone's footsteps') with the concept of excitement. This would enable lexicographers to integrate more data with minimal intervention.

Additional examples of conceptual misclassifications further demonstrate the model's limitations in interpreting idioms. Table 4 presents a set of Croatian idioms that were manually assigned to semantic fields such as perseverance, threat, or mental instability, but were misclassified by GPT-3.5-turbo due to literal interpretation or misalignment with figurative meanings. For instance, the idiom *zapeti kao sivonja* ('to be relentless in one's effort') was incorrectly associated with immobility, while *nisu sve ovce na broju komu* ('someone is a bit crazy or mentally off') was linked to incompleteness instead of the intended domain of mental instability. Such examples highlight the need for a deeper understanding of context and culturally informed processing of idioms to enable reliable classification.

Table 4: Examples of incorrect or unexpected model classifications

Idioms	Human-assigned semantic field	Conceptual misclassifications by GPT-3.5-turbo
zapeti kao sivonja 'to be relentless in one's effort'	perseverance	immobility
naći se u neobranom grožđu 'to be in a tight spot'	unfavorable situation	surprise
trese se stolica komu 'someone's position is on shaky ground'	threat	fear
nisu sve ovce na broju komu (lit. not all the sheep are accounted for) 'someone is a bit crazy or mentally off '	mental instability	incompleteness

Source: Own work

Task Two

When we carried out Task 2 for the first version of the research and asked the model to group idioms by meaning and assign names to semantic fields, we observed two recurring issues with the GPT-3.5-turbo model: crafting effective prompts and generating unique responses each time. The first issue suggests that we might have needed to instruct the model to group more semantically related idioms under a single concept rather than continually offering different concepts. However, this conflicts with the model's inherent non-deterministic nature, as it consistently produces different responses to the same prompt. Here, we will highlight the crucial parts of the results. For a group of idiomatic expressions, the model proposed the following concepts: EMOTIONS, EMOTIONAL REACTIONS, EMOTIONAL STATES, and EMOTIONAL CLOSENESS (see Table 5). On one hand, the detailed breakdown of the concept of EMOTIONS – dividing it into REACTIONS, STATES, and CLOSENESS – can be very useful, as it aligns with the further subdivision into sub-concepts considered in the manual classification. On the other hand, when considering that a user may search for dictionary entries based on a particular concept, such as HAPPINESS, it becomes clear that the broader concept of EMOTIONS, even with additional details on REACTIONS, is too abstract to serve the goal of conceptual organisation. The aim is to guide the user by offering concrete, usable information. For instance, idioms like *crven od bijesa* ('red with anger'), *kipjeti od bijesa* ('boiling with anger'), *ljut kao ris* ('angry as a lynx'), *ljut kao vrag* ('angry as the devil'), *para ide na uši komu* (lit. 'steam coming out of someone's ears'), *pao je mrak na oči komu* (lit. 'darkness fell over someone's eyes'), *poludjeti od bijesa* ('go mad with anger'), *pozelenjeti od bijesa* ('turn green with anger'), and *puknuo je film komu* (lit. 'someone's film broke') are all semantically linked to the concept of ANGER. Similarly, the model categorised the idiom *ne bi ni mrava zgazio* ('wouldn't

hurt a fly') under the concept of MERCY and EMPATHY, and *duša od čovjeka* (lit. 'a soul of a man', 'a kind-hearted person') under the concept of PERSONALITY TRAITS. Both are manually classified under the concept of KINDNESS.

Table 5: The concepts proposed by the GPT-3.5-turbo model and associated idioms

Concept created by the GPT-3.5-turbo	Associated idiom
emotions	umrijeti od smijeha 'die laughing', tresti se od bijesa 'shake with anger', zaljubiti se do ušiju 'fall head over heels in love', blagi očaj 'mild despair', duša od žene 'woman with a kind heart', srce se steže komu 'someone's heart tightens'
emotional reaction	puknuo je film komu 'someone snapped, lost it', dignuti se na stražnje noge (lit. get up on one's hind legs) 'stand up for oneself', poludjeti od bijesa 'to go mad with rage', rasplakati se kao malo dijete 'cry like a little child', plakati kao beba 'cry like a baby'
emotional condition	nervozan kao pas 'nervous as a dog', ljut kao vrag 'angry as hell', bijesan kao pas 'mad as a hornet', zaljubljen kao tele 'infatuated, puppy love', baciti u očaj koga 'to drive someone to despair'
emotional closeness	zavući se pod kožu komu 'to get under someone's skin'
negative emotions	proliti žuč 'to vent one's spleen'

Source: Own work

The second attempt produced the following results (Table 6): the model categorised the idioms *zaljubiti se do ušiju* ('fall head over heels in love') and *zaljubljen kao tele* ('infatuated, puppy love') under the concept of LOVE and ATTACHMENT, while *ljut kao vrag* ('angry as hell') and *bijesan kao pas* ('mad as a hornet') were placed under the concept of ANGER and FRUSTRATION. Unlike the results obtained in the previous attempt, this categorisation presents fully usable and well-structured semantic fields for lexicographic purposes.

This clearly demonstrates the impact of the differently structured prompt, which explicitly instructed that the concept names should be sufficiently specific while also ensuring that as many idioms as possible were grouped under the same concept if they shared a common meaning. The initial prompt used with GPT-3.5-turbo produced overly specific and inconsistent concept groupings. After manually analysing the outputs, a revised prompt was designed for GPT-4, with clearer and more targeted instructions. It explicitly directed the model to assign idioms to the same semantic field whenever they shared a common meaning and to name those fields in a manner that was sufficiently specific yet suitable for a lexicographic context. This exemplifies iterative prompt design and refinement, a prompt engineering strategy where an expert iteratively modifies prompts based on model outputs to improve task performance. Rather than relying on automated self-feedback mechanisms – as in fully autonomous

self-refinement systems (cf. Madaan et al. 2023)[28] – this study employed a human-in-the-loop approach, involving manual evaluation of initial outputs and subsequent prompt revision to better align with the intended semantic grouping behaviour. A similar refinement principle was used in the study on model truthfulness by Krishna et al., where iterative prompting strategies were developed and tested to improve the factual accuracy and reliability of LLM outputs.[29] This structured revision proved effective: the original prompt caused the model to assign a unique concept to almost every idiom, resulting in overly specific and inconsistent categories. In contrast, the revised prompt provided clearer constraints and more targeted guidance, encouraging the model to cluster idioms under shared semantic fields with labels that were both specific enough and lexicographically useful. The improved prompt, used with GPT-4o for the semantic grouping task, was as follows:

> *I will send you a list of Croatian idioms. You need to group them by meaning into semantic fields and give those fields names in Croatian. Try to group idioms with similar meanings together and give them a sufficiently specific name that describes their meaning, e.g., happiness, sadness, quarrel, obstacle, love, etc. Also, try to categorize as many idioms as possible into the same concept that they share in meaning.*

Table 6: The concepts proposed by the GPT-4o model and associated idioms

Concept created by the GPT-4	Associated idiom
emotions and reactions	umrijeti od smijeha 'die laughing', tresti se od bijesa 'shake with anger', plakati kao beba 'cry like a baby', rasplakati se kao malo dijete 'cry like a little child', poludjeti od bijesa 'to go mad with rage', proliti žuč 'to vent one's spleen'
love and attachment	zaljubiti se do ušiju 'fall head over heels in love', zaljubljen kao tele 'infatuated, puppy love'
states and feelings	blagi očaj 'mild despair', baciti u očaj koga 'to drive someone to despair', srce se steže komu 'someone's heart tightens'
defense and resistance	dignuti se na stražnje noge 'stand up for oneself', puknuo je film komu 'someone snapped, lost it'
anger and frustration	ljut kao vrag 'angry as hell', bijesan kao pas 'mad as a hornet', nervozan kao pas 'nervous as a dog'
influence and manipulation	zavući se pod kožu komu 'to get under someone's skin'
traits	duša od žene 'woman with a kind heart'

Source: Own work

28 Aman Madaan et al., "SELF-REFINE: Iterative Refinement with Self-Feedback," in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023)* (Red Hook, NY: Curran Associates Inc., 2023), 2019, 1–61, https://selfrefine.info/.

29 Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju, "Understanding the Effects of Iterative Prompting on Truthfulness," in *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)* (Vienna: JMLR.org, 2024), Paper 1024, 1–20.

Additionally, while the concepts of EMOTIONS AND REACTIONS and STATES AND FEELINGS are overly broad in lexicographic terms – and the same criticism applies to the categorisation generated by GPT-3.5 in the previous query – it is important to recognise that these concepts are not fundamentally incorrect and can still be helpful in lexicography. This is especially relevant when working with large amounts of linguistic data, as such categorisation can facilitate further manual processing. The lexicographer's task of grouping all semantically related idioms under a single concept – one that is broad enough to encompass multiple instances but specific enough to provide useful, concrete information for users – is inherently highly subjective. Therefore, this step ultimately necessitates manual intervention.

Limitations

The results of this study should be considered in light of several methodological and conceptual limitations. Firstly, although the dataset of 150 idioms was carefully chosen to reflect a range of meanings and structures, the expressive and culturally embedded nature of idioms means that high performance on this subset does not necessarily ensure applicability to the entire spectrum of Croatian phraseology. Idioms are often context-dependent, metaphorically dense, and semantically overlapping, making generalisation particularly challenging.

Secondly, although ChatGPT was used as the primary model in the later stages of the study and smaller Croatian LLMs were initially tested, the selection of general-purpose LLMs raises broader questions about their appropriateness for specialised tasks like semantic lexicographic classification. These models are not trained explicitly for idiom interpretation or lexical-semantic organisation, and their performance can differ across languages and idiomatic structures.

Finally, conceptual organisation in lexicography, especially when grouping idioms by meaning, is a highly interpretive task. There is no universally accepted or correct way to categorise idiomatic meaning, as it reflects not only linguistic but also encyclopaedic and cultural knowledge. Therefore, both the human-created reference classification and the model-produced groupings remain inherently subjective to some extent.

Conclusion

This paper examined the performance of large language models (LLMs) in analysing the semantic features of multi-word expressions with figurative meanings, particularly idioms. The study compared smaller, open-source models (CroCoV-cseBERT, bcms-bertic, and gpt2-vrabac) with more advanced proprietary models, GPT-3.5-turbo and GPT-4o. The results demonstrated a clear performance gap,

with the proprietary GPT-4o model delivering the most accurate and semantically coherent results, highlighting its improvements in linguistic comprehension and contextual reasoning.

Although LLMs, especially GPT-based models, have demonstrated potential in lexicography, it is essential to recognise the specific challenges in this highly specialised field. Lexicography involves intricate tasks such as identifying common syntax patterns, selecting collocations, and producing precise definitions and examples, which can be challenging even for human experts. While LLMs show promise in supporting dictionary development, issues persist, particularly in managing subtle phraseological meanings and ensuring consistency in semantic categorisation.

Human expertise remains essential in lexicography, as LLMs cannot yet fully replicate the depth of understanding needed for complex lexicographic tasks. Although the findings are encouraging, the study is limited by the relatively small sample of idioms and the narrow range of semantic categories examined. The models also showed a tendency to misclassify idioms by relying too heavily on literal meanings or by assigning overly broad conceptual labels. Their overall performance is further constrained by the lack of idiom-rich training corpora, particularly for under-resourced languages like Croatian.

To improve the automation of lexicographic workflows, future research should aim to enhance query precision, fine-tune LLMs on corpora rich in idioms, and develop hybrid models that blend rule-based approaches with generative AI. Moreover, expanding studies to include a variety of language models and alternative conceptual frameworks could offer deeper insights into how AI can be effectively employed in lexicographic practice, especially for under-resourced languages like Croatian.

Acknowledgements

This work has been fully supported by the Croatian Science Foundation under the project *Semantic-Syntactic Classification of Croatian Verbs (SEMTACTIC)* (IP-2022-10-8074) and by the project *Hybrid AI Approaches to Natural Language Processing and Knowledge Generation – HyAI* (uniri-iz-25-215), funded by the European Union – NextGenerationEU.

Sources and Literature

Literature

Beliga, Slobodan, and Ivana Filipović Petrović. "Large Language Models Supporting Lexicography: Conceptual Organization of Croatian Idioms." In *Proceedings of the Conference on Language Technologies and Digital Humanities*, edited by Špela Arhar Holdt and Tomaž Erjavec, 23–46. Ljubljana: Institute of Contemporary History, 2024.

Babić, Karlo, Milan Petrović, Slobodan Beliga, et al. "Characterisation of COVID-19-Related Tweets in the Croatian Language: Framework Based on the Cro-CoV-cseBERT Model." *Applied Sciences* 11 (21) (2021). https://www.mdpi.com/2076-3417/11/21/10442. doi:10.3390/app112110442.

Baisa, Vít, Marek Blahuš, Michal Cukr, et al. "Automating Dictionary Production: A Tagalog-English-Korean Dictionary from Scratch." In *Electronic Lexicography in the 21st Century (eLex 2019): Smart Lexicography, Conference Proceedings,* edited by Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, et al. Sintra, Portugal, 1-3 October 2019, 805–18. Brno: Lexical Computing, 2019.

Brown, Tom B., Benjamin Mann, Nick Ryder, et al. "Language Models Are Few-Shot Learners." In *Advances in Neural Information Processing Systems,* edited by Hugo, Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, et al. Vol. 33, 1877–1901. Curran Associates, Inc, 2020.

De Schryver, Gilles-Maurice. "Generative AI and Lexicography: The Current State of the Art Using ChatGPT." *International Journal of Lexicography* 36 (4) (2023): 355–87. https://doi.org/10.1093/ijl/ecad021.

Filipović Petrović, Ivana. 2018. *Kada se sretnu leksikografija i frazeologija: O statusu frazema u rječniku*. Zagreb: Srednja Europa.

Filipović Petrović, Ivana, Miguel López Otal, and Slobodan Beliga. "Croatian Idioms Integration: Enhancing the LIdioms Multilingual Linked Idioms Dataset." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024), edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, et al. 4106–12. Torino, Italia: ELRA and ICCL. 2024. https://aclanthology.org/2024.lrec-main.366.

Filipović Petrović, Ivana, and Jelena Parizoska. "Konceptualna Organizacija Frazeoloških Rječnika u Leksikografiji." *Filologija* 73 (2019): 27–45.

Fuertes-Olivera, Pedro. "Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes." *Lexikos* 34 (1) (2024): 123–40. https://lexikos.journals.ac.za/pub/article/view/1883. doi:10.5788/34-1-1883.

Geeraerts, Dirk. "Principles of Monolingual Lexicography." In *Wörterbücher. Ein Internationales Handbuch Zur Lexikographie*, edited by Franz Josef Hausmann, Vol. 1, 287–96. Berlin: Walter de Gruyter, 1989.

Hargraves, Orin. "Information Retrieval for Lexicographic Purposes." In *The Routledge Handbook of Lexicography*, edited by Pedro Fuertes-Olivera, 701–14. Routledge, 2018.

Jakubíček, Miloš, Vojtech Kovář, and Pavel Rychlý. "Million-Click Dictionary: Tools and Methods for Automatic Dictionary Drafting and Post-Editing." In *Book of Abstracts of the 19th EURALEX International Congress*, 65–67, 2021.

Kilgarriff, Adam, Vojtěch Baisa, Jan Bušta et al. "The Sketch Engine: Ten Years On." *Lexicography* 1 (1) (2014): 7–36. https://doi.org/10.1007/s40607-014-0009-9.

Kong, Aobo, Shiwan Zhao, Hao Chen et al. "Better Zero-Shot Reasoning with Role-Play Prompting." In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4099–113. Mexico City, Mexico: Association for Computational Linguistics, 2024. https://aclanthology.org/2024.naacl-long.228/.

Kosem, Iztok, Polona Gantar, Nataša Logar et al. "Automation of Lexicographic Work Using General and Specialized Corpora: Two Case Studies." In *Proceedings of the 16th EURALEX*

International Congress, edited by Andrea Abel, Chiara Vettori and Natascia Ralli, 355–64. Bolzano, Italy: EURAC Research, 2014.

Krishna, Satyapriya, Chirag Agarwal, and Himabindu Lakkaraju. "Understanding the Effects of Iterative Prompting on Truthfulness." In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, 1024, 1–20. Vienna, Austria: JMLR.org, 2024.

Lew, Robert. "ChatGPT as a COBUILD Lexicographer." *Humanities and Social Sciences Communications* 10 (704) (2023). doi:10.1057/s41599-023-02119-6.

Ljubešić, Nikola, and Taja Kuzman. "CLASSLA-Web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024), edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, et al., 3271–82. Torino, Italia: ELRA and ICCL, 2024. https://aclanthology.org/2024.lrec-main.291.

Ljubešić, Nikola, and Davor Lauc. "BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin, and Serbian." In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, edited by Bogdan Babych et al., 37–42. Kiyv, Ukraine: Association for Computational Linguistics, 2021. https://aclanthology.org/2021.bsnlp-1.5.

Madaan, Aman, Niket Tandon, Prakhar et al. "SELF-REFINE: Iterative Refinement with Self-Feedback." In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023), 2019:1–61.* Red Hook, NY: Curran Associates Inc., 2023. https://selfrefine.info/.

McArthur, Tom. *Worlds of Reference: Lexicography, Learning, and Language from the Clay Tablet to the Computer*. Cambridge: Cambridge University Press, 1986.

Miller, Julia. "Research in the Pipeline: Where Lexicography and Phraseology Meet." *Lexicography ASIALEX* 5, No. 1 (2018): 23–33. doi:10.1007/s40607-018-0044-z.

Moussallem, Diego, Mohamed Sherif, Diego Esteves, et al. "LIdioms: A Multilingual Linked Idioms Data Set." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), edited by Nicoletta Calzolari, et al., Miyazaki, Japan: European Language Resources Association (ELRA), 2018. https://aclanthology.org/L18-1392.

Perak, Benedikt, Slobodan Beliga, and Ana Meštrović. "Incorporating Dialect Understanding into LLM Using RAG and Prompt Engineering Techniques for Causal Common-Sense Reasoning." In *Proceedings of the 11th Workshop on NLP for Similar Languages, Varieties, and Dialects* (VarDial 2024), edited by Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, et al., 220–29. Mexico City, Mexico: ACL, 2024. https://aclanthology.org/2024.vardial-1.19. doi:10.18653/v1/2024.vardial-1.19.

Rehm, Georg, and Andy Way. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Springer Nature, 2023. https://doi.org/10.1007/978-3-031-28819-7. doi:10.1007/978-3-031-28819-7.

Rundell, Michael. "Automating the Creation of Dictionaries: Are We Nearly There?" In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography* (ASIALEX 2023 Proceedings), 1–9. Seoul, Korea: Yonsei University, 22–24 June 2023.

Tran, Hanh Thi Hong, Vid Podpečan, Mateja Jemec Tomazin, et al. "Definition Extraction for Slovene: Patterns, Transformer Classifiers, and ChatGPT." In *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*, edited by Marek Medveď, Michal Měchura, Carole Tiberius, et al., 19–38. Brno: Lexical Computing, 2023.

Ulčar, Matej, and Marko Robnik-Šikonja. "FinEst BERT and CroSloEngual BERT: Less Is More in Multilingual Models." In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings,* 104–11. Berlin, Heidelberg: Springer-Verlag, 2020. https://doi.org/10.1007/978-3-030-58323-1_11.

Škorić, Mihailo. "Novi jezički modeli za Srpski jezik." *Infoteka,* 24, 2024. https://arxiv.org/abs/2402.14379.

Online sources

Clark, Kevin, Minh-Thang Luong, Quoc V. Le, et al. "ELECTRA: Pre-training Text Encoders as Discriminators Rather than Generators." *ICLR,* 2020. https://openreview.net/pdf?id=r1xMH1BtvB.

Davies, Mark. *AI/LLM Integration with the Corpora from English-Corpora.org,* 2025. https://www.english-corpora.org/ai-llms/corpora-vs-llms.html.

Filipović Petrović, Ivana, and Jelena Parizoska. *Frazeološki rječnik hrvatskoga jezika v2*. Zagreb: Hrvatska akademija znanosti i umjetnosti, 2023. https://lexonomy.elex.is/#/frazeoloskirjecnikhr.

Ljubešić, Nikola, and Filip Klubička. *Croatian Web Corpus hrWaC 2.1,* 2016. http://hdl.handle.net/11356/1064. (Slovenian language resource repository CLARIN.SI).

Ljubešić, Nikola, Peter Rupnik, and Taja Kuzman. *Croatian Web Corpus CLASSLA-Web.hr 1.0,* 2024. http://hdl.handle.net/11356/1929. (Slovenian language resource repository CLARIN.SI).

Madaan, Aman, Niket Tandon, Prakhar Gupta, et al. *Self-Refine: Iterative Refinement with Self-Feedback,* 2023. arXiv. https://arxiv.org/abs/2303.17651.

Vossen, Piek. "ChatGPT Is a Waste of Time." *VU-Magazine,* 2022. https://vumagazine.nl/professor-piek-vossen-chatgpt-is-a-waste-of-time?lang=en.

Ivana Filipović Petrović, Slobodan Beliga

ALI LAHKO UMETNA INTELIGENCA RAZUME HRVAŠKE IDOME? OCENA VELIKIH JEZIKOVNIH MODELOV PRI LEKSIKOGRAFSKIH NALOGAH

POVZETEK

Prispevek je posvečen preučevanju možnosti uporabe velikih jezikovnih modelov, zlasti modelov GPT podjetja OpenAI, pri leksikografskem delu na področju hrvaških idiomatičnih izrazov. Študija se osredotoča na spletni *Frazeološki slovar hrvaškega jezika* in ocenjuje, kako zmogljivi so veliki jezikovni modeli pri (1) identifikaciji pomenskih ekvivalentov med idiomi ter (2) ustvarjanju konceptualnih kategorij (pomenskih polj) ter razvrščanju idiomov vanje. Končni cilj je zmanjšati količino ročnega dela pri leksikografskih delovnih postopkih ter hkrati ohraniti kakovost in zanesljivost. Raziskava se začne s primerjalnim vrednotenjem treh hrvaških odprtokodnih velikih jezikovnih modelov (Cro-CoV-cseBERT, BERTić in gpt2-vrabac) na podlagi njihove uspešnosti pri razvrščanju idiomov po semantični podobnosti. Ti modeli so kljub učenju na korpusih v hrvaškem jeziku slabše opravili naloge, ki so vključevale idiomatične pomene. Njihove omejitve so posledica neustreznih podatkov za učenje, ki ne vsebujejo dovolj idiomov, in nezmožnosti zajemanja figurativnih pomenov, ki so po svoji naravi zapleteni in odvisni od konteksta. Pri naslednjih poskusih z modeloma GPT-3.5-turbo in GPT-4o so bili doseženi bistveno boljši rezultati pri nalogah semantične podobnosti in kategorizacije. Z uporabo izpopolnjenega

inženiringa pozivov (*prompt engineering*) – predvsem načina pozivanja z igranjem vlog (*role-play prompting*) – so modeli GPT dosegli visoko stopnjo ujemanja s človeškim razvrščanjem idiomov v pomenska polja. Model GPT-4o je na primer dosegel 98-odstotno ujemanje s človeškim razvrščanjem pri nalogi razvrščanja idiomov v vnaprej določene kategorije, kot so PRIJAZNOST, NOROST in KONFLIKT. Druga glavna naloga je preverjala zmožnost modelov GPT, da samostojno razvrščajo idiome po pomenu in pomenskim poljem dodeljujejo ustrezna imena. Rezultati prvih poskusov z modelom GPT-3,5 turbo so vsebovali nedosledne in preveč specifične kategorije. Model GPT-4o pa je z izboljšanim pozivom, ki je poudarjal pojmovno skladnost in spodbujal razvrščanje v skupine po skupnem pomenu, ustvaril leksikografsko uporabne in dobro strukturirane skupine. To kaže na uspešno uporabo izpopolnjenega pozivanja – metode, pri kateri je v proces vključen človek (*human-in-the-loop*) in pri kateri se pozivi vedno znova spreminjajo na podlagi analize rezultatov modela. Kljub obetavnim rezultatom ostaja nerešenih več izzivov, saj modeli občasno napačno razvrščajo idiome zaradi zanašanja na dobesedne pomene in nepoznavanja kulturnega konteksta. Poleg tega je konceptualna organizacija idiomov še vedno subjektivna naloga, pri kateri je potrebna strokovna človeška presoja. Študija tako poudarja pomen hibridnih delovnih procesov, pri katerih so veliki jezikovni modeli leksikografom v pomoč, vendar jih ne nadomestijo. Ugotovitve prispevajo k širši razpravi o umetni inteligenci v leksikografiji, zlasti pri jezikih z nezadostnimi viri, kot je hrvaščina. Študija je pokazala, da je za čim večjo učinkovitost in kakovost razvoja slovarjev s pomočjo umetne inteligence priporočljivo natančno prilagoditi velike jezikovne modele korpusom z veliko idiomi ter izboljšati strategije oblikovanje pozivov.

Appendix. Full prompt versions used for the research

Prompt for Task one: Linking idioms to predefined semantic fields

Croatian (original):

Preuzmi ulogu leksikografa koji izrađuje novi konceptualno organizirani frazeološki rječnik hrvatskoga jezika. Odgovaraj na hrvatskom jeziku. Zadan je popis unaprijed definiranih semantičkih polja. Poveži zadani frazem s najprikladnijim semantičkim poljem s popisa. Odgovori odabirom samo jednog ponuđenog semantičkog polja.

English (translation):

Take on the role of a lexicographer creating a new conceptually organized phraseological dictionary of Croatian. Please respond in Croatian. A list of predefined semantic fields is provided. Link the given idiom to the most appropriate semantic field from the list. Respond by selecting only one of the offered semantic fields.

Prompt for Task two (first version): Grouping of idioms into concepts
Croatian (original):
Poslat ću ti popis hrvatskih frazema. Trebaš ih grupirati po značenju u semantička polja i ta polja imenovati na hrvatskom jeziku.
Pokušaj frazeme sličnog značenja grupirati zajedno i dodijeliti im dovoljno specifičan naziv koji opisuje njihovo značenje, npr. sreća, tuga, svađa, prepreka, ljubav itd.

English (translation):
I will send you a list of Croatian idioms. You need to group them by meaning into semantic fields and give those fields names in Croatian.
Try to group idioms with similar meanings together and assign them a sufficiently specific name that describes their meaning, such as happiness, sadness, quarrel, obstacle, love, etc.

Prompt for Task two (refined version): Specific grouping with emphasis on concept clarity
Croatian (original):
Pokušaj grupirati što više frazema pod isto semantičko polje ako imaju zajedničko značenje.
Koncepti neka budu konkretni i upotrebljivi, a ne preopćeniti (npr. emocije, osjećaji), osim ako to nije nužno.

English (translation):
Try to group as many idioms as possible under the same semantic field if they share a common meaning.
The concepts should be concrete and useful, rather than overly general (e.g., emotions, feelings), unless generality is necessary.

1.01 DOI: https://doi.org/10.51663/pnz.65.3.10

Matej Klemen*

Poznavanje pogostih splošnih besed v slovenščini med govorci slovenščine kot drugega in tujega jezika

IZVLEČEK

Članek predstavlja dva testa besedišča – da/ne test in pilotni test besedišča po frekvenčnih razredih, s katerima smo med govorci slovenščine kot drugega in tujega jezika preverjali poznavanje pogostih splošnih besed v slovenščini. Povzete so ugotovitve prve izvedbe da/ne testa na Mladinski poletni šoli slovenščine 2022, članek pa se osredotoča na drugo izvedbo med odraslimi tečajniki Centra za slovenščino kot drugi in tuji jezik (Filozofska fakulteta Univerze v Ljubljani) leta 2024. Cilj je bil preveriti, ali lahko da/ne test učinkovito razvršča govorce slovenščine kot drugega in tujega jezika glede na njihovo jezikovno znanje, in ugotoviti njegovo zanesljivost. Rezultati kažejo, da govorci slovenščine kot drugega in tujega jezika v večji meri poznajo bolj pogoste besede kot manj pogoste in da se poznavanje besedišča razlikuje glede na raven jezikovnega znanja. Z da/ne testom lahko dobro ločimo med govorci, ki so v slovenščini začetniki, nadaljevalci ali izpopolnjevalci, tistih, ki so na prehodu med temi ravnmi, pa s tem testom ne moremo zanesljivo uvrstiti. Rezultati so pokazali tudi, da so govorci slovanskih jezikov pri nižjih ravneh znanja dosegli boljše rezultate kot govorci neslovanskih jezikov. Pilotni test besedišča po frekvenčnih razredih potrjuje veljavnost da/ne testa, saj se njuni rezultati močno ujemajo.

Ključne besede: besedišče, da/ne test, test besedišča po frekvenčnih razredih, slovenščina kot drugi in tuji jezik

* **Lekt., Univerza v Ljubljani, Filozofska fakulteta, Center za slovenščino kot drugi in tuji jezik, Aškerčeva 2, SI-1000 Ljubljana, matej.klemen@ff.uni-lj.si; ORCID: 0009-0006-5087-9051**

ABSTRACT

KNOWLEDGE OF COMMON WORDS IN SLOVENIAN AMONG SPEAKERS OF SLOVENIAN AS A SECOND AND FOREIGN LANGUAGE

The article presents two vocabulary tests – a yes/no test and a pilot vocabulary levels test – used to evaluate the knowledge of frequently used common words in Slovenian among speakers of Slovenian as a second and foreign language. It summarises the findings from the first administration of the yes/no test at the 2022 Youth Summer School of Slovenian and focuses on the second administration involving adult learners at the Centre for Slovene as a Second and Foreign Language (Faculty of Arts, University of Ljubljana), in 2024. The study aimed to determine whether the yes/no test reliably classifies speakers of Slovenian as a second and foreign language according to their language proficiency and to assess its reliability. The results show that speakers of Slovenian as a second and foreign language are more familiar with high-frequency words than with low-frequency ones, and that vocabulary knowledge varies in relation to proficiency level. The yes/no test successfully differentiates between beginner, intermediate, and advanced learners of Slovenian, although it is less dependable when classifying speakers transitioning between these levels. The results also reveal that speakers of Slavic languages perform better than non-Slavic speakers at lower proficiency levels. The pilot vocabulary levels test supports the validity of the yes/no test, as the results of both tests show a strong correlation.

Keywords: vocabulary, yes/no test, vocabulary levels test, Slovenian as a second and foreign language

Uvod

Prispevek predstavlja dva testa besedišča, po tujih zgledih razvita za slovenščino, in sicer t. i. da/ne test in pilotni test besedišča po frekvenčnih razredih, s katerima smo med govorci slovenščine kot drugega in tujega jezika (SDTJ) preverjali poznavanje pogostih splošnih besed v slovenščini. Ker je bila prva uporaba da/ne testa med udeleženci Mladinske poletne šole (MPŠ) 2022 že podrobneje predstavljena v zborniku konference Jezikovne tehnologije in digitalna humanistika,[1] so v pričujočem prispevku rezultati in ugotovitve te izvedbe le povzeti. Osredotočamo pa se na drugo izvedbo testa, ki je bila v dveh ponovitvah izvedena med udeleženci tečajev za odrasle, ki so se slovenščino učili na Centru za slovenščino kot drugi in tuji jezik[2] (CSDTJ, Filozofska fakulteta Univerze v Ljubljani) v spomladanskem semestru leta 2024.

1 Matej Klemen, »Test poznavanja splošnih besed v slovenščini med udeleženci Mladinske poletne šole slovenščine,« v: Špela Arhar Holdt in Tomaž Erjavec, ur., *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (Ljubljana: Inštitut za novejšo zgodovino, 2024), 604–20, pridobljeno 3. 12. 2024, https://doi.org/10.5281/zenodo.13936445.

2 *Center za slovenščino kot drugi in tuji jezik*, https://centerslo.si/.

Podobno kot Read[3] smo se spraševali o dveh vidikih testa: o njegovi uporabnosti za razvrščanje tečajnikov v skupine, homogene po jezikovnem znanju, in kaj tak test pove o poznavanju besedišča. Razvijalci testa za angleščino so ugotovili, da je mogoče z da/ne testom dobro napovedati, kako učeče se uvrstiti v ustrezne skupine.[4] Enako se je ob prvi izvedbi da/ne testa za slovenščino med udeleženci MPŠ[5] pokazalo, da je mogoče na podlagi njegovih rezultatov udeležence razvrstiti po jezikovnem znanju. Pričujoči prispevek v primerjavi s prej omenjenim prinaša nove podatke in ugotovitve, pridobljene na testiranju z da/ne testom pri drugačni publiki, ki je starostno in jezikovno bolj raznolika. V testiranje so bili vključeni tudi popolni začetniki, prav tako pa smo ugotavljali, kako test rešujejo govorci slovenščini sorodnih jezikov. Prispevek predstavlja tudi nov test poznavanja splošnega besedišča, s katerim smo želeli preveriti zanesljivost odgovorov pri da/ne testu. Ob drugi izvedbi da/ne testa je bil za slovenščino namreč prvič preizkušen tudi test besedišča po frekvenčnih razredih, pripravljen po vzoru Nationovega testa Vocabulary Levels Test.[6]

Prispevek najprej predstavi zglede za pripravo obeh testov in postopek njune prilagoditve za slovenščino. V nadaljevanju so predstavljeni hipoteze, potek testiranja in rezultati. Ugotovitve so komentirane v naslednjem poglavju, prispevek pa se sklene z razmislekom o informativnosti tovrstnega testiranja in omejitvah raziskave.

Da/ne test poznavanja splošnih besed v slovenščini

Zgledi za pripravo testa

Pri da/ne testu (angl. *yes/no test*) se testirani za besede v tujem jeziku, ki so jim predstavljene brez konteksta, odločajo, ali jih poznajo ali ne. Tak test je mogoče enostavno izvesti, zlasti če je pripravljen v digitalni obliki, in tudi hitro rešiti. Neposredna zgleda za pripravo da/ne testa za slovenščino sta bila test V_YesNo[7] za angleščino in test obsega slovarja za grščino, kot sta ga pripravila Milton in Alexiou.[8] Oba izhajata iz testov, ki jih je Meara s sodelavci razvil kot alternativo klasičnim uvrstitvenim testom na tečajih angleščine.[9]

3 John Read, *Assessing Vocabulary* (Cambridge: Cambridge University Press, 2000), 127–32.

4 Paul Meara in Glyn Jones, »Vocabulary Size as a Placement Indicator,« v: Pamela Grunwell, ur., *Applied Linguistics in Society* (London: Centre for Information on Language Teaching and Research, 1988), 80–87, pridobljeno 9. 3. 2024, https://www.lognostics.co.uk/vlibrary/meara&jones1988.pdf.

5 Klemen, »Test poznavanja splošnih besed.«

6 I. S. P. Nation, »Testing and Teaching Vocabulary,« *Guidelines* 5, št. 1 (1983): 12–25.

7 Paul Meara in Imma Miralpeix, »V_YesNo v1.0« v: *Tools for Researching Vocabulary* (Bristol, Blue Ridge Summit: Multilingual Matters, 2016), 113–33, pridobljeno 9. 3. 2024, https://doi.org/10.21832/9781783096473.

8 James Milton in Thomaï Alexiou, »Developing a vocabulary size test in Greek as a foreign language,« v: Angeliki Psaltou - Joycey in Marina Mattheoudakis, ur., *Advances in Research on Language Acquisition* (Thessaloniki: Greek Applied Linguistcs Association, 2010), 307–18.

9 Paul Meara in Barbara Buxton, »An alternative to multiple choice vocabulary tests,« *Language Testing* 4, št. 2 (1987): 142–54, pridobljeno 22. 2. 2025, https://doi.org/10.1177/026553228700400202. Meara in Jones, »Vocabulary Size as a Placement Indicator.«

V_YesNo je digitalni test,[10] ki preverja poznavanje 10.000 najpogostejših besed v angleščini. Test obsega 100 v angleščini obstoječih besed (po 10 besed za vsakih 1000) in 100 v angleščini neobstoječih besed (nebesed), kar naj bi preprečevalo goljufanje oziroma kaznovalo ugibanje. Testiranemu se besede in nebesede prikazujejo posamično, ob koncu testa pa se izpiše rezultat, ki je ocena, koliko besed testirani pozna. Doseči je mogoče največ 10.000 točk. Če testirani označi, da pozna nebesedo, se njegov končni rezultat zniža, odbitek točk pa je odvisen od tega, kako testirani odgovarja na vprašanja s pravimi besedami. Testirani, ki pravilno prepoznajo večino pravih besed, so za posamezno napačno prepoznavo nebesede kaznovani blažje, medtem ko so tisti, ki pravilno prepoznajo le nekaj pravih besed, za napačno prepoznavo nebesed kaznovani strožje.[11]

Da/ne test, s katerim sta Milton in Alexiou želela oceniti obseg besedišča pri učencih grščine, je podoben testu V_YesNo. Besede zajema izmed prvih 5000 najpogostejših lem v grškem nacionalnem korpusu (Hellenic National Corpus), in sicer po 20 za vsakih 1000 besed, skupaj 100, tem pa je dodanih še 20 v grščini neobstoječih besed. Besede so testiranemu na papirju predstavljene v seznamu, brez besedilnega konteksta, odločiti pa se mora, ali jo »pozna ali zna uporabiti«.[12] Če testirani označi, da obstoječo besedo pozna, se njegov odgovor točkuje s 50 točkami, najvišje število točk je 5000. Če pa testirani označi, da pozna nebesedo, se od končnega rezultata za vsak tak odgovor odšteje 250 točk, izgubi lahko torej največ 5000 točk. Končno število točk naj bi predstavljalo oceno, kolikšen je obseg besedišča pri testiranem.

Priprava testa

Pri pripravi da/ne testa za slovenščino smo sledili načelom omenjenih testov. Besede so bile zajete iz Referenčnega seznama pogostih splošnih besed za slovenščino.[13] V tem seznamu, ki je nastal s prekrivanjem najpogostejših lem iz štirih slovenskih besedilnih korpusov (Kres, GOS, Janes, Šolar 2.0), je zbranih 4768 pogostih splošnih lem.[14] Za pripravo testa je bilo upoštevanih prvih 4000 lem, saj zadnja tisočica lem v seznamu ni popolna. V želji po kar najbolj enakomernem izboru je bila izmed vsakih zaporednih 50 lem naključno izbrana ena. Za vsakih zaporednih 1000 lem je bilo naključno izbranih 20 lem, skupaj 80. Pri preverjanju izbora so bile leme, ki lahko pripadajo različnim besednim vrstam (npr. *raven*, ki je lahko samostalnik ali pridevnik), nadomeščene z drugimi iz istega frekvenčnega razreda. Pri izboru se nismo omejevali, temveč smo vključili tudi besedne vrste, ki v nekaterih testih besedišča niso prisotne

10 Test je dostopen na *V_YesNo v1.1*, https://www.lognostics.co.uk/tools/V_YesNo/V_YesNo.htm.

11 Meara in Miralpeix, »V_YesNo v1.0.«

12 Milton in Alexiou, »Developing a vocabulary size test in Greek as a foreign language,« 318.

13 Senja Pollak et al., *Reference List of Slovene Frequent Common Words*, Slovenian language resource repository CLARIN.SI, 2020, http://hdl.handle.net/11356/1346.

14 Špela Arhar Holdt et al., »Referenčni seznam pogostih splošnih besed za slovenščino,« v: Darja Fišer in Tomaž Erjavec, ur., *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (Ljubljana: Inštitut za novejšo zgodovino, 2020), 10–15, pridobljeno 13. 5. 2021, http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Arhar-Holdt-et-al_Referencni-seznam-pogostih-splosnih-besed-za-slovenscino.pdf.

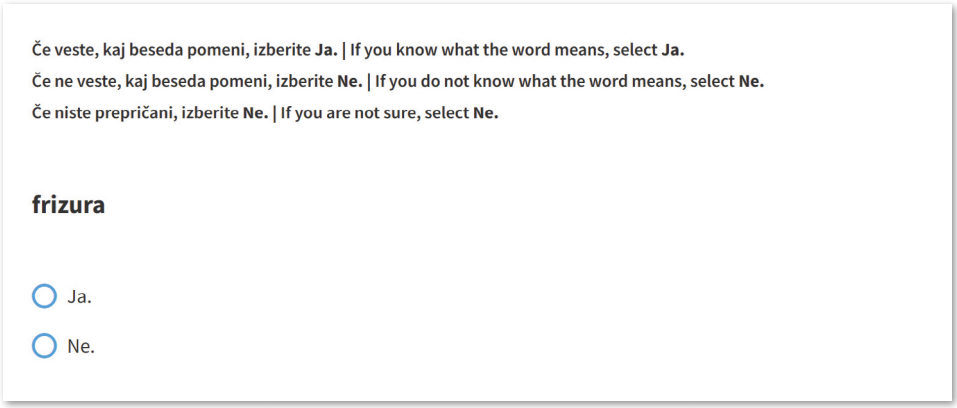
(npr. veznik *kajti*, prislovi *tako*, *večinoma* in *podobno*, ki so v izboru za prvih 1000 najpogostejših splošnih besed). Poleg izbranih besed so bile v test vključene tudi v slovenščini neobstoječe besede (npr. *posminati*, *čembrita*, *izpontivanje*, *deptanjski*), ki delujejo kot slovenske in bi lahko pripadale različnim besednim vrstam. Pri oblikovanju nebesed smo upoštevali strukturo slovenskega zloga, vanje smo vključili tipična obrazila in končnice. Razmerje med besedami in nebesedami je bilo enako kot v testu, ki sta ga zasnovala Milton in Alexiou – 5 : 1. Z izbranimi 80 obstoječimi besedami in 16 nebesedami test vsebuje 96 vprašanj, v katerih so besede zapisane v osnovni obliki.

Oblika testa

Zaradi enostavnejše izvedbe testiranja, zbiranja in analize odgovorov smo se odločili za digitalno obliko testa. Pri prvi izvedbi testa med udeleženci MPŠ je bilo uporabljeno orodje Socrative, ki ni omogočalo zbiranja dodatnih podatkov, zato je bila pri drugi izvedbi izbrana platforma 1ka.[15]

Test je sestavljal uvodni nagovor, ki je testiranim pojasnil namen testiranja, sledilo je vprašanje o strinjanju z zbiranjem osebnih podatkov, ki so bili potrebni za identifikacijo oseb, sodelujočih v dveh ponovitvah testa, zatem se je testiranim pokazalo navodilo za reševanje testa: kako bodo vprašanja strukturirana, kdaj naj izberejo gumb *Ja* in kdaj *Ne*. V navodilu so bili testiranci opozorjeni, da bodo videli v slovenščini neobstoječe besede, naj ne ugibajo in ne uporabljajo pomoči. Sledilo je 96 vprašanj, pri vsakem se je testiranemu prikazalo enako navodilo (Slika 1).[16] Vsi nagovori, navodila in vprašanja so bili napisani v slovenščini in angleščini.

Prikaz vprašanja v da/ne testu na računalniškem zaslonu



Vir: lastno delo

15 *1KA | Spletne ankete*, https://1ka.arnes.si/.

16 Ker smo želeli rezultate primerjati s testiranjem na MPŠ 2022 (Klemen, »Test poznavanja splošnih besed«), v test ni bil dodan dodaten gumb *Ne vem* ali *Nisem prepričan/a*, s katerim bi bilo mogoče znižati stopnjo ugibanja pri odgovarjanju. Xian Zhang, »The I Don't Know Option in the Vocabulary Size Test,« *TESOL Quarterly* 47, št. 4 (2013): 790–811, pridobljeno 2. 5. 2024, https://doi.org/10.1002/tesq.98.

Zaradi omejitev orodja 1ka vprašanj ni bilo mogoče prikazovati popolnoma naključno. Določena mera naključnosti je bila dosežena z uporabo treh različnih razporeditev vprašanj, pri čemer so bile besede iz različnih frekvenčnih razredov pomešane med seboj, kot v zvezi z motivacijo za reševanje pri enem od testov besedišča priporoča Nation.[17] Pri tem smo upoštevali, 1) da se test začne in konča z v slovenščini obstoječo besedo, 2) da so besede iz različnih frekvenčnih razredov razporejene po celotnem testu, 3) da se zaporedno ne pojavita besedi iz istega frekvenčnega razreda in 4) da sta med dvema nebesedama vedno vsaj dve obstoječi besedi. Testiranemu se je naključno prikazala ena od treh razporeditev vprašanj.

Točkovanje

Pri da/ne testu je bilo za vsak odgovor *Ja* pri obstoječih besedah pripisanih 50 točk. Vsaka lema v da/ne testu namreč predstavlja 50 lem, in če testirani odgovori, da jo pozna, predvidevamo, da pozna tudi preostale v njeni frekvenčni okolici.[18] Vsak odgovor, da testirani besede ne pozna (*Ne*), je dobil 0 točk. Vsaka napačna prepoznava nebesede je bila kaznovana z –250 točkami. Testirani je lahko torej s pravilnimi odgovori zbral največ 4000 točk, z napačno prepoznavo nebesed pa jih je lahko enako število izgubil.

Izračunano je bilo skupno število točk za vsakih tisoč besed oziroma vsak frekvenčni razred, skupno število točk za prave besede (v nadaljevanju *skupno število točk*), skupno število točk za nebesede in skupno korigirano število točk (od točk za obstoječe besede so bile odštete točke za nebesede, ki so bile prepoznane kot obstoječe besede; v nadaljevanju *korigirane točke*). Morebitni manjkajoči odgovori niso bili upoštevani, zanje točke niso bile prištete ali odštete.

Test receptivnega besedišča po frekvenčnih razredih

Zgled za pripravo testa

Pri pripravi testa receptivnega besedišča, torej besedišča, ki ga govorci SDTJ prepoznajo in razumejo, ne znajo pa ga nujno tudi uporabljati,[19] smo izhajali iz testa Vocabulary Levels Test (VLT), ki ga je za angleščino kot tuji jezik razvil Nation,[20]

17 Paul Nation, »The Vocabulary Size Test,« 2012, pridobljeno 21. 2. 2024, https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-information-and-specifications.pdf.

18 Philip Durrant et al., *Research Methods in Vocabulary Studies* (John Benjamins Publishing Company: 2022), 157, 158.

19 Prim. I. S. P. Nation, *Learning Vocabulary in Another Language* (Cambridge, Cambridge University Press: 2022), 52–55.

20 Nation, »Testing and Teaching Vocabulary.«

kasneje pa so ga dopolnjevali in posodabljali.[21] Gre za test, s katerim se preverja poznavanje receptivnega besedišča, in sicer samostalnikov, glagolov in pridevnikov iz različnih frekvenčnih razredov. Prve različice VLT[22] preverjajo poznavanje besed izmed najpogostejših 2000, 3000, 5000 in 10.000 besednih družin v angleščini in posebej iz nabora akademskega besedišča; verzija, ki jo je pripravil Webb s sodelavci,[23] pa izmed 1000, 2000, 3000, 4000 in 5000 besednih družin.

Vsako vprašanje sestavlja šest besed v osnovni obliki in tri definicije, testirani pa mora povezati besedo in ustrezno definicijo. V prvotni različici[24] je bilo za vsak frekvenčni razred šest vprašanj, torej 18 testiranih besed in 18 besed, ki so bile distraktorji. V različicah, ki so jih pripravili Schmitt s sodelavci[25] in Webb s sodelavci,[26] pa je bilo po deset vprašanj. VLT je bil zamišljen kot diagnostično orodje, ki naj bi testiranega usmerjalo, katero besedišče naj se uči – besedišče tistega frekvenčnega razreda, pri katerem je dosegel slabši rezultat,[27] uporablja pa se tudi za oceno obsega slovarja govorcev angleščine kot tujega jezika.[28]

Priprava testa

Za testiranje je bila pripravljena slovenska različica VLT, v tem besedilu imenovana *pilotni test besedišča po frekvenčnih razredih* oziroma piTeBeFRa. Namen ni bil, kot pri VLT, oceniti, koliko besed testirani poznajo v posameznem frekvenčnem razredu, temveč smo želeli preveriti skladnost odgovorov testiranih v da/ne testu in piTeBeFRa. Zato ni bil pripravljen »popoln« test, ki bi zajemal enako število besed iz posameznih frekvenčnih razredov, temveč so bile vanj vključene le besede, že prisotne v da/ne testu. Ker smo izhajali iz seznama pogostih splošnih besed za slovenščino,[29] je bila kot enota uporabljena lema in ne besedna družina[30] kot v prej omenjenih različicah VLT.

Izmed 80 slovenskih besed, vključenih v da/ne test, je bilo za piTeBeFRa izbranih 42 besed: 30 samostalnikov, devet glagolov in trije pridevniki.[31] Vsako vprašanje je vsebovalo tri besede iste besedne vrste iz istega frekvenčnega razreda (npr. *vodja*,

21 Norbert Schmitt, Diane Schmitt in Caroline Clapham, »Developing and Exploring the Behaviour of Two New Versions of the Vocabulary Levels Test,« *Language Testing* 18, št. 1 (2001): 55–88. Stuart Webb, Yosuke Sasao in Oliver Ballance, »The Updated Vocabulary Levels Test,« ITL – *International Journal of Applied Linguistics* 168, št. 1 (2017): 33–69, pridobljeno 13. 5. 2021, https://doi.org/10.1075/itl.168.1.02web.

22 Nation, »Testing and Teaching Vocabulary.« Schmitt, Schmitt in Clapham, »Developing and Exploring.«

23 Webb, Sasao in Ballance, »The Updated Vocabulary Levels Test.«

24 Nation, »Testing and Teaching Vocabulary.«

25 Schmitt, Schmitt in Clapham, »Developing and Exploring.«

26 Webb, Sasao in Ballance, »The Updated Vocabulary Levels Test.«

27 Nation, »Testing and Teaching Vocabulary,« 15.

28 Read, *Assessing Vocabulary*, 118.

29 Pollak et al., *Reference List*.

30 Za argumente o izboru najprimernejše enote za test gl. npr. Durrant et al., *Research Methods in Vocabulary Studies*, 158, 159.

31 Razmerje med njimi je torej 10 : 3 : 1, kar odstopa od razmerja, uporabljenega v VLT (5 : 2 : 1), ki naj bi odražalo razmerje med temi besednimi vrstami v angleščini (Webb, Sasao in Ballance, »The Updated Vocabulary Levels Test,« 34). Če bi želeli slediti razmerju med temi besednimi vrstami v slovenščini, bi bilo to glede na Referenčni seznam pogostih splošnih besed za slovenščino (Pollak idr., Reference List) približno 2 : 1 : 1.

značilnost, *dvorana*) ter tri distraktorje, ki so bili prav tako besede iste besedne vrste in se v Referenčnem seznamu pogostih splošnih besed pojavljajo do deset mest višje ali nižje ob posamezni besedi (npr. *vojak*, *storitev*, *sorodnik*), torej s podobno pogostostjo kot testirane besede. Pri izbiri besed smo upoštevali, da so bile pomensko dovolj različne, da so lahko testirani izbrali ustrezne odgovore.[32] Za testirane besede so bile oblikovane definicije. V čim večji meri smo se želeli izogniti temu, da bi piTeBeFRa preverjal tudi razumevanje definicij, zato smo jih pripravili tako, da so bile 1) čim krajše, da testiranim ni bilo treba veliko brati; 2) kadar je bilo mogoče, sestavljene iz besed na ravni A1[33] (npr. *črn – temne barve*) ali iz transparentnih besed (npr. *obzorje – horizont*);[34] 3) oblikovane tako, da so se izogibale uporabi besede z istim korenom (npr. *bolnik – ~~oseba, ki je bolna~~ – oseba, ki ni zdrava*), kar bi lahko omogočalo večje ugibanje pri odgovarjanju.[35] Definicije so razlagale pomen, naveden kot prvi v *Slovarju slovenskega knjižnega jezika*, pri čemer pa tega pomena niso nujno opisovale v celoti ali natančno (npr. *frizura – lasje*). Skupno je bilo pripravljenih 14 vprašanj: po tri za frekvenčna razreda 0–1000 in 3001–4000 ter po štiri za frekvenčna razreda 1001–2000 in 2001–3000.

Oblika testa

Test je bil prav tako naložen v orodje 1ka in izveden skupaj z drugo ponovitvijo da/ne testa. Ker je PiTeBeFRa sledil da/ne testu, se uvodne informacije niso ponavljale, temveč se je testiranim prikazala stran z navodilom za reševanje testa, ki je pojasnjevalo strukturo vprašanj in vključevalo grafični primer rešenega vprašanja. Testirane smo prosili, naj ne uporabljajo pomoči. Sledilo je 14 vprašanj, pri vsakem se je testiranemu prikazalo enako navodilo. Vsi nagovori, navodila in vprašanja so bili v slovenščini in angleščini.

V starejših različicah VLT so bila vprašanja oblikovana tako, da so bile besede in definicije postavljene v dva stolpca: v levem je bilo naštetih šest besed, v desnem pa tri definicije. Webb in sodelavci so obliko vprašanj spremenili, »da bi bila za testirane preglednejša«.[36] Vprašanja so oblikovali tabelarno: besede so prikazali v prvi vrstici, definicije pa navpično v prvem stolpcu. Njihovi postavitvi smo sledili tudi pri pripravi piTeBeFRa (Slika 2). Vprašanja v tem testu se niso prikazovala naključno, temveč glede na frekvenčne razrede od najbolj do manj frekventnih besed.

32 Nation, »Testing and Teaching Vocabulary,« 15. Webb, Sasao in Ballance, »The Updated Vocabulary Levels Test,« 36.

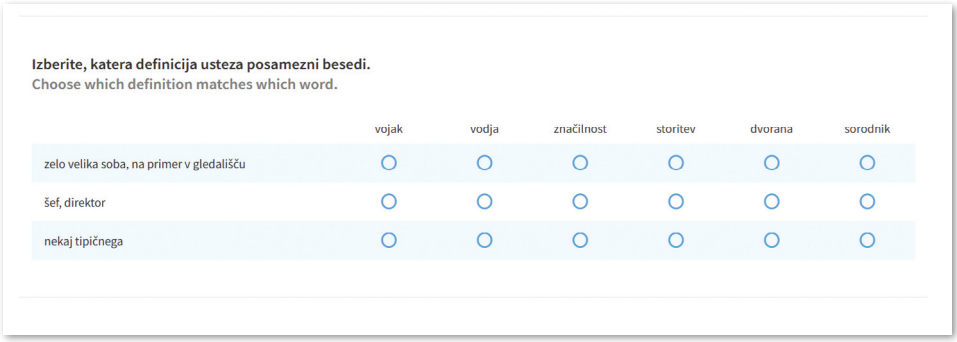
33 Matej Klemen, Špela Arhar Holdt in Senja Pollak, *Core Vocabulary for Slovenian as L2 1.0*, Slovenian Language Resource Repository CLARIN.SI, 2022, pridobljeno 18. 11. 2022, http://hdl.handle.net/11356/1697.

34 Želeli smo uporabiti čim bolj enostavno in čim večjemu številu govorcev SDTJ znano besedišče. V definicijah nismo sledili načelu Nationa (»Testing and Teaching Vocabulary,« 15), da bi besede razlagali z besedami iz višjih frekvenčnih razredov (da bi bile npr. definicije besed, ki sodijo v tretjo tisočico, sestavljene z besedami, ki sodijo v prvo in drugo tisočico).

35 Prim. Schmitt, Schmitt in Clapham, »Developing and Exploring,« 59.

36 Webb, Sasao in Ballance, »The Updated Vocabulary Levels Test,« 37.

Prikaz vprašanja v piTeBeFRa na računalniškem zaslonu



Vir: lastno delo

Točkovanje

Pri piTeBeFRa je bila testiranemu za vsako ustrezno povezano besedo in definicijo dodeljena ena točka, nato pa je bilo izračunano število doseženih točk. Pri napačnih odgovorih se točke niso odštevale, manjkajoči odgovori niso bili upoštevani.

Prva izvedba da/ne testa med udeleženci Mladinske poletne šole slovenščine

Ta razdelek povzema ugotovitve o izvedbi da/ne testa med udeleženci MPŠ leta 2022, ki so že bile obširneje predstavljene.[37] Testiranje z da/ne testom je bilo izvedeno dvakrat: 5. julija 2022 (T1-MPŠ) in 15. julija 2022 (T2-MPŠ), prvi in zadnji dan pouka. V testiranje so bili vključeni učeči se, stari od 13 do 18 let, z različnim jezikovnim znanjem. Bili so začetniki, nadaljevalci in izpopolnjevalci, ki so bili na podlagi »klasičnega« uvrstitvenega testa razvrščeni v skupine od tri do devet. V testiranje niso bili vključeni popolni začetniki v slovenščini.

Rezultati so pokazali, da so testirani v večji meri poznali frekventnejše besede in za vsakih naslednjih tisoč dosegli nekoliko nižji rezultat (Tabela 1). Pri obeh testiranjih se je pokazal padajoči profil poznavanja besedišča, ki se je začel stopničasto spuščati pri poznavanju tretje in četrte tisočice besed. Analize so potrdile, da razlika med prvima dvema frekvenčnima razredoma ni bila statistično značilna ($p_{bonferroni} = 1{,}0$ pri T1-MPŠ oziroma $p_{bonferroni} = 0{,}8$ pri T2-MPŠ), vse druge razlike pa so bile statistično značilne ($p_{bonferroni} < 0{,}05$).

37 Klemen, »Test poznavanja splošnih besed.«

Tabela 1: Število točk, doseženih pri T1-MPŠ in T2-MPŠ, za posameznih tisoč besed

Testiranje	Frekvenčni razred	N	M	Mdn	SD
T1-MPŠ	1–1000	48	794	850	189
	1001–2000	48	789	850	221
	2001–3000	48	728	750	234
	3001–4000	48	646	650	206
T2-MPŠ	1–1000	44	842	875	166
	1001–2000	44	816	900	225
	2001–3000	44	738	800	248
	3001–4000	44	684	700	252

Vir: lastno delo

Po pričakovanjih so se pri T1-MPŠ skupine z različnim jezikovnim znanjem med seboj statistično značilno razlikovale v doseženem rezultatu ($p < 0{,}001$): začetniki so dosegli nižji rezultat kot nadaljevalci, ti pa nižjega kot izpopolnjevalci. Rezultati da/ne testa pri T1-MPŠ so bili močno povezani z rezultati uvrstitvenega testa (za skupne točke: $r = 0{,}78$, $p < 0{,}001$; za korigirane točke $r = 0{,}78$, $p < 0{,}001$). Tudi povezanost med rezultatom da/ne testa pri T1-MPŠ in tem, v katero od skupin (od tri do devet) so bili testirani uvrščeni, se je izkazala za visoko (za skupne točke: $r = 0{,}737$, $p < 0{,}001$; za korigirane točke: $r = 0{,}732$, $p < 0{,}001$) in je bila le nekoliko nižja kot povezanost med številom točk na uvrstitvenem testu in razvrstitvijo v skupino ($r = 0{,}842$, $p < 0{,}001$). S tem se je izkazalo, da bi udeležence MPŠ lahko podobno ustrezno kot z uvrstitvenim testom razvrstili v skupine, homogene po jezikovnem znanju, le na podlagi rezultatov da/ne testa.

Primerjava odgovorov vseh 25 udeležencev MPŠ, ki so sodelovali v obeh testiranjih, je pokazala rahlo izboljšanje poznavanja v slovenščini obstoječih besed ($M = 3108 : 3296$, $Mdn = 3450 : 3750$), pri prepoznavanju nebesed pa so dosegli malenkost slabši rezultat ($M = -320 : -440$, $Mdn = 0 : -250$). T-test je za te testirane pokazal statistično značilne razlike med prvim in drugim testiranjem pri skupnih točkah ter rezultati za prvih, drugih in četrtih tisoč besed ($p < 0{,}05$), ne pa za korigirane točke ($p = 0{,}565$) in za tretjih tisoč besed ($p = 0{,}108$).

Domneva, da bo do največjega izboljšanja rezultata pri drugem testiranju prišlo pri testiranih z nižjim rezultatom pri prvem testiranju, se ni potrdila. Za skupini nadaljevalcev in izpopolnjevalcev, ki so sodelovali tako pri T1-MPŠ kot pri T2-MPŠ, za posameznih tisoč besed ni bilo ugotovljeno, da bi bile vrednosti statistično značilno drugačne ($p > 0{,}05$); statistično značilno se je spremenilo le skupno število točk pri skupini izpopolnjevalcev ($p = 0{,}013$).

Prva izvedba da/ne testa je torej pokazala, da so testirani v večji meri poznali bolj frekventne besede kot manj frekventne, da je z njim mogoče govorce SDTJ razvrstiti glede na njihovo jezikovno znanje in da bi ga lahko na MPŠ uporabili kot alternativo uvrstitvenemu testu.

Druga izvedba da/ne testa in prva izvedba piTeBeFRa

Razlogi za izvedbo testiranja

Prva izvedba je bila omejena na majhno, starostno sorazmerno homogeno skupino testiranih. V drugi izvedbi smo zato želeli preveriti, kako test deluje pri drugačni publiki, in sicer pri odraslih učečih se SDTJ. Želeli smo si starostno in glede prvega jezika testiranih raznoliko skupino.

Vsi testirani na MPŠ so imeli vsaj nekaj predznanja slovenščine, zato nas je pri drugi izvedbi zanimalo, kako da/ne test deluje pri popolnih začetnikih. Poleg tega smo želeli preveriti, kako so rezultati da/ne testa povezani z jezikovnim znanjem testiranih po *Skupnem evropskem jezikovnem okviru* (SEJO)[38] in kako so povezani z njihovim prvim jezikom.

Ker da/ne test meri le prepoznavo besed, ne pa tudi poznavanja njihovega pomena, in ker lahko taki testi precenjujejo znanje testiranih,[39] smo želeli preveriti, ali testirani poznajo pomen besed, vključenih v da/ne test. Zato je bil ob drugi izvedbi pilotno preizkušen tudi test besedišča po frekvenčnih razredih piTeBeFRa.

Zanimalo nas je, ali je da/ne test uporaben za spremljanje napredka učečih se slovenščine pri poznavanju besedišča v daljšem časovnem obdobju. V prvi izvedbi testa na MPŠ po 40-urnem tečaju in dveh tednih učenja slovenščine nismo zaznali statistično značilnega napredka v vseh postavkah (skupne točke, korigirane točke in rezultati za vse frekvenčne razrede), zato smo se pri drugi izvedbi odločili za testiranje na daljših tečajih (drugo ponovitev smo izvedli na vsaj 80-urnih tečajih) in v daljšem časovnem obdobju (približno 3 meseci).

Hipoteze

Na podlagi prve izvedbe da/ne testa so bile oblikovane naslednje hipoteze:

Hipoteza 1: *Poznavanje besed glede na njihovo pogostost in profili poznavanja besedišča*
Predvidevamo, da bo tudi druga izvedba da/ne testa potrdila, da testirani v večji meri poznajo bolj frekventne besede kot manj frekventne (hipoteza 1.1).
Poleg tega predvidevamo, da se bodo za testirane z različnimi ravnmi jezikovnega znanja (začetniki, nadaljevalci, izpopolnjevalci) pokazali različni profili poznavanja frekvenčnih razredov (hipoteza 1.2), in sicer pričakujemo, da se bo pri začetnikih profil začel stopničasto spuščati že pri drugi tisočici najfrekventnejših besed, pri nadaljevalcih in izpopolnjevalcih pa pozneje (hipoteza 1.3).

38 Svet Evrope, *Skupni evropski jezikovni okvir: učenje, poučevanje, ocenjevanje* (Ljubljana: Ministrstvo RS za šolstvo in šport, Urad za razvoj šolstva, 2011).

39 Gl. Durrant et al., *Research Methods in Vocabulary Studies*, 178.

Hipoteza 2: *Da/ne test kot orodje za razvrščanje udeležencev po jezikovnem znanju*
Tako kot pri prvi izvedbi pričakujemo, da bo mogoče tudi odrasle testirane na podlagi rezultatov da/ne testa razvrstiti glede na njihovo jezikovno znanje primerljivo kot z uvrstitvenim testiranjem (hipoteza 2.1).
Poleg tega predvidevamo, da bodo rezultati da/ne testa povezani z ravnjo jezikovnega znanja testiranih po SEJO, pri čemer pričakujemo, da bodo testirani, ki svoje znanje ocenjujejo višje (npr. na ravni B2), dosegli boljše rezultate v primerjavi s tistimi, ki svoje znanje ocenjujejo nižje (npr. na ravni A2) (hipoteza 2.2).

Hipoteza 3: *Rezultati da/ne testa glede na prvi jezik testiranih*
Pričakujemo, da se bodo pri da/ne testu pokazale razlike med govorci slovenščini sorodnih jezikov in drugimi govorci SDTJ.

Hipoteza 4: *Primerjava rezultatov prvega in drugega testiranja z da/ne testom*
Z dvema ponovitvama testa, ob začetku in koncu semestra, želimo ugotoviti, kako se napredek v jezikovnem znanju odraža v rezultatih da/ne testa. Predvidevamo, da bodo testirani pri drugem testiranju dosegli statistično pomembno višji rezultat (hipoteza 4.1), spremembe pa bodo večje predvsem pri tistih, ki so pri prvem testiranju dosegli nižji rezultat (hipoteza 4.2). Čeprav te hipoteze pri prvi izvedbi testa na MPŠ ni bilo mogoče potrditi, predvidevamo, da bo potrjena pri odraslih in vključenih popolnih začetnikih.

Hipoteza 5: *Ujemanje rezultatov da/ne testa in piTeBeFRa*
Glede na ugotovitve tujih študij[40] pričakujemo močno povezanost med rezultati da/ne testa in piTeBeFRa.

Potek testiranja in testirani

K reševanju testov so bili po elektronski pošti povabljeni vsi, ki so v spomladanskem semestru 2024 obiskovali tečaje CSDTJ za odrasle (Tabela 2). Izvedeni sta bili dve testiranji: prvo ob začetku semestra (T1), v katerem smo izvedli da/ne test, in drugo ob koncu semestra (T2), ko sta bila zaporedno izvedena da/ne test in piTeBeFRa.

40 Akira Mochida in Michael Harrington, »The Yes/No Test as a Measure of Receptive Vocabulary Knowledge,« *Language Testing* 23, št. 1 (2006): 73–98, pridobljeno 17. 2. 2025, https://doi.org/10.1191/0265532206lt321oa.

Tabela 2: Tečaji CSDTJ in število testiranih pri T1 in T2, ki so obiskovali posamezni tečaj.

Tečaj	Izvedba	Št. ur pouka	Št. testiranih pri T1	Št. testiranih pri T2
Jutranji tečaj	v živo	80	14	5
Popoldanski tečaj	prek videokonference	80	34	18
Spomladanska šola	v živo	170	19	8
Slovenščina za študente	v živo	48	3	
Tečaj za zaposlene na UL	prek videokonference	72	14	

Vir: lastno delo

Pri T1 so testirani odgovarjali od 26. februarja do 10. marca 2024, pri T2 pa med 11. in 17. junijem 2024. Respondenti v obeh testiranjih so test reševali na računalnikih (T1: 53,6 %, T2: 48,4 %), telefonih (T1: 45,2 %, T2: 51,6 %) ali tablicah (T1: 1,2 %). Če pri izračunu povprečja ne upoštevamo osamelcev, so testirani pri T1 da/ne test reševali povprečno 8 min 44 s, pri T2 pa so oba testa skupaj (da/ne test in piTeBeFRa) reševali povprečno 19 min 53 s. Med obiskovanjem tečajev so testirani živeli tako v Sloveniji kot zunaj nje.

Pri T1 je 94 učečih se soglašalo z zbiranjem osebnih podatkov in začelo reševati test. Iz analize so bili izločeni vsi testi, pri katerih je bilo izpolnjevanje prekinjeno in niso bili izpolnjeni do konca, vključeni pa so bili tisti, pri katerih je bilo neodgovorjeno samo posamezno vprašanje,[41] testirani pa so test rešili do konca. Tako je bilo v T1 pridobljenih 84 ustrezno rešenih testov. Med temi 84 testiranci je bilo 35 moških in 49 žensk. V času reševanja testa so bili stari od 16 do 67 let (35,7 % jih je bilo starih od 31 do 40 let, 22,6 % od 21 do 30 in 19 % od 41 do 50 let). 74 (88,09 %) jih je bilo univerzitetno izobraženih, 9 (10,71 %) srednješolsko, eden pa je imel osnovnošolsko izobrazbo. Govorili so 25 različnih prvih jezikov, in sicer angleščino (15), ruščino (13), hrvaščino (7), nemščino (5), srbščino (5), madžarščino (4), italijanščino (3), japonščino (3), makedonščino (3), španščino (3), ukrajinščino (3), bosanščino (2), francoščino (2), litovščino (2), nizozemščino (2), romunščino (2), urdujščino (2) in po eden bolgarščino, češčino, finščino, hindijščino, ruandščino, tajščino, turščino in vietnamščino.

Njihovo znanje slovenščine je bilo ocenjeno na podlagi uvrstitvenega testiranja; učitelji, ki so ga izvajali, so razlikovali med začetniki (Z) in boljšimi začetniki (Z+),

41 Manjkajoči odgovori so bili redki: osem testiranih ni odgovorilo na eno vprašanje, devet testiranih na dve vprašanji in po en testirani na tri oz. štiri od 96 vprašanj.

pri tistih z nadaljevalnim znanjem med nižjimi nadaljevalci (N–), nadaljevalci (N) in višjimi nadaljevalci (N+), kot izpopolnjevalce (I) so označili tiste z najvišjim znanjem (Tabela 3).[42]

Tabela 3: Število testiranih pri T1 glede na ocenjeno jezikovno znanje

Jezikovno znanje	Število testiranih
Izpopolnjevalci (I)	7
Višji nadaljevalci (N+)	13
Nadaljevalci (N)	14
Nižji nadaljevalci (N–)	12
Boljši začetniki (Z+)	23
Začetniki (Z)	15

Vir: lastno delo

Pri T2 so prav tako sodelovali udeleženci tečajev CSDTJ. V T2 je 36 učečih se soglašalo z zbiranjem osebnih podatkov in začelo reševati test. Iz analize so bili izločeni vsi testi, pri katerih je bilo izpolnjevanje prekinjeno in niso bili izpolnjeni do konca ter odgovori enega testiranega, ki ni odgovoril na nobeno vprašanje pri da/ne testu, ohranjeni pa so bili tisti, pri katerih so bila neodgovorjena posamezna vprašanja.[43] Tako je bilo pri T2 pridobljenih 31 ustrezno rešenih testov. Med temi 31 testiranci je bilo 12 moških in 19 žensk. V času reševanja testa so bili stari od 18 do 63 let (25,8 % jih je bilo starih od 41 do 50 let, po 22,6 % pa od 21 do 30 in od 31 do 40 let). 27 (84,4 %) je bilo univerzitetno izobraženih, štirje (15,6 %) pa srednješolsko. Govorili so 16 različnih prvih jezikov, in sicer angleščino (6), nemščino (3), ruščino (3), srbščino (3), hrvaščino (2), italijanščino (2), litovščino (2), makedonščino (2), po eden pa bosanščino, češčino, finščino, madžarščino, malgaščino, portugalščino, španščino in vietnamščino.

Ob zaključku obeh testov pri T2 so odgovorili tudi na vprašanje, kako ocenjujejo svoje znanje slovenščine po SEJO: 26 jih je svoje znanje ocenilo z ravnmi od A1 do C1, pet pa jih je izbralo odgovor *Ne vem* (Tabela 4).

42 Te oznake niso usklajene z ravnmi jezikovnega znanja po SEJO. Med boljše začetnike (Z+) in boljše nadaljevalce (N+) smo glede na to, kateri učbenik so uporabljali, prišteli tudi govorce slovanskih jezikov, ki so jih učitelji med uvrstitvenim testiranjem označili kot *Slovane začetnike* in *Slovane nadaljevalce*.

43 Pri da/ne testu so bili manjkajoči odgovori redki: osem testiranih ni odgovorilo na eno vprašanje, dva testirana na dve vprašanji in en testirani na tri od 96 vprašanj; pri piTeBeFRa pa je bilo manjkajočih odgovorov več: trije testirani niso odgovorili na eno vprašanje, po dva testirana na 15 oz. 25 vprašanj, po en testirani pa na 3, 4, 5, 8, 10, 11, 16, 17, 27, 29 ali 41 od 42 vprašanj.

Tabela 4: Število testiranih pri T2 glede na samooceno ravni jezikovnega znanja po SEJO

Jezikovno znanje	Število testiranih
C1	1
B2	5
B1	8
A2	3
A1	9
Ne vem.	5

Vir: lastno delo

Za primerjalno analizo da/ne testa v T1 in T2 so bili upoštevani samo odgovori oseb, ki so sodelovale v obeh testiranjih. Takih je bilo 25.

Rezultati

V tem razdelku rezultate prikazujemo glede na vrstni red hipotez. Rezultate smo statistično obdelali s programom Jamovi (verzija 2.3).[44]

Poznavanje besed glede na njihovo pogostost in profili poznavanja besedišča

Če pogledamo vse testirance, ki so sodelovali pri T1 in T2, lahko vidimo, da so v povprečju v večji meri poznali frekventnejše besede in za vsakih naslednjih tisoč dosegli nekoliko nižji rezultat (Tabela 5). V obeh testiranjih se je pokazal padajoči profil poznavanja besedišča (grafa 1 in 2). Analiza variance (ANOVA) za odvisne vzorce je pokazala, da se rezultati za posameznih tisoč besed pri T1 in T2 statistično pomembno razlikujejo (T1: ob upoštevanju Huynh-Feldtovega popravka $F(2{,}47, 204{,}61) = 88{,}7$, $p < 0{,}001$; T2: $F(3, 90) = 36$, $p < 0{,}001$). Primerjava povprečnih dosežkov za zaporedne frekvenčne razrede s post hoc testi je razkrila statistično značilne razlike med njimi (T1: za frekvenčna razreda 1–1000 in 1001–2000 je $p_{bonferroni} = 0{,}002$, za preostale pa $p_{bonferroni} < 0{,}001$; T2: za frekvenčna razreda 1–1000 in 1001–2000 je $p_{bonferroni} = 0{,}024$, za preostale pa $p_{bonferroni} < 0{,}001$), razen med frekvenčnima razredoma 2001–3000 in 3001–4000 pri T2 ($p_{bonferroni} = 0{,}595$).

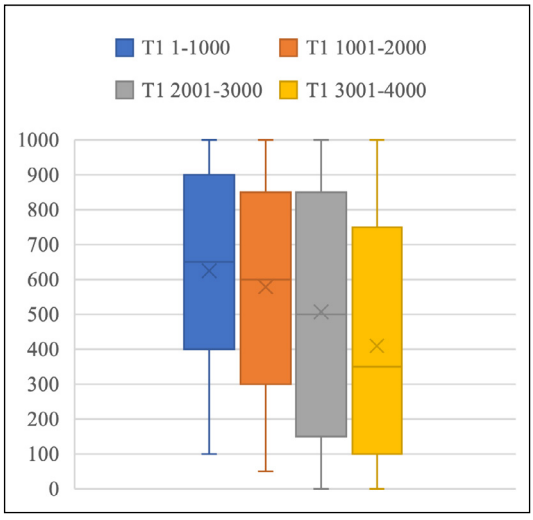
44 *Jamovi – open statistical software for the desktop and cloud*, https://www.jamovi.org/.

Tabela 5: Število točk, doseženih pri T1 in T2, za posamezne frekvenčne razrede

Testiranje	Frekvenčni razred	N	M	Mdn	SD	Minimum	Maksimum
T1	1–1000	84	624	650	282	100	1000
	1001–2000	84	578	600	302	50	1000
	2001–3000	84	507	500	348	0	1000
	3001–4000	84	409	350	321	0	1000
T2	1–1000	31	742	800	227	250	1000
	1001–2000	31	685	750	275	150	1000
	2001–3000	31	574	500	321	100	950
	3001–4000	31	534	600	341	50	1000

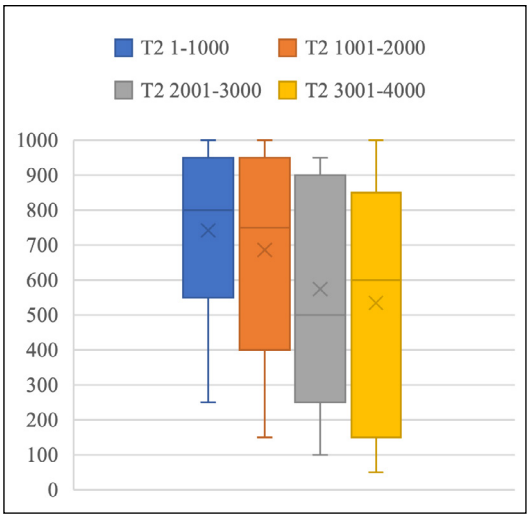
Vir: lastno delo

Graf 1: Število doseženih točk za posamezne frekvenčne razrede pri T1 (N = 84)



Vir: lastno delo

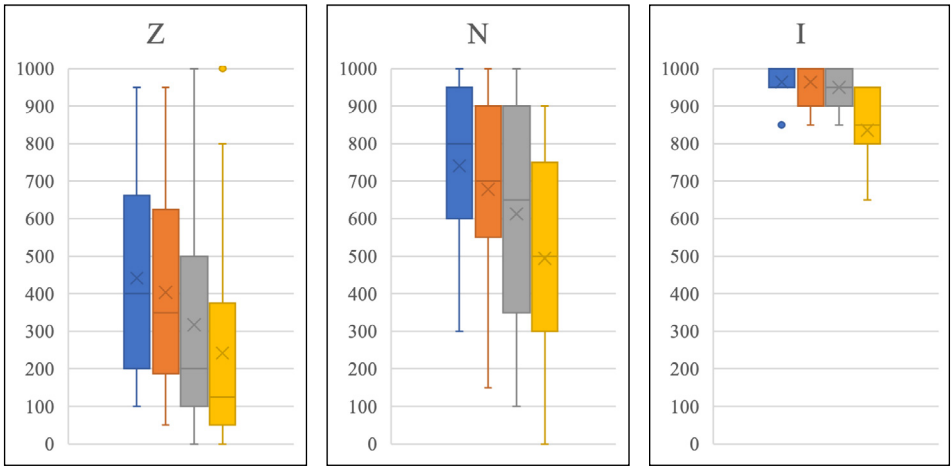
Graf 2: Število doseženih točk za posamezne frekvenčne razrede pri T2 (N = 31)



Vir: lastno delo

Ocene jezikovnega znanja, ki so jih testiranci dobili na podlagi uvrstitvenega testiranja, smo razdelili v tri kategorije: začetniki (Z), nadaljevalci (N) in izpopolnjevalci (I). Pri skupini Z je analiza variance za odvisne vzorce (*repeated measures ANOVA*) pokazala, da se rezultati za posameznih tisoč besed pri T1 med seboj statistično pomembno razlikujejo (ob upoštevanju Greenhouse-Geisserjevega popravka $F(2{,}27, 83{,}91) = 44{,}3$, $p < 0{,}001$). Post hoc test je potrdil statistično značilne razlike med vsemi frekvenčnimi razredi ($p_{bonferroni} < 0{,}001$) razen med prvima (1–1000 in 1001–2000), kjer razlika ni bila statistično značilna ($p_{bonferroni} = 0{,}147$). Tudi pri skupini N je analiza variance za odvisne vzorce pokazala, da se rezultati za posameznih tisoč besed med seboj statistično pomembno razlikujejo (ob upoštevanju Greenhouse-Geisserjevega popravka $F(2{,}39, 90{,}64) = 42{,}1$, $p < 0{,}001$), post hoc test pa je potrdil statistično značilno razliko med vsemi zaporednimi frekvenčnimi razredi (med prvima dvema: $p_{bonferroni} = 0{,}024$, med drugim in tretjim: $p_{bonferroni} = 0{,}006$, med tretjim in četrtim: $p_{bonferroni} < 0{,}001$). Pri skupini I je analiza variance za odvisne vzorce prav tako pokazala, da se rezultati za posameznih tisoč besed med seboj statistično pomembno razlikujejo (ob upoštevanju Greenhouse-Geisserjevega popravka $F(1{,}55, 9{,}32) = 7{,}26$, $p = 0{,}016$). Post hoc test ni potrdil statistično značilnih razlik med posameznimi frekvenčnimi razredi ($p > 0{,}05$), skoraj značilna pa je bila razlika med razredoma 1–1000 in 3001–4000 ($p_{bonferroni} = 0{,}057$).

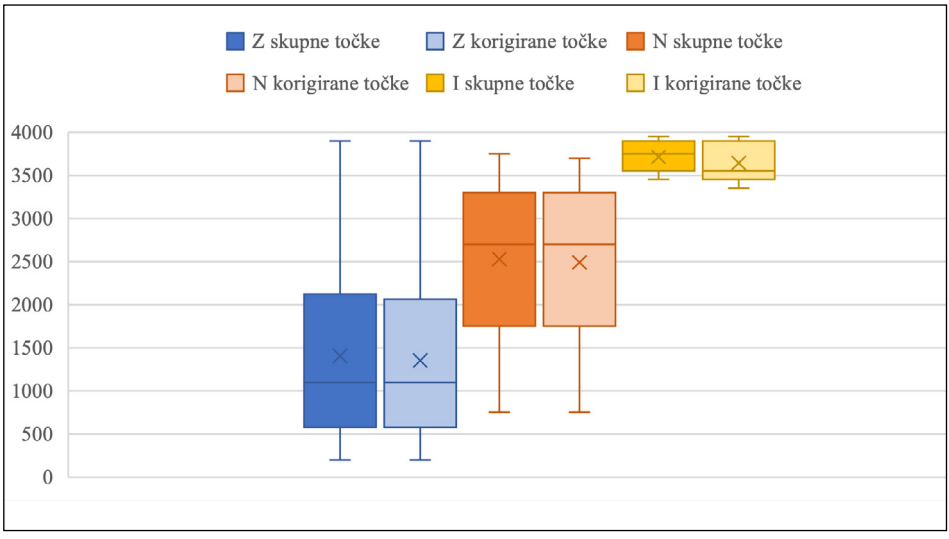
Grafi 3, 4, 5: Število doseženih točk za posamezne frekvenčne razrede pri T1 za začetnike (Z: N = 38), nadaljevalce (N: N = 39) in izpopolnjevalce (I: N = 7)



Vir: lastno delo

Da/ne test kot orodje za razvrščanje udeležencev po jezikovnem znanju

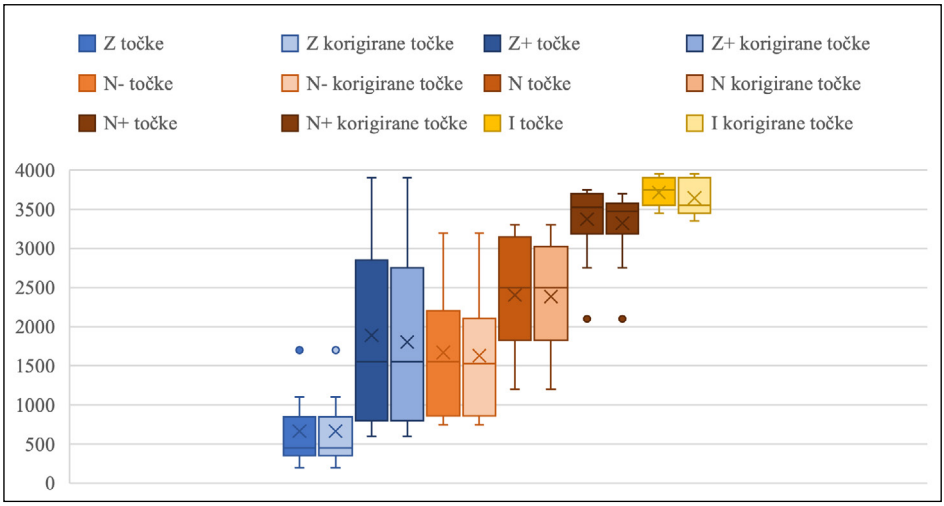
Graf 6: Skupno število točk in korigirane točke pri T1 glede na ocene jezikovnega znanja (Z: N = 39; N: N = 39; I: N = 7)



Vir: lastno delo

Kot je razvidno že iz Grafov 3, 4 in 5, so se dosežki testiranih na različnih ravneh jezikovnega znanja razlikovali. Skupno število doseženih točk in korigirane točke za te tri skupine prikazuje Graf 6. Izvedena je bila enosmerna ANOVA za primerjavo skupin glede na število doseženih točk in glede na korigirane točke pri T1. Rezultati so pokazali statistično značilne razlike med skupinami Z, N in I (skupne točke: $F(2, 50{,}2) = 85{,}4$, $p < 0{,}001$; korigirane točke: $F(2, 45{,}7) = 82{,}5$, $p < 0{,}001$). Tukeyjev post hoc test je pokazal, da so bile razlike statistično značilne med vsemi pari skupin ($p < 0{,}05$).

Graf 7: Skupno število točk in korigirane točke pri T1 glede na oznake jezikovnega znanja (Z: N = 15; Z+: N = 23; N–: N = 12; N: N = 13; N+: N = 14; I: N = 7)



Vir: lastno delo

Primerjali smo tudi rezultate glede na natančneje opredeljeno jezikovno znanje (Z, Z+, N–, N, N+, I). Graf 7 kaže, da razlike med vsemi skupinami niso zelo izrazite. Z enosmerno analizo variance (ANOVA) so bile potrjene statistično značilne razlike med skupinami (točke: $F(5, 34{,}2) = 91{,}6$, $p < 0{,}001$; korigirane točke: $F(5, 33{,}6) = 82{,}4$, $p < 0{,}001$). Tukeyjev post hoc test je pokazal pomembne razlike med večino parov skupin ($p < 0{,}05$). Statistično značilnih razlik pa ni bilo med naslednjimi skupinami pri skupnih točkah: Z+ in N– ($p = 0{,}967$), N– in N ($p = 0{,}177$) in N+ in I ($p = 0{,}931$); pri korigiranih točkah: Z+ in N– ($p = 0{,}986$), Z+ in N ($p = 0{,}22$), N– in N ($p = 0{,}123$), N+ in I ($p = 0{,}935$).

Pri T2 so testirani ocenili svoje znanje slovenščine po SEJO. Kot kaže Tabela 6, so tisti, ki so svoje znanje ocenili višje, dosegli boljši rezultat.

Tabela 6: Skupno število točk in korigirane točke pri T2 glede na raven jezikovnega znanja

	Raven po SEJO	N	M	Mdn	SD	Minimum	Maksimum
T2 skupne točke	A1	9	1478	1200	734	550	2950
	A2	3	2433	2200	1168	1400	3700
	B1	8	3369	3350	345	2800	3850
	B2	5	3500	3850	728	2200	3850
	C1	1	3900	3900	/	3900	3900
T2 korigirane točke	A1	9	1339	1100	749	300	2950
	A2	3	2017	2200	548	1400	2450
	B1	8	2775	3150	1088	350	3750
	B2	5	3450	3750	706	2200	3850
	C1	1	3900	3900	/	3900	3900

Vir: lastno delo

Ker je bil v vzorcu le en testirani, ki je svoje znanje ocenil z ravnjo C1, je bil izločen iz analize variance, upoštevani pa so bili le testiranci, ki so se opredelili na ravneh od A1 do B2. Rezultati enosmerne analize variance kažejo, da obstajajo statistično značilne razlike glede na raven znanja po SEJO pri skupnih točkah ($p = 0{,}003$) in korigiranih točkah ($p = 0{,}006$). Tukeyjev post hoc test je pokazal, da so testirani na ravni A1 dosegli bistveno nižje rezultate kot tisti na B1 in B2, pri čemer so razlike močno statistično značilne ($p < 0{,}01$ pri skupnih točkah, $p < 0{,}05$ pri korigiranih točkah). Med testiranimi na ravni A2 in preostalimi ni statistično pomembnih razlik, kar kaže, da njihovi rezultati variirajo in niso enoznačno višji od nižjih ali nižji od višjih ravni, prav tako ni statistično pomembne razlike med testiranimi na ravneh B1 in B2.

Za izbrane tečaje[45] je bila izračunana korelacija med številom točk, doseženih na da/ne testu pri T1, in razvrstitvijo testiranih v skupino. Pri Spomladanski šoli, kjer je pouk potekal v petih skupinah in jo je obiskovalo 19 testiranih, se je pokazala močna pozitivna povezanost (za skupne in korigirane točke: $\tau = 0{,}777$, $p < 0{,}001$). Podobno velja za Popoldanski tečaj, kjer je pouk potekal v sedmih skupinah in ga je obiskovalo 34 testiranih (za skupne točke: $\tau = 0{,}718$, $p < 0{,}001$, za korigirane točke: $\tau = 0{,}752$, $p < 0{,}001$). Pri Jutranjem tečaju, kjer je pouk potekal v treh skupinah in ga je obiskovalo 14 testiranih, pa se je pokazala srednje močna pozitivna korelacija (za skupne točke: $\tau = 0{,}571$, $p = 0{,}011$, za korigirane točke: $\tau = 0{,}568$, $p = 0{,}011$). Tudi multinominalna logistična regresija je potrdila, da višje skupne točke na da/ne testu povečujejo verjetnost uvrstitve v višjo skupino (Spomladanska šola – skupne točke: $R^2_{McF} = 0{,}527$, korigirane točke: $R^2_{McF} = 0{,}528$; Popoldanski tečaj – skupne točke: $R^2_{McF} = 0{,}349$, korigirane točke: $R^2_{McF} = 0{,}375$; Jutranji tečaj – skupne točke: $R^2_{McF} = 0{,}501$, korigirane točke: $R^2_{McF} = 0{,}499$; za vse $p < 0{,}001$).

45 Za tečaj za študente korelacija ni bila izračunana, saj so ga obiskovali le trije testirani. Prav tako ni bila izračunana korelacija za tečaj za zaposlene na UL, saj skupine niso bile razporejene po jezikovnem znanju (npr. ocena jezikovnega znanja v skupini 3 je bila višja od tistega v skupini 4).

Rezultati da/ne testa glede na prvi jezik testiranih

Preverili smo, ali pri doseženem rezultatu na da/ne testu pri T1 obstajajo razlike glede na to, ali je prvi jezik testiranega slovanski ali neslovanski, in glede na ocenjeno jezikovno znanje (Z, N, I). Slovanski govorci so v povprečju dosegli višje rezultate kot neslovanski (v povprečju za 967 točk pri skupnih točkah oziroma 858 točk pri korigiranih; Tabela 7). Za statistično analizo[46] je bila uporabljena dvofaktorska analiza variance (ANOVA). Rezultati kažejo statistično značilne razlike med skupinami Z, N in I (za skupne točke: $F(2,78) = 23{,}37$, $p < 0{,}001$, $\eta^2 p = 0{,}375$), prav tako se kot pomembna kaže pripadnost jezikovni skupini (za skupne točke: $F(1,78) = 20{,}01$, $p < 0{,}001$, $\eta^2 p = 0{,}204$). Zaznana je bila tudi statistično značilna interakcija med ocenjenim jezikovnim znanjem in jezikovno pripadnostjo (za skupne točke: $F(2,78) = 5{,}30$, $p = 0{,}007$, $\eta^2 p = 0{,}120$). Rezultati post hoc testov so pokazali, da so v skupini Z slovanski govorci dosegli bistveno višje rezultate kot neslovanski; razlika v povprečjih znaša 1815 točk in je statistično značilna ($p < 0{,}001$, $d = 2{,}53$). Razlika v skupini N je nekoliko manjša (razlika v povprečjih znaša 1148 točk), a je statistično značilna ($p < 0{,}001$, $d = 1{,}60$). Pri skupini I pa razlike med slovanskimi in neslovanskimi govorci niso bile statistično značilne ($p = 1{,}000$, $d = -0{,}09$), kar pomeni, da jezikovna pripadnost pri tej ravni znanja ni več pomemben dejavnik; slovanski govorci v tej skupini so dosegli celo malenkost nižji rezultat.

Tabela 7: Skupno število točk in korigirane točke pri T1 glede na oceno jezikovnega znanja za govorce slovanskih in neslovanskih jezikov

	Jezikovno znanje	J1 je slovanski	N	M	Mdn	SD	Minimum	Maksimum
T1 skupne točke	Z	da	11	2695	2850	857	1250	3900
		ne	27	880	700	594	200	2500
	N	da	19	3116	3300	631	1900	3750
		ne	20	1968	1775	930	750	3300
	I	da	4	3688	3700	221	3450	3900
		ne	3	3750	3750	200	3550	3950
T2 korigirane točke	Z	da	11	2536	2750	851	1250	3900
		ne	27	870	700	586	200	2500
	N	da	19	3050	3300	647	1700	3700
		ne	20	1955	1775	913	750	3300
	I	da	4	3563	3500	239	3350	3900
		ne	3	3750	3750	200	3550	3950

Vir: lastno delo

46 Ker so rezultati analiz za korigirane točke zelo podobni rezultatom za skupne točke, so v nadaljevanju navedeni samo rezultati za skupne točke.

Primerjava rezultatov prvega in drugega testiranja z da/ne testom

Odgovori 25 testiranih, ki so sodelovali v obeh testiranjih z da/ne testom, kažejo, da so pri T2 poznali več v slovenščini obstoječih besed (M = 1928 : 2334), nekoliko slabši pa so bili pri prepoznavanju nebesed (M = –20 : –180). T-test za ponovljene meritve je potrdil statistično značilno izboljšanje za skupne (za 406 besed, $p < 0{,}001$) in korigirane točke (za 246 besed, $p = 0{,}008$) ter za posamezne frekvenčne razrede ($p < 0{,}01$).

Rezultate testiranih smo opazovali tudi glede na njihovo jezikovno znanje (Z, N, I), kot je bilo ocenjeno ob začetku tečaja (Tabela 8). Pri skupini Z (N = 10) se je rezultat statistično značilno izboljšal tako pri skupnih točkah kot za vsak frekvenčni razred ($p < 0{,}05$). Največji napredek so dosegli pri bolj frekventnih besedah. Pri korigiranih točkah se je njihov rezultat sicer povečal, a ni bil statistično značilen ($p = 0{,}148$). Nadaljevalci (N = 13) so svoj rezultat statistično značilno izboljšali v vseh opazovanih postavkah razen v tretjem frekvenčnem razredu (2001–3000). Rezultat v skupini I (N = 2) pri skupnih točkah se je le malenkost povečal, pri korigiranih pa je bil celo slabši. Pri teh točkah in tudi za posamezne frekvenčne razrede ni bilo statistično značilnih razlik ($p > 0{,}05$).

Tabela 8: : Rezultati t-testa za testirane, ki so sodelovali tako pri T1 kot pri T2, glede na jezikovno znanje.

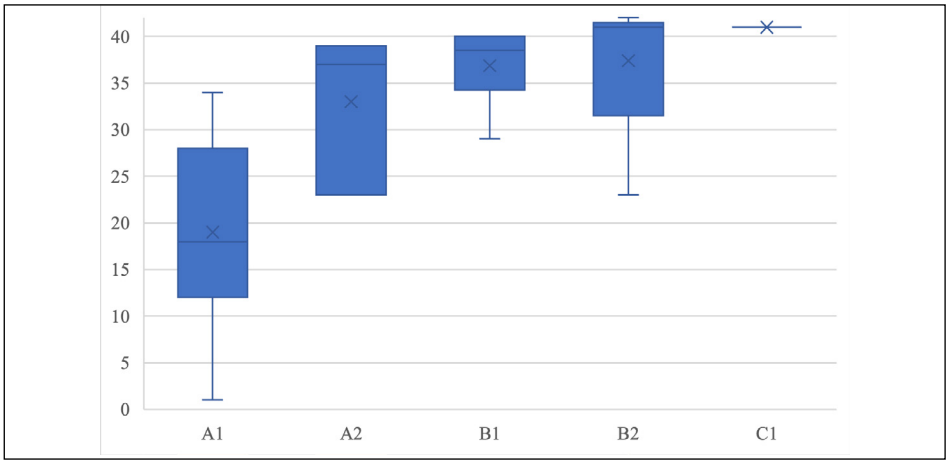
	Z		N		I	
	t (p)	razlika med povprečji (SE)	t (p)	razlika med povprečji (SE)	t (p)	razlika med povprečji (SE)
skupne točke	4,98 (p < 0,001)	415 (83,3)	4,42 (p < 0,001)	450 (101,7)	0,33 (p = 0,795)	75 (225,0)
korigirane točke	1,58 (p = 0,148)	215 (135,8)	3,99 (p = 0,002)	373,1 (93,5)	–1,54 (p = 0,366)	–425 (275,0)
1–1000	3,78 (p = 0,004)	145 (38,3)	4,16 (p = 0,001)	100 (24,0)	–1,00 (p = 0,500)	–25 (25,0)
1001–2000	4,12 (p = 0,003)	115 (27,9)	3,06 (p = 0,010)	123,1 (40,3)	1,00 (p = 0,500)	75 (75,0)
2001–3000	3,16 (p = 0,012)	85 (26,9)	1,97 (p = 0,072)	65,4 (33,2)	0,33 (p = 0,795)	25 (75,0)
3001–4000	3,77 (p = 0,004)	70 (18,6)	4,62 (p < 0,001)	161,5 (35,0)	0,00 (p = 1,000)	0 (50,0)

Vir: lastno delo

Ujemanje rezultatov da/ne testa in piTeBeFRa

Testirani, ki so odgovarjali na piTeBeFRa (N = 31), so v povprečju pravilno odgovorili na 29,9 od 42 vprašanj (*Mdn* = 34, *SD* = 11,09, *min* = 1, *maks* = 42). Tisti, ki so svoje znanje slovenščine po SEJO ocenili višje, so pravilno odgovorili na več vprašanj (Graf 8).

Graf 8: Rezultati testiranih na piTeBeFRa glede na samooceno ravni jezikovnega znanja po SEJO (N = 26)



Vir: lastno delo

Tabela 9: Kontingenčna tabela za odgovore pri da/ne testu in piTeBeFRa

Da/ne test \ piTeBeFRa	Pravilno	Napačno	Ni odgovora
Ja	747	45	57
Ne	175	76	195
Ni odgovora	5	0	2

Vir: lastno delo

Preverili smo, kako se odgovori pri da/ne testu in odgovori pri piTeBeFRa ujemajo (Tabela 9). Delež popolnega ujemanja odgovorov med odgovori pri da/ne testu in odgovori pri piTeBeFRa (1) znaša 78,91 odstotka, za posamezne testirane pa sega od 38,1 do 100 odstotkov (*M* = 78,26, *Mdn* = 80,49).

$$(1)\ \text{Popolno ujemanje}(\%) = \left(\frac{(\text{Ja \& Pravilno} + \text{Ne \& Napačno})}{\text{Ja \& (Pravilno ali Napačno)} + \text{Ne \& (Pravilno ali Napačno)}}\right) \times 100$$

Pokazala se je visoka pozitivna korelacija med rezultatom da/ne testa in številom pravilnih odgovorov na piTeBeFRa (za skupne točke: $r_S(29) = 0{,}882$, $p < 0{,}001$; za korigirane točke: $r_S(29) = 0{,}846$, $p < 0{,}001$).

Diskusija

Rezultati da/ne testa pri T1 in T2 pri odraslih učečih se SDTJ kažejo, da ti v večji meri poznajo bolj frekventne splošne besede v slovenščini kot manj frekventne (hipoteza 1.1), kar je skladno z ugotovitvami za mlajše učeče se.[47] Iz profilov štirih frekvenčnih razredov (Grafi 3, 4 in 5) je mogoče opaziti, da ta ugotovitev drži za učeče se z različnim jezikovnim znanjem, pričakovano pa tisti z višjim znanjem (N in I) poznajo več besed in so njihovi rezultati za vse frekvenčne razrede višji. Tako se je potrdila tudi domneva, da so profili poznavanja različnih frekvenčnih razredov pri testiranih z različnim jezikovnim znanjem (Z, N, I) različni (hipoteza 1.2). Vsi trije profili so »tipični« v tem, da se spuščajo proti manj frekventnim besedam:[48] pri skupini Z se je pokazal bolj enakomerno padajoč stopničasti profil kot pri skupinah N in I. Profil pri skupini I niti ni več stopničast in se spusti le pri zadnjem frekvenčnem razredu. Analize pa niso potrdile, da bi bile pri skupini Z statistično značilne razlike pri rezultatih že med prvim in drugim frekvenčnim razredom (hipoteza 1.3). Primerjava rezultatov za zaporedne frekvenčne razrede pokaže, da se pri skupini Z statistično značilne razlike izrazijo med drugim in tretjim ter tretjim in četrtim frekvenčnim razredom, pri skupini N med vsemi zaporednimi razredi, medtem ko pri skupini I ni statistično značilnih razlik. Primerjava razlik med povprečji za zaporedne razrede pokaže trend, da se večje razlike, ki so pri skupinah Z in N tudi statistično značilne, oziroma padci zamikajo v desno, k manj frekventnim besedam (pri Z: 38 → 87 → 75, pri N: 63 → 65 → 118, pri I: 0 → 14 → 114). Tako domnevamo, da bi se pri skupini izpopolnjevalcev statistično značilna razlika pokazala pri še manj frekventnih besedah, denimo pri peti ali šesti tisočici besed, ki pa v test nista bili vključeni. Ker je bilo v raziskavo vključenih 38 začetnikov in 39 nadaljevalcev, je mogoče sorazmerno zanesljivo trditi, da trend pri teh dveh skupinah drži. Izpopolnjevalcev pa je bilo pri T1 le sedem, zato bi bilo treba rezultate potrditi v prihodnji študiji z večjim vzorcem.

Tako kot pri prvi izvedbi da/ne testa[49] se je tudi ob drugi izkazalo, da je rezultat močno povezan s splošnim jezikovnim znanjem. Rezultati potrjujejo tudi, da je mogoče z da/ne testom dobro ločiti med posamezniki, ki so začetniki, nadaljevalci ali izpopolnjevalci, kar prav tako potrjuje prejšnje ugotovitve. Pri preverjanju razlikovanja na podrobnejši ravni, kjer so testirani znotraj treh večjih skupin razdeljeni na bolj in manj jezikovno zmožne, post hoc testi niso potrdili razlik med vsemi

47 Klemen, »Test poznavanja splošnih besed.«

48 Paul Meara, *EFL Vocabulary Tests*, 2. izdaja (Swansea: _lognostics, 2010), 5, 6, pridobljeno 13. 5. 2021, https://www.lognostics.co.uk/vlibrary/meara1992z.pdf.

49 Klemen, »Test poznavanja splošnih besed.«

zaporednimi pari skupin, kar nakazuje, da so povprečne vrednosti točk pri T1 med njimi podobne. Na podlagi dobljenih rezultatov testiranih na prehodu med stopnjami (Z+ ↔ N–, N+ ↔ I) s tem testom ni mogoče zanesljivo uvrstiti. Hipoteza 2.1 je tako potrjena le delno.

Pri T2 so testirani samoocenili svoje znanje slovenščine po SEJO. Zaradi manjšega vzorca je bilo analizo ANOVA mogoče izvesti le za ravni od A1 do B2. Rezultati analize variance v grobem podpirajo hipotezo 2.2: višja raven jezikovnega znanja praviloma pomeni boljši rezultat. Ta trend je opazen pri večini ravni, pri testiranih na ravni A2 pa v primerjavi z A1 in B1 ni bilo statistično značilnih razlik. Opozoriti je treba, da je bilo pri T2 število testiranih na posamezni ravni zelo majhno, sploh za raven A2 (N = 3), zato so ti rezultati predvsem orientacijske narave. Prav tako je treba opomniti, da zanesljivost samoocen jezikovnega znanja ostaja vprašljiva, saj so lahko subjektivne in ne odražajo nujno dejanske jezikovne zmožnosti. Kljub temu rezultati kažejo podoben trend kot rezultati na T1 in ocene jezikovnega znanja pri uvrstitvenem testiranju (Z, Z+, N– itn.). Za večjo zanesljivost bi bilo smiselno te ugotovitve preveriti na večjem vzorcu in z objektivnejšo oceno ravni znanja, na primer z izpitom jezikovnega znanja, umerjenim na SEJO.

Primerjava rezultatov govorcev različnih jezikov kaže, da slovanski govorci v primerjavi z govorci drugih jezikov na da/ne testu dosegajo boljše rezultate med Z in N, med I pa te razlike ni več. Take razlike so pričakovane, saj lahko govorci slovanskih, zlasti južnoslovanskih jezikov, ki si s slovenščino delijo del besedišča ali je njihovo besedišče podobno, zaradi pozitivnega transferja lažje sklepajo o pomenu besed v slovenščini.[50] To pa pomeni, da višji rezultat na da/ne testu pri slovanskih govorcih na začetnih ravneh učenja slovenščine ne pomeni nujno tudi boljše splošne jezikovne zmožnosti v slovenščini. Rezultati v glavnem potrjujejo domneve (hipoteza 3), zdi pa se, da se glede poznavanja splošnega besedišča razlike med govorci različnih jezikov z izboljšanjem jezikovne zmožnosti v slovenščini zmanjšujejo.

Ob upoštevanju povedanega, korelacijskih koeficientov med oceno jezikovnega znanja in točkami na da/ne testu pri T1 ter rezultatov multinominalne logistične regresije za izbrane tri tečaje (Spomladanska šola, Popoldanski tečaj in Jutranji tečaj) se kaže, da bi da/ne test lahko uporabili za razvrščanje udeležencev različnih tečajev v skupine s primerljivim jezikovnim znanjem. Kot smo že opozorili,[51] bi bilo za natančnejšo sliko o dejanski jezikovni zmožnosti test smiselno dopolniti z nalogo za samostojno produkcijo. Rezultati pričujoče raziskave nakazujejo, da bi bila takšna naloga potrebna zlasti pri govorcih slovanskih jezikov.

Da/ne test je bil na omenjenih treh tečajih izveden dvakrat: ob začetku in zaključku tečaja. Rezultati kažejo, da so testiranci pri T2 v povprečju dosegli višji rezultat kot pri T1 v vseh opazovanih postavkah (skupne točke, korigirane točke in rezultati za posamezne frekvenčne razrede) (hipoteza 4.1). Pri primerjavi rezultatov glede na ocenjeno

50 Tatjana Balažic Bulc, »Jezikovni prenos pri učenju sorodnih jezikov (na primeru slovenščine in srbohrvaščine),« *Jezik in slovstvo* 49, št. 3–4 (2004): 77–89, pridobljeno 12. 2. 2025, https://doi.org/10.4312/jis.49.3-4.77-89.

51 Klemen, »Test poznavanja splošnih besed,« 616.

znanje ob začetku tečaja je bilo ugotovljeno, da so pri poznavanju splošnega besedišča najbolj napredovali N (450 besed), malo manj pa Z (415 besed) – oboji so rezultate statistično značilno izboljšali v večini opazovanih postavk –, pri I pa večje razlike ni bilo opaziti (75 besed). Čeprav so bili v testiranje vključeni tudi popolni začetniki, se hipoteza o največjem napredku pri učečih se z nižjim znanjem ni potrdila. Velika omejitev tega dela raziskave je majhno število testiranih (le dva izpopolnjevalca), zaradi česar rezultatov ni mogoče posplošiti.

Tak rezultat pa najverjetneje ne odraža dejanskega napredka v poznavanju besedišča. Določene besede, ki se na Referenčnem seznamu pogostih splošnih besed za slovenščino pojavljajo med najfrekventnejšimi, namreč ne sodijo med najbolj frekventne v kontekstu SDTJ,[52] kar pomeni, da se testirani začetniki z njimi najverjetneje niso srečali pri pouku. Če bi želeli z da/ne testom opazovati jezikovni napredek glede na pri pouku obravnavano besedišče, bi bilo da/ne test bolj smiselno oblikovati na podlagi seznama jedrnega besedišča za slovenščino kot tuji jezik.[53]

S piTeBeFRa smo želeli preveriti, ali testirani, ki v da/ne testu trdijo, da poznajo določeno kombinacijo črk kot slovensko besedo, poznajo tudi njen pomen. Potrdilo se je, da so odgovori testiranih pri obeh testih v veliki večini (78,26 %) enaki in da so rezultati obeh testov močno povezani. Korelacijski koeficient je bil tako kot v raziskavi Mochide in Harringtona o povezanosti točk, doseženih na da/ne testu, in rezultata VLT višji od 0,8.[54]

Pri piTeBeFRa nas je presenetil velik delež neodgovorjenih vprašanj. Na podlagi pogoste kombinacije odgovorov *Ne* pri da/ne testu in manjkajočega odgovora pri piTeBeFra je mogoče sklepati, da testirani, ki besede niso poznali, pri piTeBeFRa niso želeli tvegati oziroma niso ugibali o pomenu besede. Opozoriti je treba še, da neodgovorjeno vprašanje ne pomeni nujno, da testirani ne pozna besede. Ker so bile nekatere besede, vključene v piTeBeFRa, večpomenske, je mogoče domnevati, da ne pozna pomena, predstavljenega v definiciji, morda pa pozna katerega drugega.

Sklep

Prispevek predstavi dva testa, s katerima smo med govorci SDTJ preverjali poznavanje pogostih splošnih besed v slovenščini: da/ne test in pilotni test besedišča po frekvenčnih razredih (piTeBeFRa). Rezultati da/ne testa so potrdili, da testirani v večji meri poznajo bolj frekventne besede kot manj frekventne. Pokazalo se je, da je test uporaben za razvrščanje učečih se glede na njihovo jezikovno znanje. Ugotovili smo, da se pri nižjih ravneh znanja (Z in N) kažejo razlike v poznavanju besedišča med govorci slovanskih jezikov in preostalimi. V raziskavi nas je zanimal tudi napredek tečajnikov v

52 Med tistimi, ki sodijo v jedrno besedišče za ravni A1, A2 in B1 (Klemen, Arhar Holdt in Pollak, *Core Vocabulary*), je 39 odstotkov takih, ki jih ni mogoče najti med pogostimi splošnimi besedami.

53 Klemen, Arhar Holdt in Pollak, *Core Vocabulary*.

54 Mochida in Harrington, »The Yes/No Test,« 87.

enem semestru. Testirani, ki so sodelovali v dveh izvedbah testiranja – na začetku in na koncu semestra –, so ob koncu semestra izboljšali svoj rezultat, zlasti v skupini Z in N. Primerjava rezultatov da/ne testa in piTeBeFRa pa je potrdila zanesljivost da/ne testa.

Izvedeni testiranji tako dopolnjujeta ugotovitve prve izvedbe da/ne testa poznavanja splošnih besed v slovenščini na MPŠ leta 2022.[55] Test je bil tokrat preizkušen pri odraslih govorcih SDTJ, v testiranje pa so bili vključeni tudi začetniki. Kljub temu da gre za skupino govorcev različnih prvih jezikov, ki je starostno dovolj raznolika, rezultatov ni mogoče posploševati na vse odrasle govorce SDTJ. Velika večina testiranih v pričujoči raziskavi je bila namreč visoko izobražena in vzorec s tega vidika ni reprezentativen. Upoštevati je treba tudi dejstvo, da so bili testiranci zaradi vključenosti v učni proces vajeni reševanja različnih testov in nalog.

Da bi lahko ugotovitve posplošili, bi bilo treba take teste izvesti na večjem vzorcu in med različnimi publikami. Pri nekaterih od njih bi lahko – kot kažejo izkušnje z izpiti iz znanja slovenščine na vstopni ravni – težave povzročil že kognitivno zahtevnejši format vprašanj pri piTeBeFRa.[56]

Vpliv na rezultat piTeBeFRa je imelo lahko tudi prikazovanje vprašanj po vrsti glede na frekvenčne razrede. Da bi ta vpliv zmanjšali, bi bilo smiselno razviti računalniški program za izvedbo testa, v katerem bi z menjavanjem bolj frekventnih in manj frekventnih besed v zaporednih vprašanjih vplivali tudi na motivacijo za reševanje.[57]

Čeprav so rezultati piTeBeFRa potrdili, da testiranci poznajo vsaj en pomen večine besed, za katere v da/ne testu trdijo, da jih poznajo, je glede obsega besedišča mogoče le okvirno ugotoviti, da začetniki, ki so bili pri T1 verjetno na ravneh do nizke A2 po SEJO,[58] poznajo okoli 1000–1500 besed od 4000 najpogostejših v slovenščini, nadaljevalci, ki so bili na ravneh od A2 do B1, okoli 2200–2800 besed, izpopolnjevalci, ki so bili na ravneh od nizke B2 in višje, pa okoli 3500–3800 besed. Te ocene so približne, saj je bilo zlasti v skupini začetnikov in nadaljevalcev mogoče opaziti veliko raznolikost rezultatov.

Nikakor pa ni mogoče trditi, da s tako pripravljenima testoma preverjamo celoten obseg besedišča govorcev SDTJ. Testa namreč zajemata besedišče iz sorazmerno majhnega nabora 4000 lem, testirani pa so gotovo poznali tudi besede, ki se niso uvrstile na Referenčni seznam pogostih splošnih besed za slovenščino. V obeh testih smo preverjali poznavanje enobesednih poimenovanj, v prihodnje pa bi bilo smiselno razviti tudi test, s katerim bi bilo mogoče preverjati, koliko in kako govorci SDTJ poznajo tudi večbesedna poimenovanja, saj so kolokacije, frazemi, leksikalni koščki ipd. pomemben del slovarja.[59]

55 Klemen, »Test poznavanja splošnih besed.«

56 Gl. Ina Ferbežar in Mateja Eniko, »'Lah blatschem gotovina?': jezikovni profil uporabnika slovenščine na najnižji ravni,« v: Nataša Pirih Svetina in Ina Ferbežar, ur., *Na stičišču svetov: slovenščina kot drugi in tuji jezik. Obdobja 41* (Ljubljana: Založba Univerze, 2022), 99–108, pridobljeno 11. 4. 2025, https://doi.org/10.4312/Obdobja.41.99-108.

57 Nation, »The Vocabulary Size Test.«

58 Raven jezikovnega znanja je mogoče le približno oceniti glede na učbenik, ki so ga uporabljali.

59 Durrant et al., *Research Methods in Vocabulary Studies*, 15–19.

V prihodnje bi bilo treba v da/ne test vključiti večje število besed iz posameznih frekvenčnih razredov (Gyllstad in sodelavci priporočajo 3 odstotke)[60] in ga pripraviti tudi za nadaljnje frekvenčne razrede. Test besedišča po frekvenčnih razredih pa bi bilo treba dopolniti, da bi zajemal enako število besed iz posameznega frekvenčnega razreda, in njegovo veljavnost preveriti tudi z drugimi oblikami preverjanja poznavanja besed, npr. z intervjuji.[61] Nato pa bi bilo smiselno preveriti njegovo praktično vrednost, denimo kot orodje za spremljanje napredka pri učenju besedišča. Pri tem bi veljalo razmisliti, ali ne bi bilo bolj smiselno pripraviti testa besedišča, ki bi besede zajemal iz seznama besed za posamezne ravni jezikovnega znanja po SEJO, na primer iz seznama jedrnega besedišča za slovenščino.[62] S tem bi organizatorji tečajev in učitelji SDTJ dobili uporabnejša orodja, raziskovalci pa boljši vpogled v obseg receptivnega besedišča govorcev SDTJ.

Zahvala

Zahvaljujem se Jani Kete Matičič, vodji programa Tečaji slovenščine, in lekt. Tanji Jerman, vodji učiteljev, ter vsem učiteljicam, učiteljem in učečim se na Centru za slovenščino kot drugi in tuji jezik, ki so mi omogočili izvedbo testiranj. Recenzentoma hvala za natančno branje.

Viri in literatura

Arhar Holdt, Špela, Senja Pollak, Marko Robnik Šikonja in Simon Krek. »Referenčni seznam pogostih splošnih besed za slovenščino.« V: *Jezikovne tehnologije in digitalna humanistika: zbornik konference*, uredila Darja Fišer in Tomaž Erjavec, 10–15. Ljubljana: Inštitut za novejšo zgodovino, 2020. Pridobljeno 13. 5. 2021. http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Arhar-Holdt-et-al_Referencni-seznam-pogostih-splosnih-besed-za-slovenscino.pdf.

Balažic Bulc, Tatjana. »Jezikovni prenos pri učenju sorodnih jezikov (na primeru slovenščine in srbohrvaščine).« *Jezik in slovstvo* 49, št. 3–4 (2004): 77–89. Pridobljeno 12. 2. 2025. https://doi.org/10.4312/jis.49.3-4.77-89.

Durrant, Philip, Anna Siyanova-Chanturia, Benjamin Kremmel in Suhad Sonbul. *Research Methods in Vocabulary Studies*. John Benjamins Publishing Company, 2022. https://doi.org/10.1075/rmal.2.

Ferbežar, Ina in Mateja Eniko. »'Lah blatschem gotovina?': jezikovni profil uporabnika slovenščine na najnižji ravni.« V: *Na stičišču svetov: slovenščina kot drugi in tuji jezik. Obdobja 41*, uredili Nataša Pirih Svetina in Ina Ferbežar, 99–108. Ljubljana: Založba Univerze, 2022. Pridobljeno 11. 4. 2025. https://doi.org/10.4312/Obdobja.41.99-108.

Gyllstad, Henrik, Laura Vilkaitė in Norbert Schmitt. »Assessing Vocabulary Size through Multiple-Choice Formats: Issues with Guessing and Sampling Rates.« *ITL - International Journal of Applied*

60 Henrik Gyllstad, Laura Vilkaitė in Norbert Schmitt, »Assessing Vocabulary Size through Multiple-Choice Formats: Issues with Guessing and Sampling Rates,« *ITL – International Journal of Applied Linguistics* 166, št. 2 (2015): 278–306, pridobljeno 25. 2. 2025, https://doi.org/10.1075/itl.166.2.04gyl.

61 Prim. ibidem.

62 Klemen, Arhar Holdt in Pollak, *Core Vocabulary*.

Linguistics 166, št. 2 (2015): 278–306. Pridobljeno 25. 2. 2025. https://doi.org/10.1075/itl.166.2.04gyl.

Klemen, Matej. »Test poznavanja splošnih besed v slovenščini med udeleženci Mladinske poletne šole slovenščine.« V: *Jezikovne tehnologije in digitalna humanistika: zbornik konference*, uredila Špela Arhar Holdt in Tomaž Erjavec, 604–20. Ljubljana: Inštitut za novejšo zgodovino, 2024. Pridobljeno 3. 12. 2024. https://doi.org/10.5281/zenodo.13936445.

Klemen, Matej, Špela Arhar Holdt in Senja Pollak. *Core Vocabulary for Slovenian as L2 1.0*. Slovenian language resource repository CLARIN.SI, 2022. Pridobljeno 18. 11. 2022. http://hdl.handle.net/11356/1697.

Meara, Paul. *EFL Vocabulary Tests*. Druga izdaja. Swansea: _lognostics, 2010. Pridobljeno 13. 5. 2021. https://www.lognostics.co.uk/vlibrary/meara1992z.pdf.

Meara, Paul in Barbara Buxton. »An Alternative to Multiple Choice Vocabulary Tests.« *Language Testing* 4, št. 2 (1987): 142–54. Pridobljeno 22. 2. 2025. https://doi.org/10.1177/026553228700400202.

Meara, Paul in Glyn Jones. »Vocabulary Size as a Placement Indicator.« V: *Applied Linguistics in Society*, uredila Pamela Grunwell, 80–87. London: Centre for Information on Language Teaching and Research, 1988. Pridobljeno 9. 3. 2024. https://www.lognostics.co.uk/vlibrary/meara&jones1988.pdf.

Meara, Paul in Imma Miralpeix. »V_YesNo v1.0.« V: *Tools for Researching Vocabulary*, 113–33. Bristol, Blue Ridge Summit: Multilingual Matters, 2016. Pridobljeno 9. 3. 2024. https://doi.org/10.21832/9781783096473

Milton, James in Thomaï Alexiou. »Developing a Vocabulary Size Test in Greek as a Foreign Language.« V: *Advances in Research on Language Acquisition*, uredili Angeliki Psaltou - Joycey in Marina Mattheoudakis, 307–18. Thessaloniki: Greek Applied Linguistcs Association, 2010.

Mochida, Akira in Michael Harrington. »The Yes/No Test as a Measure of Receptive Vocabulary Knowledge.« *Language Testing* 23, št. 1 (2006): 73–98. Pridobljeno 17. 2. 2025. https://doi.org/10.1191/0265532206lt321oa.

Nation, I. S. P. »Testing and Teaching Vocabulary.« *Guidelines* 5, št. 1 (1983): 12–25.

Nation, I. S. P. *Learning Vocabulary in Another Language*. Tretja izdaja. Cambridge: Cambridge University Press, 2022. https://doi.org/10.1017/9781009093873.

Nation, Paul. »The Vocabulary Size Test.« 2012. Pridobljeno 21. 2. 2024. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-information-and-specifications.pdf.

Pollak, Senja, Špela Arhar Holdt, Simon Krek in Marko Robnik-Šikonja. *Reference List of Slovene Frequent Common Words*. Slovenian language resource repository CLARIN.SI, 2020. http://hdl.handle.net/11356/1346.

Read, John. *Assessing Vocabulary*. Cambridge: Cambridge University Press, 2000.

Schmitt, Norbert Diane Schmitt in Caroline Clapham. »Developing and Exploring the Behaviour of Two New Versions of the Vocabulary Levels Test.« *Language Testing* 18, št. 1 (2001): 55–88.

Svet Evrope. *Skupni evropski jezikovni okvir: učenje, poučevanje, ocenjevanje.* Ljubljana: Ministrstvo RS za šolstvo in šport, Urad za razvoj šolstva, 2011.

Webb, Stuart, Yosuke Sasao in Oliver Ballance. »The Updated Vocabulary Levels Test.« *ITL - International Journal of Applied Linguistics* 168, št. 1 (2017): 33–69. Pridobljeno 13. 5. 2021. https://doi.org/10.1075/itl.168.1.02web.

Zhang, Xian. »The I Don't Know Option in the Vocabulary Size Test.« *TESOL Quarterly* 47, št. 4 (2013): 790–811. Pridobljeno 2. 5. 2024. https://doi.org/10.1002/tesq.98.

Matej Klemen

KNOWLEDGE OF COMMON WORDS IN SLOVENIAN AMONG SPEAKERS OF SLOVENIAN AS A SECOND AND FOREIGN LANGUAGE

SUMMARY

This article presents two vocabulary tests developed for Slovenian as a second (L2) and foreign language (FL), based on similar tests for other languages: the yes/no test and a pilot vocabulary levels test. The study examines the familiarity with common words in Slovenian among L2 and FL learners and evaluates the effectiveness of the yes/no test in classifying learners by language proficiency.

The first administration of the yes/no test took place at the 2022 Youth Summer School of Slovenian, involving participants aged 13 to 18. The findings indicated that learners recognised frequent words more effectively than less common ones, and the test successfully distinguished between learners with different levels of Slovenian proficiency. This article focuses on the second administration of the yes/no test among adult learners at the Centre for Slovene as a Second and Foreign Language at the Faculty of Arts, University of Ljubljana in 2024. Unlike the first administration, the second included absolute beginners as well. Additionally, a pilot vocabulary levels test was introduced to validate the results of the yes/no test.

The results confirmed that speakers of Slovenian as an L2 and FL are more familiar with high-frequency words than with low-frequency ones. The yes/no test proved useful in classifying learners at broader levels (beginner, intermediate, advanced) but was less precise for those transitioning between these three levels. It also revealed that Slavic language speakers performed better at lower levels than non-Slavic speakers, likely due to the linguistic similarities. However, no significant differences were observed among advanced learners.

The study also examined progress over a semester-long course. Learners who participated in two test administrations (at the beginning and end of the semester) showed significant improvement, particularly beginners and intermediate learners.

The pilot vocabulary levels test showed a strong correlation with the yes/no test results, confirming its validity.

The study suggests further refinements for both tests, such as including more words and expanding the test across various frequency levels.

1.01 DOI: https://doi.org/10.51663/pnz.65.3.11

Jernej Kosi*

The Breadbasket of Slovenia: The Genealogy of a Metonym and Its Role in Nation-Building

IZVLEČEK

ŽITNICA SLOVENIJE: GENEALOGIJA METONIMIJE IN NJEN POMEN V PROCESU GRADNJE NACIJE

Članek obravnava genealogijo in vlogo, ki jo je imela in jo še zmeraj ima pri gradnji naroda metonimija »žitnica Slovenije«. Prekmurje, ki leži na severovzhodu Slovenije, so politiki, znanstveniki in novinarji pogosto opisovali kot žitnico države – deželo agrarnega izobilja, ki zagotavlja najpomembnejša žita in živila. Ta naziv, utemeljen na rodovitni prsti, ugodnem podnebju in obsežni pridelavi pšenice, koruze in krompirja, se je uveljavil po vključitvi regije v jugoslovansko državo po prvi svetovni vojni.

Pred letom 1919 se je izraz žitnica v slovenski tiskani kulturi pojavljal v splošnem prenesenem pomenu, specifična povezava s Prekmurjem pa se je pojavila šele v povezavi z razpadom Avstro-Ogrske in teritorialno reorganizacijo, ki je sledila. Raba metonimije je imela dvojno vlogo: služila je kot gospodarski opis in kot simbolni instrument nacionalne integracije. Slovenski uradniki in intelektualci, ki v glavnem niso bili seznanjeni s prekmursko realnostjo, so poudarjali kmetijsko bogastvo pokrajine, da bi upravičili njeno vključitev kot del slovenskega ozemlja v okviru novonastale Kraljevine Srbov, Hrvatov in Slovencev.

Medvojna publicistika je popularizirala podobo Prekmurja kot »naše žitnice« in jo vtisnila v slovensko nacionalno imaginacijo, in to kljub trdovratni revščini, prekomerni naseljenosti in pomanjkanju hrane v tej regiji. Ta paradoks – med pripovedjo o izobilju in izkušnjo pomanjkanja – ponazarja, kako je metonimija služila kot orodje nacionalne integracije.

* **PhD, Assoc. Prof., University of Ljubljana, Faculty of Arts, Department of History, Aškerčeva ulica 2, SI-1000 Ljubljana, jernej.kosi@ff.uni-lj.si; ORCID: 0000-0003-3260-3431**

Obenem z zamegljevanjem strukturnih šibkosti regije je spodbujala njeno simbolno apropriacijo in pripadnost, povezujoč Prekmurje s slovensko nacijo. Metonimija je še zmeraj v rabi. Zasledimo jo v političnih in akademskih kontekstih, kar odraža njeno aktualnost.

Ključne besede: Prekmurje (Slovenija), žitnica, narodotvorni diskurzi, postimperialne tranzicije, obmejna območja

ABSTRACT

This article traces the genealogy and nation-building role of the phrase "breadbasket of Slovenia" as a metonym for the Prekmurje region. Located in northeastern Slovenia, Prekmurje has often been portrayed by politicians, scientists, and journalists as the country's breadbasket: a land of agricultural abundance that provides essential grain and foodstuffs. This designation—grounded in fertile soil, a favorable climate, and its significant production of wheat, corn, and potatoes—became prominent after the region's incorporation into the Yugoslav state following the First World War.

Before 1919, the term žitnica ("breadbasket" and "granary" in English) appeared in Slovenian print culture in a broader figurative sense, but its specific association with Prekmurje emerged in the context of Austria-Hungary's collapse and the territorial reorganization that followed. That metonymy fulfilled a dual purpose: it served as an economic descriptor and as a symbolic instrument of national integration. Slovenian officials and intellectuals, largely unfamiliar with Prekmurje's realities, emphasized its agricultural wealth to justify its incorporation into the Slovenian territory of the newly created Kingdom of Serbs, Croats, and Slovenes.

Interwar journalism popularized the image of Prekmurje as "our breadbasket," embedding it in the Slovenian national imagination despite the region's persistent poverty, overpopulation, and food insecurity. This paradox—between the narrative of abundance and the lived experience of deprivation—illustrates how the breadbasket trope functioned as a tool of national integration. While obscuring structural fragilities, it fostered symbolic ownership and belonging, binding Prekmurje to the Slovenian nation. Persisting into the present, the metonym is still invoked across political and academic contexts, attesting to its ongoing significance.

Keywords: Prekmurje (Slovenia), breadbasket, nation-building discourses, post-imperial transitions, borderlands

Introduction

Prekmurje, a region in the northeasternmost part of Slovenia, is "our breadbasket." Slovenian parliamentarians have left no doubt about this during the recent debates. Positioned at the crossroads of Hungary, Austria, Croatia, and Slovenia this territory has been recently cast in the National Assembly as the area of Slovenia that provides bread "for us." Prekmurje is much more than just a region of opportunity, observed MP Nataša Sukič in November 2021: "After all, this is our country's breadbasket."[1]

Similarly, a year earlier, an MP representing the Hungarian national minority residing in Prekmurje emphasized that the proposed document was "important for the further development of Prekmurje as a breadbasket."[2] In March 2018, following six hours of debate over Slovenia's future at the plenary session, the Minister for the Environment and Spatial Planning, Irena Majcen, spoke as a guest and pointedly referred to Slovenia's breadbasket. Climate change would require investments in Prekmurje's irrigation if we wanted the region's rich soil to keep generating abundant yields, she argued.[3] Lastly, Franc Breznik, a member of the Slovenian Democratic Party, also stated at a meeting of the Parliamentary Committee on Economic Affairs in September 2021 that the plains of Prekmurje are an outstanding agricultural area—the breadbasket of Slovenia.[4]

Admittedly, these statements amount only to anecdotal evidence. Nevertheless, they represent a body of assertions voiced by Slovenian parliamentarians across a wide spectrum of worldviews and political convictions. Judging from available sources, at least in the last decade, there have been no voices in parliament questioning the economic role and symbolic significance of Prekmurje as Slovenia's breadbasket.[5]

Slovenian experts and scientists also occasionally use the phrase "breadbasket of Slovenia" or "Slovenian breadbasket" to denote Prekmurje. It is somewhat unexpected to come across antonomasia—a change of a proper name with the phrase—in texts where clarity and unambiguity are required.[6] Yet in a Slovene-speaking academic context, in contrast to parliamentary dialogues, such usage is often grounded in solid agronomic and statistical facts. As a result, the phrase functions not merely as a rhetorical

1 "27. redna seja Državnega zbora (25.11.2021)," https://www.dz-rs.si/wps/portal/Home/seje/izbranaSeja?seja=sEUeepT0199ZjKc4uJB30g&uid=F712F2333CEF8BC3C125878B003A9B78&mandat=VIII.

2 "Sejni zapisi Državnega zbora. 15. seja (27., 28. in 29. januar 2020)," https://fotogalerija.dz-rs.si/datoteke/Publikacije/Sejni_zapisi_Drzavnega_zbora/2018-2022/2020_01_27_S_15.pdf.

3 "Sejni zapisi državnega zbora. 55. izredna seja (16. marec 2018)," https://fotogalerija.dz-rs.si/datoteke/Publikacije/Sejni_zapisi_Drzavnega_zbora/2014-2018/2018_03_16_IS_55.pdf.

4 "Odbor za gospodarstvo. 15. redna seja (9. september 2021)," https://www.dz-rs.si/wps/portal/Home/seje/izbranaSejaDt/!ut/p/z1/04_Sj9CPykssy0xPLMnMz0vMAfIjo8zivSy9Hb283Q0N_L0NzA0CQ0xMQy28LA3c3U30w1EVuBsFmRoEuhg5-QYbGBsEBxvpRxGj3wAHcDQgTj8eBVH4jS_IDQ0NdVRUBAAe3pc5/dz/d5/L2dBISEvZ0FBIS9nQSEh/?seja=pjXJQDtOFA4g4-bXoV7DyQ&uid=33667F925DEC984DC125871B001E512F&mandat=VIII.

5 See Andrej Pančur, Katja Meden, Tomaž Erjavec, Mihael Ojsteršek, Mojca Šorn, and Neja Blaj Hribar, *Slovenian Parliamentary Corpus (1990–2022) siParl 4.0* (Ljubljana: Institute of Contemporary History, 2024), http://hdl.handle.net/11356/1936 (accessed online).

6 On cases of antonomasia in the Croatian language see Ana Grgić and Davor Nikolić, "'Ovaj grad zovu još i…' – o antonomazijama za toponime," *Folia onomastica Croatica* 23 (2014): 77–94.

device, but also as an implicit analytical category that shapes scholarly interpretations.[7] As highlighted by the authors of the "first regional geographical monograph of Slovenia," the lowland area of Prekmurje along the Mura/Mur River (together with a narrow strip of flat land on the other, Styrian bank) "is rightly considered the breadbasket of Slovenia, as in 1993 it produced a good third of all wheat, slightly less corn, and a tenth of all potatoes in Slovenia."[8]

Since at least the late 1980s, the notion of Prekmurje as the breadbasket of Slovenia has rested on convincing empirical evidence. Due to fertile soil, a favorable continental climate, and a landscape with substantial flat areas, conditions here are well suited for agriculture. Already before 1919, the landed estates had concentrated on intensive grain production, while in the socialist era after 1945, systematic reclamation of swampy areas and the introduction of heavy mechanization further intensified agricultural production.[9] Even the transitional period from the late socialist to the post-socialist economy did not diminish the importance of agriculture in Prekmurje. On the contrary, unlike other parts of Slovenia, which experienced a significant decline in farming, the area of arable land cultivated here increased by 14%.[10] The statistical region of Pomurje—which roughly overlaps with Prekmurje—is still Slovenia's principal agricultural zone. Although it covers only 6.6% of the country's territory, it accounts for more than a fifth of all arable land, on which almost half of all wheat and almost a third of all corn in Slovenia is grown today.[11]

Representations of Prekmurje, however, are marked by a paradox. Although the region is frequently portrayed as the "breadbasket of Slovenia," it is also closely associated with socio-economic fragility characteristic of geographically remote and underdeveloped areas. Underneath the solid statistical certainties and picturesque views of

7 Maks Wraber, "Gozdna vegetacijska slika in gozdnogojitveni problemi Prekmurja," *Geografski vestnik* 23 (1951): 179. Etelka Korpič-Horvat, *Zaposlovanje in deagrarizacija pomurskega prebivalstva* (Murska Sobota: Pomurska založba, 1992), 190. Oto Luthar, ed., *Prekmurje za radovedneže in ljubitelje* (Ljubljana: Založba ZRC, ZRC SAZU, 2010), 15, 16. Marijan M. Klemenčič et al., *Življenjska (ne)moč obrobnih podeželskih območij v Sloveniji* (Ljubljana: Znanstvena založba Filozofske fakultete, 2018), 13. Stanka Dešnik, "Barve trideželnega parka," in Darja Senčur Peček, ed., *Vanekovo stoletje: Ob stoletnici rojstva dr. Vaneka Šiftarja* (Murska Sobota: Univerzitetna založba Univerze, 2019), 225. Božo Repe, "*Vsakdo mora imeti priliko, da udejstvi vse svoje telesne in duševne moči.": Milko Brezigar in prvi slovenski program narodnega gospodarstva* (Ljubljana: Založba Univerze, 2023), 52, 63.

8 Drago Perko and Milan Orožen Adamič, eds., *Slovenia: Landscapes and People*, 3rd ed. (Ljubljana: Mladinska knjiga, 2001), 575.

9 On the pre-1919 aristocratic estates in Prekmurje and the interwar land reform, see Miroslav Kokolj, *Prekmurski Slovenci: od narodne osvoboditve do nacistične okupacije, 1919–1941* (Murska Sobota: Pomurska založba, 1984), 483–591. For an overview of the developments in the region's agriculture after 1945, see Korpič-Horvat, *Zaposlovanje in deagrarizacija pomurskega prebivalstva*, 153–73. For a vivid description of the survival strategies of the rural population during the socialist era, drawing in part on the testimonies of farmers from Prekmurje see Polona Sitar, "Agrikulturna modernizacija in življenjski svet podjetnih polkmetov. Integrirana kmečka ekonomija v socialistični Sloveniji," Prispevki za novejšo zgodovino 61, No. 2 (2021): 142–68. Lev Centrih and Polona Sitar, *Pol kmet, pol proletarec: integrirana kmečka ekonomija v socialistični Sloveniji, 1945–1991* (Koper: Založba Univerze na Primorskem, 2023), 151–228.

10 Tomaž Cunder, "Kmetijstvo v Pomurju danes in jutri," in Tatjana Kikec, ed., *Pomurje [Elektronski vir]: trajnostni regional razvoj ob reki Muri: zbornik / 20. zborovanje slovenskih geografov, Ljutomer – Murska Sobota, 26.–28. marec 2009* (Ljubljana: Zveza geografov Slovenije; Društvo geografov Pomurja, 2009), 146, https://www.drustvo-geografov-pomurja.si/projekti/zborovanje/Zbornik_geografov_POMURJE_2009.pdf.

11 *Regionalni razvojni program Pomurske regije 2021–2027* (Murska Sobota: Razvojni center Murska Sobota, 13 June 2022), 5, https://www.rcms.si/upload/files/RRP_Pomurje_2021-2027_13-6-2022.pdf.

grain fields on fertile plains, conveying an image of stability, security, and abundance, lies a long-standing and empirically well-documented reality of social vulnerability and economic fragility. The region was already overpopulated at the end of the nineteenth century, forcing landless farmers and members of smallholding families to seasonal and permanent migrations.[12] Similar conditions continued in interwar Yugoslavia: the land reform only accelerated the fragmentation of landholdings and the emergence of dwarf farms.[13] The miserable living conditions began very slowly to improve only after 1945. In the communist era, Prekmurje underwent gradual industrialization and deagrarianization, even though (hidden) rural overpopulation was still present. Until the collapse of socialist Yugoslavia, Prekmurje remained among the least developed regions of Slovenia, marked by high levels of unemployment and population decline.[14] Economic underdevelopment extended into the post-socialist transition. In 2003, the Pomurje region achieved only two-thirds of the average Slovenian GDP per capita. What is more, the global financial and economic crisis of 2008 hit the region, with its predominantly traditional labor-intensive industrial production, with full force.[15]

This article examines how the narrative of regional abundance has persisted in Prekmurje despite the region's long history of socio-economic vulnerability. The aim of the study is to trace the genealogy of Prekmurje's construction as the "breadbasket of Slovenia" and to explore what the origins of this phrase reveal about national(ist) visions and aspirations projected onto the region. It demonstrates that the metonymy crystallized in the immediate aftermath of the region's incorporation into the Yugoslav state in 1919 arose initially from the fascination of Slovenian officials from less fertile Cisleithanian lands. The term gained wider popularity because it served as a means of symbolically integrating a contested borderland into the imagined Slovenian national space. By analyzing references to Prekmurje as a Slovenian breadbasket in official documents and the press after the Yugoslav occupation and throughout the interwar decades, the study shows how the metonym functioned both as an economic designation grounded in fertile land and as a tool of nation-building in the context of post-imperial territorial and political reconfigurations.

12 Janez Malačič, "Demografski razvoj v Prekmurju 1919–2019: upadanje prebivalstva ter modernizacija razvoja," in Peter Štih et al., eds., *"Mi vsi živeti ščemo": Prekmurje 1919: okoliščine, dogajanje, posledice* (Ljubljana: Slovenska akademija znanosti in umetnosti, 2020), 353–55. Kokolj, *Prekmurski Slovenci*, 608–25.

13 Kokolj, *Prekmurski Slovenci*, 589–91.

14 Korpič-Horvat, *Zaposlovanje in deagrarizacija pomurskega prebivalstva*, 117–32.

15 Aleksander Lorenčič, *Prelom s starim in začetek novega: Tranzicija slovenskega gospodarstva iz socializma v kapitalizem* (1990–2004) (Ljubljana: Inštitut za novejšo zgodovino, 2012), 341, 451, http://hdl.handle.net/11686/38023. After restructuring in the last decade, however, the regional economy is stable, export-oriented, technologically advanced, and marked by record revenues, low unemployment, and rising productivity. See *Regionalni razvojni program pomurske regije 2021–2027* (Murska Sobota: Razvojni center Murska Sobota, 2022), 14.

Žitnica as a 'Breadbasket': Nineteenth-century Beginnings

The Slovenian equivalent of the word 'breadbasket' is '*žitnica*', a term that, much like its English counterpart, carries multiple meanings in both modern and historical lexicographical sources and dictionaries. In its literal sense, it refers to a place for storing grain, a granary. As a granary, the word '*žitnica*' is embedded in the vocabulary of material culture, storage management, and architectural features of warehouse spaces. Historically, the term also denoted a form of feudal tax, while since the nineteenth century it has been employed as a metonym.[16] In a figurative meaning, *žitnica* is what is referred to in Merriam Webster as a breadbasket: "a major cereal-producing region."[17] Dictionary evidence suggests that the figurative use of the word entered Slovenian language relatively late. While '*žitnica*' already denoted both a warehouse and a feudal tax in printed texts from the late eighteenth century, its figurative sense was still missing more than a century later from Pleteršnik's Slovenian–German dictionary.[18]

In contrast to Pleteršnik's dictionary, Slovenian newspapers and periodicals from the *Corpus of Slovenian Periodicals (1771–1914)* attest that by the second half of the nineteenth century the word '*žitnica*' was already employed as a figure of speech.[19] Nevertheless, many Slovenian authors continued to employ the term primarily in its literal sense, denoting enclosed structures on farms and landed estates where various types of grain were stored after the harvest.[20] The prevalence of this type of usage is not surprising due to the socioeconomic structure of present-day Slovenian territory, which was predominantly linked to agricultural production at least until the mid-twentieth century.

One defining feature of nineteenth-century Slovenian print culture was the strong focus of many authors on the world of agriculture and peasant society. Within this framework, Slovenian newspapers and journals up to the outbreak of the First World War mentioned '*žitnica*' in numerous pedagogical articles that gave normative descriptions of granaries: dry, airy, cool, clean spaces built from durable materials to prevent the grain from spoiling. Authors aimed at a rural readership further

16 See the entry for *"žitnica"* in *Slovar slovenskega knjižnega jezika*, 2nd ed., supplemented and partly revised edition, www.fran.si, accessed 4 September 2025. For historical usage of the word, see entry *"žitnica"* in Maks Pleteršnik, *Slovensko-nemški slovar*, www.fran.si, accessed 20 September 2025.

17 *Merriam-Webster.com Dictionary*, s.v. "breadbasket," accessed 20 September 2025, https://www.merriam-webster.com/dictionary/breadbasket.

18 For usage in the late eighteenth century, see entry *"žitnica"* in Marko Snoj, *Slovar Pohlinovega jezika*, www.fran.si, accessed 20 September 2025.

19 For the corpus, see Filip Dobranić, Bojan Evkoski in Nikola Ljubešić, »Corpus of Slovenian periodicals (1771–1914) sPeriodika 1.0« (2023), *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1881.

20 See for instance: "Grozen požar v Mokronogu," *Slovenec*, 21 Augusta 1911, 3. "Dražbeni oklic," 14 February 1914, 7. "Poučno potovanje učencev kmetijske šole na Grmu," *Narodni gospodar*, 10 August 1908, 244. "Spodnje Libuče," *Naš dom*, 27 July 1905, 4. "Setev je pred durmi," *Kmetovalec*, 15 January 1891, 2. "Gospodarske stavbe," *Novice*, 30 December 1898, 516. "Požarna kronika," *Ljubljanski list*, 16 June 1884, 4.

offered practical advice on pest control and on safeguarding stored grain from theft.[21] Yet the word '*žitnica*' was not confined exclusively to farm-level storage buildings. It also designated larger supply infrastructures—feudal, provincial, and municipal storage—activated by the administrative apparatus as mechanisms of collective support in times of food crisis.[22]

Beside predominant usage in agricultural contexts, from the mid-nineteenth century onward, the term '*žitnica*' also appeared in the public sphere in a figurative sense as breadbasket. Numerous geographical locations around the world have been referred to as breadbaskets. Unsurprisingly, the first chronological mention from the mid-nineteenth century refers to a region of global historical relevance. In 1850, an author writing in a magazine for Slovenian schoolchildren described Egypt as "the breadbasket for many countries with scarce vegetation."[23] During this period, the term "breadbasket" often referred to Tsarist Russia.[24] Given the Habsburg context, Hungary was also described as a breadbasket in Slovenian newspapers, as was the Hungarian province of the Banat.[25] Sicily, Eastern Rumelia, the Kosovo Plain, and Skadar were also portrayed as breadbaskets supplying wider regions or political entities.[26]

Yet, the term's usage in journalistic reports and descriptions was not neutral. The expression "breadbasket" was not employed merely as an objective marker of a region's agricultural potential to produce more grain than the local population could consume. Rather, regions were described as breadbaskets "for someone" or "from someone." This usage underscored the relational dynamics of food production and distribution, situating agricultural regions within wider political, economic, and imperial frameworks. In the second half of the nineteenth century, Russia was more than just a breadbasket: it was the "breadbasket of Europe." Meanwhile, Hungary was considered the "breadbasket of Austria" or the "breadbasket of our Empire," a label that some authors also associated with the Banat. The Kosovo Plain was dubbed "the breadbasket of European Turkey," while Skadar had a similar function for Montenegro and Albania.

21 See for instance "Vprašanja in odgovori," *Narodni gospodar*, 10 October 1903, 297. "Žužek," *Gospodarski glasnik za Štajersko*, 1 January 1912, 137. "Vprašanja in odgovori," *Narodni gospodar,* 25 August 1901, 248. "Vprašanja in odgovori," *Kmetovalec*, 31 December 1894, 191. "Gospodarska skušnja," *Novice*, 17 June 1874, 189. "Kako shranjujemo žito?," *Novice*, 28 February 1902, 84. "Srenja pod lipo," *Besednik. Kratkočasen in podučen list za slovensko ljudstvo*, 25 October 1870, 158.

22 "Iz Ljubljane," *Novice*, 21 February 1866, 68. "Henrik I., ptičar," *Vertec*, 1 June 1882, 91. "Iz Glamoča v Bosni," *Slovenski narod*, 18 April 1879, 4. "Zadnji Sorgo umrl," *Slovenski narod*, 5 December 1912, 2. "Vincencij Vovk," Zgodnja danica, 4 October 1872, 320. "Od hranjenja žita," *Pravi Slovenec*, listi za podučenje naroda, 6 August 1849, 169–71.

23 "Božja roka ali nevarno kopanje v Nilu," *Vedež: časopis za šolsko mladost*, 3 January 1850, 1.

24 "Podraženje kruha," *Edinost*, 25 September 1907, 1. "Rusko žito," *Slovenec*, 3 June 1912, 6.

25 For Hungary: "Gospodarske stvari," *Slovenski gospodar*, 25 September 1873, 316. "Razmera obrništva do kmetijstva," *Novice gospodarske, obrniške in narode*, 5 October 1864, 1. "Zveza obrtništva in kmetijstva," *Novice gospodarske, obrtniške in narode*, 23 November 1881, 1. "Pšenične cene po dobrih letinah," *Glasnik Avstrijske krščanske tobačne delavske zveze*, 14 January 1911, 7. For the Banat "Iz Gradca do Sarajeva," *Slovenski gospodar*, 26 December 1878, 422. "Železnic nam je treba," *Kmetijske in rokodelske novice*, 9 November 1864, 364.

26 For Sicily, "Slavia italiana," *Ljubljanski zvon*, 1 December 1884, 766. "Nekdaj in sedaj," *Dolenjske novice*, 15 May 1885, 76. "Na mestih strašnega potresa," *Slovenec*, 16 January 1909, 1. "Kmetijstvo," *Narodni gospodar*, 10 November 1900, 343. For Rumelia, "O balkanskih zadevah," *Slovenskih narod*, 14 October 1885, 2. For the Kosovo Plain, "Donavsko-adrijanska železnica," *Slovenski narod*, 17 March 1908, 1. "Iz gospodarskega življenja v Makedoniji," *Edinost*, 16 February 1913, 5. For Skadar, "Pismo s Cetinja," *Slovenec*, 9 December 1912, 1.

At the same time, writers could have invested their own beliefs, expectations, and visions regarding the production, circulation, and consumption of food in their use of the term breadbasket. In spring 1912, an author writing in the Slovenian liberal paper *Slovenski narod* argued that the marsh south of Ljubljana should be drained so that it could become the breadbasket of the capital of Carniola.[27] On the other hand, agriculture in the Italian region of Apulia—"once the breadbasket of Italy"—was declining due to high taxes according to a report published in 1889 in the same paper.[28] The content of articles in which the term appears could also have been influenced by an awareness of global and regional production asymmetries, as well as fears and hopes linked to the (in)ability of individual states to reach autarky. In 1902, the author of a short article on Russia highlighted the geopolitical importance of Russian agriculture, the cornerstone of the country's power, since "Russia, along with America, is considered the largest breadbasket, supplying the whole world with food."[29]

By the outbreak of the First World War, as evidence from the *Corpus of Slovenian Periodicals (1771–1914)* shows, the term "breadbasket" was already circulating in Slovenian public discourse, mainly in relation to supply, self-sufficiency, and production asymmetries. Yet the figure of speech that would later frame Prekmurje as Slovenia's "breadbasket" only appeared after the collapse of Austria-Hungary, in the context of Yugoslav occupation and the political transitions of 1918–19. In other words, Prekmurje was invented as the breadbasket of Slovenia as a result of the collapse of the centuries-old Habsburg imperial order in Central Europe.[30]

Prekmurje Becomes the Slovenian Breadbasket: The Invention of a Figure of Speech

After the beginning of the collapse of the Austria-Hungary, the population of the future Prekmurje region experienced a turbulent acceleration of history.[31] Ultimately, in the context of the post-imperial land grab and the creation of new state borders, the Yugoslav army occupied the two westernmost parts of the Hungarian counties Zala and Vas in August 1919. The Treaty of Trianon in July 1920 merely formalized

27 "Ljubljanski občni svet," *Slovenski narod*, 17 May 1912, 2.

28 "Vnanje države," *Slovenski narod*, 16 April 1889, 3.

29 "Skrb Rusije za kmetijstvo," *Slovenski gospodar*, 3 June 1902, 2.

30 Illustrative in this regard is the fact that Prekmurje was not included in the first Slovenian national economic program. See Božo Repe, "*Vsakdo mora imeti priliko, da udejstvi vse svoje telesne in duševne moči." Milko Brezigar in prvi slovenski program narodnega gospodarstva* (Ljubljana: Založba Univerze, 2023).

31 Miroslav Kokolj, "Prekmurje v prevratnih letih 1918–1919," in Janko Liška, ed., *Revolucionarno vrenje v Pomurju v letih 1918–1920* (Murska Sobota: Pomurska založba, 1981). György Feiszt, "Revolucionarni pokret u Prekmurju od 1918. do 1919.," in Branimir Bunjac, ed., *Pomurje 1914–1920.: zbornik radova = Mura mente 1914-1920.* (Povijesno društvo Međimurske županije, 2011). László Göncz, *A Muravidék útja a délszláv királyságba: a tájegység története az első világháború végétől a jugoszláv megszállásig: (1918 ősze – 1919 augusztusa)* (Magyar Nyugat Könyvkiadó, 2024). László Göncz, "Utrinki iz zgodovine Beltincev v t. i. prevratnem obdobju (od oktobra 1918 do jugoslovanske zasedbe Prekmurja)," in Sonja Novak-Lukanovič and Barbara Riman, eds., *Raznolikost v raziskovanju etničnosti: izbrani pogledi III* (Ljubljana: Inštitut za narodnostna vprašanja, 2023).

the authority of the South Slavic state over this area, predominantly populated by a Slavophone population. Within the Yugoslav state framework, the region received a new official administrative name—Prekmurje—and was subordinated to regional administrative bodies in Ljubljana.[32]

Long before Prekmurje was constructed as "Slovenia's breadbasket," the region had already been imagined as part of the Slovenian national space. The association of Prekmurje with Slovenia's breadbasket was thus not only a product of post-imperial political circumstances but was also connected to older ethnolinguistic appropriations traceable to the mid-nineteenth century. From the 1840s onward, Slovenian national activists occasionally referred to the territory as belonging to the broader Slovenian cultural sphere. Their claims rested primarily on linguistic grounds: in the widely accepted classification of languages, the dialects of the area were identified as Slovenian. Guided by the principles of ethnolinguistic nationalism, activists, intellectuals, and at times even state officials therefore categorized the Slavic-speaking population of the region as part of the Slovenian nation. Nevertheless, until the 1920s most of the local Slavophones did not identify themselves culturally or politically with the Slovenian nation.[33] Even so, linguistic assumptions and beliefs formed the core of the ethnographic evidence that Yugoslav diplomats presented at the Paris Peace Conference to justify both the Slovenian character of this area and the legitimacy of the Yugoslav occupation of Prekmurje.[34]

Immediately after the Yugoslav military occupation of Prekmurje in August 1919, numerous Slovenian officials and politicians crossed the former internal border between the Austrian and Hungarian parts of the empire and set foot in the region for the first time. What they encountered was a landscape and population largely unfamiliar to them. As former residents of the Austrian half of the monarchy, they had little understanding of the political, cultural, and socioeconomic conditions that prevailed on the Hungarian side. This was not surprising. In the final decades of Austria-Hungary, the Slovenian speaking society in Cisleithania possessed only rudimentary information about the region, provided by travelers or the few local correspondents for Slovenian newspapers. Yet once confronted with Prekmurje itself, Slovenian politicians and officials found a territory markedly different from most other Slovenian regions: one distinguished by fertile soil and abundant harvests. In the context of post-war shortages, poverty, and recurrent famine, these qualities offered both a justification

32 Kokolj, Prekmurski Slovenci, 11–33. See also Peter Štih et al., eds., *"Mi vsi živeti ščemo" : Prekmurje 1919: okoliščine, dogajanje, posledice : zbornik prispevkov mednarodnega in interdisciplinarnega posveta na Slovenski akademiji znanosti in umetnosti, Ljubljana, 29.-30. maj 2019* (Ljubljana: Slovenska akademija znanosti in umetnosti, 2020).

33 Jernej Kosi, "'However, the language here is changing gradually, and in the presence of so many local dialects the Croatian and its kindred Slovenian world cannot be separated very precisely': drawing the Slovenian-Croatian national border in the territory of the present-day Prekmurje region," *Prispevki za novejšo zgodovino* 57, No. 2 (2017): 33–50. Jernej Kosi, "The imagined Slovene nation and local categories of identification: 'Slovenes' in the Kingdom of Hungary and Postwar Prekmurje," *Austrian History Yearbook* 49 (2018): 87–102.

34 The most important contribution in this regard was the work of Matija Slavič, who participated in the Paris Peace Conference as the Yugoslav expert on Prekmurje. See Matija Slavič, *Prekmurje* (Slovenska krščansko-socialna zveza, 1921) and Matija Slavič, "Prekmurske meje v diplomaciji," in Vilko Novak, ed., *Slovenska krajina: zbornik ob petnajstletnici osvobojenja* (Ljubljana: Konzorcij, 1935).

and a means of substantiating Prekmurje's place within the broader Slovenian national framework. It was therefore not only possible but, from their perspective, necessary to reimagine Prekmurje as a newly acquired territory of strategic and symbolic value within the emerging South Slavic state.

Two official reports produced shortly after the Yugoslav occupation—one political, the other administrative—played a key role in shaping the vision of Prekmurje's fertile fields as a resource for the Slovenian national community. In October 1919, Albin Prepeluh, Minister of Welfare in the Regional Government for Slovenia, visited Prekmurje. In his report, which he prepared for discussion at a meeting of the Government, he described the situation in Prekmurje in detail. The very first topic he addressed in his report concerned the economic potential of Prekmurje. Prepeluh emphasized that for Slovenia, "Prekmurje is a major acquisition in economic terms. The fertile land there can serve as a regional breadbasket and thus compensate for the loss of many Slovenian villages elsewhere."[35] Accordingly, Prepeluh believed that Prekmurje should remain in Slovenian hands and that its food potential should be realized by improving communication links and implementing land reform.[36]

A month later, in November 1919, the itinerant agricultural teacher Franc Vojsk submitted a parallel report to the Regional Government in Ljubljana, reinforcing Prepeluh's view of Prekmurje as economically underdeveloped but agriculturally rich—an assessment that further anchored the emerging breadbasket narrative. On the one hand, he stressed that the region was almost entirely without industry and severely overpopulated in relation to its economic resources, forcing many inhabitants into emigration or seasonal work. On the other hand, he emphasized the fertility of the lowland soil, the favourable climate, and the region's capacity to produce abundant harvests. According to Vojsk, substantial resources were already available but failed to reach the Slovenian market; with improved agricultural education and railway links, he maintained, output could expand even further. Agriculture, he pointed out, was the region's most significant strength amid general underdevelopment and would enable Prekmurje to assume an important role in Slovenia's economic life, thereby reinforcing the emerging image of the region as the country's breadbasket.[37]

In their reports, Prepeluh and Vojsk articulated two perceptions of Prekmurje that would remain influential throughout the interwar period and beyond. On the one hand, they stressed the region's geography, fertile land, and evident economic potential; on the other, they emphasized its wider national significance, grounded in the abundance of grain and favourable conditions for farming. These views crystallized in the metonymy of the "breadbasket." By presenting Prekmurje as both agriculturally rich and nationally indispensable, such portrayals helped to legitimize its incorporation into the South Slavic state and to establish the idea of the "breadbasket of Slovenia" as a recurring theme in interwar journalistic and public discourse.

35 SI AS 60, box Prekmurje IV/V, map V (1919–1925), nr. 12943/1919.

36 Peter Ribnikar, ed., *Sejni zapisniki Narodne vlade Slovencev, Hrvatov in Srbov v Ljubljani in Deželnih vlad za Slovenijo: 1918–1921, vol. 2, od 28. feb. 1919 do 5. nov. 1919* (Ljubljana: Arhiv Republike Slovenije, 1999), 386.

37 SI AS 60, box Prekmurje V, No. 13535 (20 November 1919).

Prekmurje as the Slovenian Breadbasket: Explication and Popularization in Interwar Journalistic Discourse

Building on the views articulated by Prepeluh, journalists in Slovenian interwar magazines and newspapers reinforced the metonym of the "breadbasket of Slovenia" to depict Prekmurje as a land of exceptional fertility and agricultural abundance, thereby contributing to the popularization of the trope. Already in the earliest texts, writers claimed that Prekmurje was a "real breadbasket, where the fields yield in abundance."[38] Such depictions centered on the region's natural resources, believed to provide favorable conditions for extraordinary harvests. Some authors even compared the fertility of Prekmurje's arable land and its economic potential to that of Banat, long regarded as the gold standard for grain production on both sides of the First World War.[39] The flat lowland between the Mura/Mur River in the west and south and the surrounding hills in the north and east was described as "a single large breadbasket with enormous grain reserves" or as "a fertile field—a breadbasket," both formulations underscoring the exceptional agricultural potential attributed to Prekmurje.[40]

The authors did not justify the use of the term "breadbasket" to describe Prekmurje solely on the basis of general descriptions of agricultural abundance. They often referred to specific impressive quantities of agricultural produce that exceeded regional needs. According to an article published in the newspaper *Slovenski narod* in September 1919, it would be possible to export large quantities of grain from Prekmurje: up to 1,000 wagons of cereals per year.[41] In addition to grain, abundant fruit harvests, especially apples, would also be made available for trade, with more than 1,000 train cars of apples also available for transport.[42] Fruit was soon to become an "important export item," claimed the author of an article entitled "On the Economic Situation in Prekmurje" in 1924.[43] More than this, Prekmurje was also said to be a "land of poultry," where "thousands of geese, ducks, and other poultry" grazed in the meadows, while livestock farming was also "very well developed."[44] The term "breadbasket" did not have a monolithic meaning, as it could be used to describe very different aspects of Prekmurje's agricultural and socio-economic reality. It could refer to the entire region, where Prekmurje was described as "our breadbasket with its 100,000 souls," a term that encompassed not only the abundance of grain but also the broader agricultural wealth, including livestock, fruit growing, and beekeeping.

38 "Naše prekmurje," *Sokolič. List za sokolski naraščaj*, September 1919 (No. 5-6), 71.

39 "Prekmurje," *Mladost. Glasilo slovenskih orlov*, July/August/September 1919, 92, 93. "Narodno-gospodarski položaj Slovenskih Goric," *Trgovski list. Časopis za trgovino, industrijo in obrt*, 12 May 1921, 2.

40 "Zanimivosti iz Prekmurja," *Male novice*, 21 August 1919, 1. "Orlovski poročevalec," *Mladost, orlovsko glasilo*, 1920 (No. 6), 93.

41 "Gospodarske, kulturne in politične težnje muropoljskih Slovencev," *Slovenski narod*, 4 September 1919, 2–3.

42 "Zanimivosti iz Prekmurja," 1.

43 "O gospodarskih razmerah v Prekmurju," *Slovenec*, 26 November 1924, 5.

44 "Zanimivosti iz Prekmurja," 1.

More often, however, the term was applied specifically to the flat and fertile part of the region, distinguishing it from the hilly, economically less active, poorer, and more forested part of Prekmurje.[45]

In the interwar period, references to the breadbasket went beyond depictions of fertility to affirm Prekmurje's place within the Slovenian national community. The region was portrayed not just as a breadbasket, but as *our* breadbasket—the breadbasket of Slovenia. In the journal piece cited above, Prekmurje was not described only as a "real breadbasket, where the fields yield in abundance." It was depicted as much more than that:

> Prekmurje is a rich land. It is a real breadbasket, where the fields yield in abundance. Until now, these fields were in the hands of Hungarian magnates, and Slovenian farmers worked for foreigners. Now they belong to us, and Slovenian farmers will work here for their own benefit, and their harvest will richly reward their efforts.[46]

Interwar journalism was instrumental in naturalizing the image of Prekmurje as Slovenia's breadbasket, presenting it alternatively as an existing reality or projecting it as a future ideal. With the implementation of the land reform, Prekmurje's lowlands could become "a breadbasket for the less fertile parts of Slovenia."[47] This notion appeared repeatedly in the interwar Slovenian press. In September 1919, *Slovenski narod* emphasized that Prekmurje's agrarian resources and wealth were renowned far and wide and that it could "become a true breadbasket for Slovenia."[48] The following year, *Slovenec* described the region simply as "the breadbasket for Slovenia," while in 1921 the *Trgovski list* compared it to the Banat, calling it "a rich breadbasket of Slovenia."[49]

In the course of the interwar years, the breadbasket metonymy resurfaced periodically in the press, often tied to modernization and visions of Prekmurje's future. *Jugoslavija* stressed in early 1922 that "once Prekmurje gains a railway connection to Slovenia, it will become our breadbasket, and industry will also flourish there."[50] By the mid-1920s, the expression was already established: *Slovenec* remarked that "Prekmurje is often called the breadbasket of Slovenia," while *Narodni dnevnik* described "our fertile Prekmurje, which in due time may become the breadbasket of Slovenia."[51] The association continued into the late interwar years. *Neodvisnost* in 1937 claimed that "Prekmurje will be Slovenia's breadbasket. Therein lies its solution."[52]

45 "O gospodarskih razmerah v Prekmurju," 5.
46 "Naše Prekmurje," 71.
47 "Prekmurje," 92, 93.
48 "Gospodarske, kulturne in politične težnje," 2, 3.
49 "Izpraznitev Radgone in Prekmurje," *Slovenec*, 30 July 1920, 2. "Narodno-gospodarski položaj Slovenskih Goric," 2.
50 "Jugoslovanska kreditna banka," *Jugoslavija*, 8 January 1922, 3.
51 "O gospodarskih razmerah v Prekmurju," *Slovenec*, 26 November 1924, 5. "Prekmurska železnica," *Narodni dnevnik. Neodvisen političen list*, 6 March 1924, 2.
52 "Po Prekmurju," *Neodvisnost. Tednik za vsa javna vprašanja*, 13 March 1937, 4.

Shortly before the outbreak of the Second World War, *Večernik* highlighted the grain-producing districts of Murska Sobota and Lendava, concluding simply: "This is our breadbasket."[53]

Although not omnipresent, such references show that the metonymy of the breadbasket persisted in interwar discourse, where it served simultaneously as an economic claim, a political argument, and a symbol of national belonging that linked Prekmurje to the Slovenian national community. Its political resonance is evident in the words of Anton Korošec, leader of the Slovene People's Party (*Slovenska ljudska stranka*, SLS), who in July 1923—when his party was in opposition and promoting an autonomist program for Slovenia—stressed Prekmurje's importance for the Slovenian nation. In a speech reported in *Slovenec*, Korošec opposed Croatian politician Stjepan Radić's attempts to incorporate Prekmurje into a Croatian administrative framework, declaring: "We will not give away our Prekmurje, which will be our breadbasket, and which is and will remain Slovenian."[54] His statement illustrates how the metonymy of the breadbasket could be employed both as a practical and symbolic claim to secure Prekmurje's place within Slovenia.

The representation of Prekmurje as a breadbasket during the interwar period, however, was a symbolic designation with little grounding in the region's actual socio-economic conditions or levels of agricultural productivity. This characterization preceded material realities; only gradually would Prekmurje approach levels of output of national significance, let alone sustained surpluses beyond local consumption. Behind the image of abundance, many inhabitants lived at or below the subsistence threshold. By late 1919, administrative reports depicted acute shortages and hunger in the uplands: prices had spiked, essentials such as salt, sugar, matches, kerosene, and tobacco were scarce, clothing and footwear were in short supply, and many lacked even basic garments and adequate food. In 1920, approximately one-fifth of the population depended on state food relief, even as agricultural surpluses continued to be exported from the region. Even in years of favourable harvests, Prekmurje remained only marginally self-sufficient, with part of its grain production directed elsewhere. The discontinuation of customary seasonal labour arrangements—through which workers had traditionally been remunerated in grain—further exacerbated rural insecurity, pushing many households toward acute subsistence pressure and the risk of hunger. By the mid-1930s, contemporary surveys revealed persistently meagre and nutritionally inadequate diets, with most smallholdings unable to sustain household subsistence. Local accounts pointed to widespread deprivation—entire communities lacking bread for extended periods, children fainting from hunger, and families enduring prolonged shortages of basic foodstuffs—while health data indicated that the region suffered

53 "Raznolikost slovenještajerskega kmetijstva," *Večernik*, 4 August 1940, 11.
54 "Seja vodstva S.L.S. v Celju," *Slovenec*, 10 July 1923, 1.

from the highest overall and infant mortality rates, as well as the greatest prevalence of tuberculosis within the Drava Banovina. The breadbasket image thus masked deep structural vulnerability, transforming scarcity into a narrative of plenty.[55]

Conclusion

Since its invention in the post-imperial context of 1919, the notion of Prekmurje as Slovenia's breadbasket has endured across political, scholarly, and journalistic discourses. Recent parliamentary debates, the writings of Slovenian scientists and experts, and interwar journalistic representations of the region all testify to the lasting appeal of this designation. What began as an improvised figure of speech quickly developed into a recurring trope, invoked by individuals with different political and professional backgrounds, who nonetheless agreed in treating Prekmurje's agricultural wealth as nationally significant.

While grounded in empirical evidence of the region's agricultural abundance and productive potential, the breadbasket designation ultimately served as a symbolic vehicle for national integration. In the aftermath of the collapse of Austria-Hungary, Slovenian officials, experts, and journalists employed the image not only to describe fertile soil but to affirm Prekmurje's place in the newly established South Slavic state. By presenting the region as a source of nourishment "for us," the trope helped legitimize the incorporation of a contested borderland into the Slovenian national space.

The durability of this image is paradoxical, as it overlays Prekmurje's long-standing socio-economic fragility with a narrative of stability and abundance. Beneath depictions of impressive yields lay a persistent reality of scarcity, poverty, and economic underdevelopment that marked the region well into the late twentieth century. The breadbasket trope thus obscured structural vulnerabilities while highlighting agrarian productivity and agricultural abundance, binding Prekmurje to the Slovenian nation through an idealized vision that stood in stark contrast to the lived experience of its inhabitants. The discursive construction of Prekmurje as the Slovenian breadbasket therefore reveals how nationalist imagination endowed the post-imperial landgrab with ideological coherence and moral purpose — a vision that imagined Prekmurje as a breadbasket well before it became a region of genuine national importance in food production, with outputs exceeding local needs.

55 For additional discussion and further empirical evidence on food access and deprivation in interwar Prekmurje (including secondary literature), see Jernej Kosi, "'Yugoslavia has nothing. Yugoslavia has no bread. But Hungary gives us bread': Access to food and (dis)loyalty in a 'redeemed' Yugoslav borderland," *Austrian History Yearbook* 55 (2024): 283–97.

Acknowledgement

I gratefully acknowledge the financial support of the Slovenian Research and Innovation Agency (ARIS) (research core funding No. P6-0235 and project No. N6-0190, *Nourishing Victory: Food Supply and Post-Imperial Transition in Slovenia and the Czech Lands, 1918–1923*).

Sources and Literature

Archival sources

SI AS – Archives of the Republic of Slovenia:
SI AS 60 – Pokrajinska uprava za Slovenijo, Predsedstvo (1918–1924).

Literature

Centrih, Lev, and Polona Sitar. *Pol kmet, pol proletarec: integrirana kmečka ekonomija v socialistični Sloveniji, 1945–1991*. Koper: Založba Univerze na Primorskem, 2023.

Cunder, Tomaž. "Kmetijstvo v Pomurju danes in jutri." In *Pomurje [Elektronski vir]: trajnostni regionalni razvoj ob reki Muri: zbornik / 20. zborovanje slovenskih geografov, Ljutomer – Murska Sobota, 26.–28. marec 2009*, edited by Tatjana Kikec, 143–56. Ljubljana: Zveza geografov Slovenije; Društvo geografov Pomurja, 2009. https://www.drustvo-geografov-pomurja.si/projekti/zborovanje/Zbornik_geografov_POMURJE_2009.pdf.

Dešnik, Stanka. "Barve trideželnega parka." In *Vanekovo stoletje: Ob stoletnici rojstva dr. Vaneka Šiftarja*, edited by Darja Senčur Peček, 225. Murska Sobota: Univerzitetna založba Univerze, 2019.

Dobranić, Filip, Bojan Evkoski in Nikola Ljubešić. »Corpus of Slovenian periodicals (1771–1914) sPeriodika 1.0« (2023). *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1881.

Feiszt, György. "Revolucionarni pokret u Prekmurju od 1918. do 1919." In *Pomurje 1914–1920.: zbornik radova = Mura mente 1914–1920.*, edited by Branimir Bunjac. Povijesno društvo Međimurske županije, 2011.

Grgić, Ana, in Davor Nikolić. "'Ovaj grad zovu još i…' – o antonomazijama za toponime." *Folia onomastica Croatica* 23 (2014): 77–94.

Göncz, László. *A Muravidék útja a délszláv királyságba: a tájegység története az első világháború végétől a jugoszláv megszállásig (1918 ősze – 1919 augusztusa)*. Magyar Nyugat Könyvkiadó, 2024.

Göncz, László. "Utrinki iz zgodovine Beltincev v t. i. prevratnem obdobju (od oktobra 1918 do jugoslovanske zasedbe Prekmurja)." In *Raznolikost v raziskovanju etničnosti: izbrani pogledi III*, edited by Sonja Novak-Lukanovič and Barbara Riman. Ljubljana: Inštitut za narodnostna vprašanja, 2023.

Klemenčič, Marijan M., et al. *Življenjska (ne)moč obrobnih podeželskih območij v Sloveniji*. Ljubljana: Znanstvena založba Filozofske fakultete, 2018.

Kokolj, Miroslav. *Prekmurski Slovenci: od narodne osvoboditve do nacistične okupacije, 1919–1941*. Murska Sobota: Pomurska založba, 1984.

Kokolj, Miroslav. "Prekmurje v prevratnih letih 1918–1919." In *Revolucionarno vrenje v Pomurju v letih 1918–1920*, edited by Janko Liška. Murska Sobota: Pomurska založba, 1981.
Korpič-Horvat, Etelka. *Zaposlovanje in deagrarizacija pomurskega prebivalstva*. Murska Sobota: Pomurska založba, 1992.
Kosi, Jernej. "The Imagined Slovene Nation and Local Categories of Identification: 'Slovenes' in the Kingdom of Hungary and Postwar Prekmurje." *Austrian History Yearbook* 49 (2018): 87–102.
Kosi, Jernej. "'Yugoslavia Has Nothing. Yugoslavia Has No Bread. But Hungary Gives Us Bread': Access to Food and (Dis)Loyalty in a 'Redeemed' Yugoslav Borderland." *Austrian History Yearbook* 55 (2024): 283–97.
Kosi, Jernej. "'However, the language here is changing gradually...': Drawing the Slovenian-Croatian National Border in the Territory of the Present-Day Prekmurje Region." *Prispevki za novejšo zgodovino* 57, No. 2 (2017): 33–50.
Lorenčič, Aleksander. *Prelom s starim in začetek novega: Tranzicija slovenskega gospodarstva iz socializma v kapitalizem (1990–2004)*. Ljubljana: Inštitut za novejšo zgodovino, 2012. http://hdl.handle.net/11686/38023.
Luthar, Oto, ed. *Prekmurje za radovedneže in ljubitelje*. Ljubljana: Založba ZRC, ZRC SAZU, 2010.
"Mi vsi živeti ščemo": Prekmurje 1919: okoliščine, dogajanje, posledice, edited by Peter Štih et al. Ljubljana: SAZU, 2020.
Malačič, Janez. "Demografski razvoj v Prekmurju 1919–2019." In *'Mi vsi živeti ščemo': Prekmurje 1919: okoliščine, dogajanje, posledice*, edited by Peter Štih et al., 351–77. Ljubljana: SAZU, 2020.
Pančur, Andrej, Katja Meden, Tomaž Erjavec, Mihael Ojsteršek, Mojca Šorn, and Neja Blaj Hribar. *Slovenian Parliamentary Corpus (1990–2022) siParl 4.0*. Ljubljana: Institute of Contemporary History, 2024. http://hdl.handle.net/11356/1936.
Slavič, Matija. *Prekmurje*. Ljubljana: Slovenska krščansko-socialna zveza, 1921.
Slavič, Matija. "Prekmurske meje v diplomaciji." In *Slovenska krajina: zbornik ob petnajstletnici osvobojenja*, edited by Vilko Novak. Ljubljana: Konzorcij, 1935.
Slovenia: Landscapes and People, 3rd edition, edited by Drago Perko in Milan Orožen Adamič. Ljubljana: Mladinska knjiga, 2001.
Wraber, Maks. "Gozdna vegetacijska slika in gozdnogojitveni problemi Prekmurja." *Geografski vestnik* 23 (1951): 179–230.

Online sources

Državni zbor Republike Slovenije, www.dz-rs.si.
Merriam-Webster.com Dictionary, https://www.merriam-webster.com/dictionary/breadbasket.
Pleteršnik, Maks. *Slovensko-nemški slovar*, www.fran.si.
Slovar slovenskega knjižnega jezika, 2nd edition, supplemented and partly revised edition, www.fran.si.
Snoj, Marko. *Slovar Pohlinovega jezika*, www.fran.si.

Periodicals

Besednik. Kratkočasen in podučen list za slovensko ljudstvo, 1870.
Dolenjske novice, 1885.
Edinost, 1907, 1913.
Glasnik Avstrijske krščanske tobačne delavske zveze, 1911.
Gospodarski glasnik za Štajersko, 1912.
Jugoslavija, 1922.
Kmetijske in rokodelske novice, 1864, 1866, 1874, 1881, 1898, 1902
Kmetovalec, 1891, 1894.
Ljubljanski list, 1884.

Ljubljanski zvon, 1884.
Male novice, 1919.
Mladost. Glasilo slovenskih orlov, 1919.
Mladost, orlovsko glasilo, 1920.
Naš dom, 1905.
Narodni dnevnik. Neodvisen političen list, 1924.
Narodni gospodar, 1900, 1901, 1903, 1908.
Neodvisnost. Tednik za vsa javna vprašanja, 1937.
Pravi Slovenec, listi za podučenje naroda, 1849.
Slovenski gospodar, 1873, 1878, 1902.
Slovenec, 1909, 1911, 1912, 1920, 1923, 1924.
Slovenski narod, 1879, 1885, 1889, 1908, 1912, 1919.
Sokolič. List za sokolski naraščaj, 1919.
Trgovski list. Časopis za trgovino, industrijo in obrt, 1921.
Večernik, 1940.
Vedež: časopis za šolsko mladost, 1850.
Vertec, 1882.
Zgodnja danica, 1872.

Printed sources

Repe, Božo. "*Vsakdo mora imeti priliko, da udejstvi vse svoje telesne in duševne moči." Milko Brezigar in prvi slovenski program narodnega gospodarstva*. Ljubljana: Založba Univerze, 2023.

Sejni zapisniki Narodne vlade Slovencev, Hrvatov in Srbov v Ljubljani in Deželnih vlad za Slovenijo: 1918–1921, *vol. 2,* edited by Peter Ribnikar. Ljubljana: Arhiv Republike Slovenije, 1999.

Jernej Kosi

ŽITNICA SLOVENIJE: GENEALOGIJA METONIMIJE IN NJEN POMEN V PROCESU GRADNJE NACIJE

POVZETEK

Članek analizira nastanek, utrjevanje in pomensko delovanje označevalca »žitnica Slovenije« za Prekmurje ter pojasnjuje, zakaj se je podoba obilja utrdila navkljub dolgotrajni socioekonomski ranljivosti regije od konca 19. stoletja naprej. Izhodišče predstavlja analiza političnih poročil, administrativne dokumentacije in medvojnega časopisja, dopolnjena z interpretacijo strokovnih besedil, v katerih je termin dobil status implicitne analitične kategorije. V 19. stoletju se je izraz »žitnica« v slovenščini v prvi vrsti nanašal na prostore in postopke skladiščenja žita, preneseni pomen (»območje obilne pridelave žit«) pa je praviloma zadeval »neslovenska« območja (denimo Rusijo kot »žitnico Evrope«, Ogrsko, Banat ipd.). Prekmurje se

kot »slovenska žitnica« pojavi šele po jugoslovanski okupaciji leta 1919, ko je nova politična konfiguracija terjala simbolno utemeljitev vključitve nedavno okupiranega večjezičnega in večkulturnega mejnega območja v slovenski nacionalni imaginarij.

Začetke rabe metonimije strukturirata poročili Albina Prepeluha in Franca Vojska iz jeseni 1919. Še zlasti v Prepeluhovem poročilu, namenjenem upravnim telesom v Ljubljani, je bilo Prekmurje predstavljeno kot gospodarska »pridobitev« z izrazitim agrarnim potencialom, ki naj bi v preskrbovalnem smislu nadomestilo »izgubljene« ali manj rodovitne dele Slovenije. Takšno razumevanje je v dvajsetih in tridesetih letih razširil tudi časopisno-revijalni diskurz: Prekmurje je bilo opisano kot »naša žitnica«, pogosto v primerjavi z Banatom, z navajanjem presežkov žit, sadja, perutnine ter z napovedmi učinkov zemljiške reforme, prometne modernizacije in kmetijskega izobraževanja. Metonimija je delovala v dvojnem smislu: kot opis rodovitne nižinske krajine in kot sredstvo nacionalne integracije v kontekstu postimperialnega preoblikovanja meja.

Besedilo obenem poudarja razkorak med retoriko obilja in materialno realnostjo. Administrativna poročila in javnozdravstveni kazalniki za medvojno obdobje razkrivajo razdrobljeno posest, prikrito agrarno prenaseljenost, prehransko negotovost, sezonske in trajne migracije, skromne in prehransko neustrezne diete ter nadpovprečno umrljivost. Tudi po letu 1945 je regija kljub industrializaciji in deagrarizaciji ostala razvojno šibka, kar se je nadaljevalo v postsocialistični tranziciji; kriza po letu 2008 je dodatno razgalila strukturno ranljivost. Hkrati pa statistika potrjuje nadpovprečni agrarni pomen Pomurja v slovenskem kontekstu, kar je omogočalo vztrajanje diskurza o » slovenski žitnici«.

Prispevek pokaže, da je metonimija »žitnica Slovenije« delovala kot učinkovita simbolna tehnologija, ki je prepletala empirične prvine (rodovitna tla, ravninski relief, kontinentalno podnebje) z nacionalnimi in političnimi projekcijami. Vztrajnost izraza skozi različna politična obdobja – od medvojne Jugoslavije do sodobnih parlamentarnih debat – razkriva njegovo sposobnost prekrivanja odročnosti in socialne zapostavljenosti s predstavo o stabilnosti, samozadostnosti in »našosti«. Podoba Prekmurja kot »žitnice« je v resnici vzniknila še pred tem, ko je bil agrarni potencial Prekmurja, ki bi upravičeval rabo takega označevalca, sploh materializiran. To pa pomeni, da je bil nastanek metonimije »žitnica Slovenije« prej simbolni odgovor na postimperialno preurejanje prostora kakor pa odsev trajno presežnih agrarnih kapacitet. V tem smislu je metonimija ustvarila interpretativno ogrodje za vpis regije v slovenski nacionalni imaginarij.

1.02 DOI: https://doi.org/10.51663/pnz.65.3.12

Nik Obid*

Vsakdanji in banalni nacionalizem med strukturo in delovanjem

IZVLEČEK

Klasična delitev sociooloških pristopov ločuje med strukturalnimi teorijami in teorijami družbenega delovanja, ki glede na večjo vlogo akterja ali strukture razlagajo kompleksno funkcioniranje družbe. Omenjena delitev vpliva tudi na preučevanje nacionalizma, ki je največkrat obravnavan kot produkt (elitnih) strukturnih sil, ki prek različnih mehanizmov posredujejo ide(ologi)jo nacionalizma med prebivalstvo. Prispevek se osredotoča na povezanost sociooloških teorij, ki poudarjajo prepletenost vloge strukture in delovanja, s teorijama vsakdanjega in banalnega nacionalizma, ki temeljita na preučevanju nacionalizma in njegovega vpliva na ravni posameznih članov družbe.

Ključne besede: vsakdanji nacionalizem, banalni nacionalizem, struktura, delovanje, identiteta

ABSTRACT

EVERYDAY AND BANAL NATIONALISM BETWEEN STRUCTURE AND AGENCY

The classic division of sociological approaches distinguishes between structural theories and theories of social action, which explain society's complex functioning based on the primary role of either actors or structures. This division also influences the study of nationalism, which is

* **Mladi raziskovalec, asistent, Inštitut za slovensko izseljenstvo in migracije, ZRC SAZU, Novi trg 2, SI-1000 Ljubljana, nik.obid@zrc-sazu.si**

most often seen as a product of (elite) structural forces that disseminate the idea (or ideology) of nationalism among the population through various mechanisms. The contribution focuses on the relationship between the sociological theories, which emphasise the interconnected roles of structure and agency, and the theories of everyday and banal nationalism, which study nationalism and its influence on individual members of society.

Keywords: everyday nationalism, banal nationalism, structure, agency, identity

Uvod

Analizirati kulturne spore o nacionalni identiteti brez razumevanja, kako ljudje prevzamejo in živijo takšne identitete ter kako identiteta nato oblikuje njihovo ravnanje, nam ne pomaga pri razumevanju nacionalizma.[1]

Preučevanje nacionalizma se je dolga leta osredotočalo na analizo njegove strukturne narave in celostnega vpliva na delovanje družbe in posameznikov. To je povsem razumljivo, saj se z vplivom ideologije nacionalizma srečujemo na vseh poteh vsakdanjega življenja. Omejuje ali razširja nam pogled na svet, spreminja njegovo razumevanje, utrjuje ali razbija notranjo koherentnost družb, neti vojne ali spore ter krepi vključene ali ponižuje izključene iz nekega domnevno zaokroženega družbenega kroga. Gre namreč za zelo fluiden ideološki pojem, ki se na eni strani pogosto izkazuje kot izredno trdna podstat družbe, po drugi pa je filozofsko in sociološko izredno šibek ter na trenutke celo kontradiktoren in nekoherenten.[2] Ker je pojavnost naroda in nacionalne identitete kompleksna in variabilna[3], odvisna pa je od številnih družbenih dejavnikov, je težko oblikovati teoretični pristop, ki bi ga za(ob)jel v enotno sliko.

Prav zaradi kompleksnosti njegovega preučevanja so se izoblikovali različni raziskovalni pristopi, katerih cilj je razumevanje procesa usklajevanja politične skupnosti (države) s kulturno enoto oziroma narodom.[4] Raziskovalne paradigme, kot so modernizem, etnosimbolizem in evolucionizem, so bile največkrat povezane z analizo družbenih struktur, saj so se vpliva nacionalizma, njegove pojavnosti in ideološke moči lotevali od zgoraj navzdol.[5] Poudarjale so njegov izvor in politično-ideološke dimenzije, povezane z rastjo industrijskega kapitalizma ter oblikovanjem moderne nacionalne države, ali pa so ga preučevale kot kulturni konstrukt kolektivne pripadnosti,

1 Stephen Reicher in Nick Hopkins, *Self and Nation: Categorization, Contestation and Mobilization* (Thousand Oakes: SAGE, 2001), 3.

2 Damjan Mandelc, *Na mejah nacije: teorije in prakse nacionalizma* (Ljubljana: Filozofska fakulteta, 2011), 45.

3 Anthony Smith, *National Identity* (London: Penguin, 1991), 143.

4 Jon E. Fox in Cynthia Miller-Idriss, »Everday nationhood,« *Ethnicities* 8, št. 4 (2008): 536, pridobljeno 5. 6. 2025, https://doi.org/10.1177/1468796808088925.

5 Jon E. Fox in Maarten Van Ginderachter, »Everday nationalism's evidence problem,« *Nations and Nationalism* 24, št. 3 (2018): 547, pridobljeno 5. 6. 2025, https://doi.org/10.1111/nana.12418.

uresničen in legitimiran skozi institucionalne in diskurzivne prakse.[6] Osrediščene so (bile) predvsem okoli načina, »kako se narodi oblikujejo, niso pa zmogle razložiti tega, kako ljudje oblikujejo narod«.[7] A v kompleksnem sodobnem svetu to nikakor ne zadošča, saj celostne analize nacionalizma ne more biti brez preučevanja uveljavljenih vsakdanjih praks in navad, prek katerih ide(ologi)ja prodira v življenja navadnih ljudi, ter njihovega vsakodnevnega osmišljanja nacionalnih kategorij. Zato bi lahko trdili, da analiza nacionalizma sledi klasični dilemi sociološke znanosti, ki poskuša s konfliktom med strukturalnimi teorijami in teorijami družbenega delovanja zajeti kompleksnost družbe kot celote.[8] Od Billigove znamenite knjige o banalnem nacionalizmu[9] dalje je vse bolj jasno, da temeljite analize ne more biti brez razumevanja njegove (re)produkcije na ravni posameznih članov družbe.

V nasprotju s klasičnimi teorijami nacionalizma banalni in vsakdanji nacionalizem poudarjata drugačne raziskovalne perspektive, pri čemer se v raziskovalnih izhodiščih temeljno razlikujeta. Vsakdanji nacionalizem (pri nekaterih avtorjih tudi vsakdanja narodnost – »everyday nationhood«)[10] je osredotočen na preučevanje posameznikovega razumevanja, uporabe in predvsem *(so)oblikovanja* nacionalnih kategorij v vsakdanjih življenjskih praksah, pogovorih in potrošniških navadah.[11] Raziskovalci banalnega nacionalizma po drugi strani analizirajo, kako nacionalizem v vsakdanje življenje prodira večinoma neopazno prek prevladujočih diskurzov (npr. tistih, ki jih posredujejo mediji, institucije, politični govori, šolski kurikulumi, rituali, dogodki ali simboli) ter tako na impliciten način reproducira idejo naroda in nacionalne države.[12] Kot trdi Fox, je bistvena razlika v raziskovalnem izhodišču, saj se banalni nacionalizem osredotoča na ideologijo nacionalizma in njegovo prodiranje od zgoraj navzdol, medtem ko so za raziskovalce vsakdanjega nacionalizma osnova preučevanja navadni ljudje, ki počnejo vsakdanje stvari.[13] Pri obeh pristopih je bistveno preučevanje nacionalizma v kontekstu situacij na ravni akterja, kjer ta zavedno ali nezavedno upošteva nacionalne kategorije in deluje v njihovem okviru oziroma se nanje sklicuje in jih uporablja za lastne namene.[14]

Razumevanje teh procesov je izredno pomembno v času pospešene globalizacije in razvoja sodobnih individualističnih družb, kjer se vprašanja o (samo)ohranjanju narodov in nacionalizma pozornemu opazovalcu pojavljajo vsakodnevno. Pri obravnavi

6 Fox in Miller-Idriss, »Everday nationhood,« 536. Mandelc, *Na mejah nacije*, 45, 46.

7 Fox in Ginderachter, »Everday nationalism's evidence problem,« 1.

8 Haralambos in Holborn, *Sociologija: teme in pogledi* (Ljubljana: Državna založba Slovenije, 1999), 874, 875.

9 Billig, *Banal nationalism* (London: SAGE Publications, 1995).

10 Fox in Miller-Idriss, »Everyday nationhood.«

11 Fox in Ginderachter, »Everday nationalism's evidence problem,« 1, 2.

12 Michael Skey, »The national in everday life: A critical engagement with Michael Billig's thesis of Banal Nationalism,« *The Sociological Review* 57, št. 2 (2009): 331, 332, pridobljeno 7. 6. 2025, https://doi.org/10.1111/j.1467-954X.2009.01832.x.

13 Jon E. Fox, »Banal nationalism in everyday life,« *Nations and Nationalism* 24, št. 4 (2018): 863, pridobljeno 7. 6. 2025, doi: 10.1111/nana.12458.

14 Jonathan Hearn in Marco Antonsich, »Theoretical and methodological considerations for the study of banal and everyday nationalism,« *Nations and Nationalism* 24, št. 3 (2018): 1, 2, pridobljeno 7. 6. 2025, https://doi.org/10.1111/nana.12419.

obeh pristopov vznikajo številne teoretske in metodološke dileme, pa tudi vprašanja o objektih raziskovanja in znanstvenem doprinosu k študijam nacionalizma.[15] V članku vsakdanji in banalni nacionalizem teoretsko sociološko kontekstualiziram in povežem s temeljnimi deli ter izbranimi članki, ki obe paradigmi implementirajo v praksi. V prvem delu predstavim izbrane sociološke teorije, ki poskušajo preseči strogo delitev med strukturo in delovanjem, pri čemer sem še posebno pozoren na vidike, pomembne za razumevanje reprodukcije nacionalizma na ravni posameznih članov družbe. V drugem delu članka obravnavane sociološke teorije, ki temeljijo na preučevanju odnosa, kompleksnosti in prepletenosti strukture in delovanja, povežem s teoretičnimi dilemami obeh pristopov v sodobni in historični perspektivi ter s tem prikažem pomembnost omenjenih teorij za preučevanje kompleksne narave nacionalizma in njegovih praktičnih vsakdanjih dimenzij.

Teoretični sociološki premisleki

Preučevanje nacionalizma, ki predstavlja eno izmed prevladujočih ideologij v modernem času, obsega najrazličnejše družboslovne premisleke in metodološke pristope. Razprave pogosto zajemajo zgodovinske in politološke raziskovalne pristope, še zanimivejše pa so sociološke obravnave, ki raziskujejo, kako se preučevanja nacionalizma lotiti z družbeno-analitične perspektive. Slovenski sociolog dr. Rudi Rizman kot bistvene našteva štiri sklope njegovega preučevanja: komunikacijski, marksistični, psihološki in funkcionalni.[16] Prvi obsega analizo nacionalizma s teorijo sistema notranjih komunikacij, pri čemer slednja zagotavlja občutek skupne identitete, medtem ko marksistični ponuja premislek o njem skozi prizmo razrednega konflikta oziroma ekonomskih konfliktov v družbi. Psihološki analitični pristop nacionalizem razume predvsem na podlagi istovetenja s širšimi družbenimi cilji, ki so pogosto posledica socialno-ekonomskih in političnih sprememb, funkcionalni pristop pa zajema temeljne premisleke o nacionalizmu kot ideološki podstati prehoda med tradicionalnimi in modernimi identitetami.[17]

Temeljna ideja sodobnih socioloških teorij je predvsem preusmeritev raziskovalnega fokusa z enega na več različnih vidikov družbenega življenja, ki so med

15 Eleanor Knott, »Everyday nationalism. A review of the literature,« *Studies of National Movements* 3 (2015), pridobljeno 8. 6. 2025, https://test.snm.nise.eu/index.php/studies/article/view/0308s. J. Paul Goode in David R. Stroup, »Everyday nationalism: constructivism for the masses,« *Social Sciences Quarterly* 96, št. 3 (2015), pridobljeno 8. 6. 2025, doi: 10.1111/ssqu.12188. Anthony Smith, »The limits of everyday nationhood,« *Ethnicities* 8, št. 4 (2018), pridobljeno 8. 6. 2025, https://doi.org/10.1177/1468796808080040102. Jon E. Fox in Cynthia Miller Idriss, »The 'here and now' of everyday nationhood,« *Ethnicities* 8, št. 4 (2018): 573–76, pridobljeno 11. 6. 2025, https://doi.org/10.1177/1468796808088925. Sophie Duchesne, »Who is afraid of banal nationalism?,« *Nations and Nationalism* 24, št. 4 (2018), pridobljeno 9. 6. 2025, doi: 10.1111/nana.12457. Hearn in Antonsich, »Theoretical and methodological considerations.« Fox in Ginderachter, »Everday nationalism's evidence problem.« Fox, »Banal nationalism in everyday life.« Fox in Miller-Idriss, »Everday nationhood.«

16 Rudi Rizman, ur., *Študije o etnonacionalizmu* (Ljubljana: KRT, 1991), 27–31.

17 Mandelc, *Na mejah nacije*, 77.

seboj povezani.[18] Temu sledita tudi koncepta preučevanja banalnih in vsakdanjih vidikov nacionalizma, ki poskušata (strukturno) pogojene nacionalne diskurze elit[19] razumeti v dinamiki posameznikovega vsakdanjega življenja.[20] Koncepti sociologov, ki so poskušali preseči klasično delitev med strukturo in delovanjem, so močno vplivali tudi na analizo samega nacionalizma. Za sociologijo istočasna pomembnost mikro in makro pristopov, ki sta ju v svojih delih analizirala ameriška sociologa Jeffrey C. Alexander in Randall Collins[21], je v marsikaterem vidiku zelo podobna teoretičnim konceptom sociologov, ki poudarjajo pomen povezanosti strukturnih dejavnikov družbe in akterjevega (so)ustvarjanja družbenega sveta, med katere štejemo predvsem Pierra Bourdieuja, Margaret Archer in Anthonyja Giddensa.[22] To je še posebej uporabno pri preučevanju nacionalizma, pri katerem mora družboslovna znanost presegati osredotočanje zgolj na njegov strukturni vpliv, kar je poudaril že znameniti zgodovinar Eric Hobsbawm, ki je opozoril na nujnost obravnave dualnosti nacionalizma, saj ga naj ne bi bilo mogoče razumeti brez analize mikroravni.[23]

Kot prvega med pomembnejšimi predstavniki takih pristopov lahko omenimo znanega francoskega sociologa Pierra Bourdieuja, ki je poskušal v svojem delu *Outline of a Theory of Practice* premostiti dilemo med vplivom struktur ali delovanjem akterja na družbeno dinamiko.[24] V ta namen je razvil termina »doxa« in »habitus«, s katerima je pojasnjeval akterjevo zavedno in nezavedno delovanje kot posledico socializacije in vpliva zunanjih družbenih struktur,[25] a ob tem ohranil njegovo moč pri njihovi spremembi. S tem je močno povezana tudi njegova analiza akterjevega delovanja na družbenem prizorišču – to naj bi bilo odvisno od različnih pridobljenih oblik kapitala (ekonomskega, kulturnega, družbenega in simbolnega), ki jih posameznik uporabi na polju, kjer zaseda svoj položaj.[26] Z omenjenimi teoretičnimi koncepti je Bourdieu poskušal preseči klasično sociološko past med pretiranim poudarjanjem strukture ali delovanja v družbeni dinamiki, zaradi česar je njegova teorija vsekakor zelo zanimiva za preučevanje vpliva nacionalizma na vsakdanji ravni. Če vemo, da sta banalni in vsakdanji nacionalizem zelo povezana z rutinami, ki jih akter ponavlja v svojem vsakdanjem življenju ter pogosto reproducira v situacijah, v katerih nezavedno ve, kako ravnati,[27] predstavljajo takšni koncepti izredno pomembno raziskovalno osnovo.

18 Ibidem.

19 Z elitami v primeru nacionalizma poimenujem predvsem družbene skupine, ki posedujejo lastnosti, ki jih uvrščajo višje na družbeni lestvici. Lastnosti vključujejo politično, administrativno ali versko moč, višjo izobrazbo, premožnost, družbeni položaj ali slavo. Pomembna je tudi njihova umeščenost v geografski (državni, regionalni ali lokalni) kontekst. – Več o tem Joseph Whitmeyer, »Elites and popular nationalism,« *British Journal of Sociology* 53, št. 3 (2002): 322, pridobljeno 12. 7. 2025, https://doi.org/10.1080/0007131022000000536.

20 Hearn in Antonsich, »Theoretical and methodological considerations,« 1.

21 Mandelc, *Na mejah nacije*, 78.

22 Pip Jones in Liz Bradbury, *Indroducing Social Theory* (Cambridge: Polity Press, 2018), pogl. 7, pridobljeno 10. 6. 2025, https://research-ebsco-com.nukweb.nuk.uni-lj.si/c/rsg6t3/search/details/ecieo34n3v?db=nlebk.

23 Eric Hobsbawm, *Nations and Nationalism since 1780: Programme, Myth, Reality* (Cambridge: Cambridge University Press, 1992), 10.

24 Pierre Bourdieu, *Outline of a Theory of Practice* (Cambridge: Cambridge University Press, 1997).

25 Ibid.

26 Jones in Bradbury, *Introducing Social Theory*, pogl. 7.

27 Tim Edensor in Shanti Sumartojo, »Geographies of everyday nationhood: experiencing multiculturalism in Melbourne,« *Nations and Nationalism* 24, št. 3 (2018): 555, pridobljeno 12. 6. 2025, doi: 10.1111/nana.12421.

Pri sociološki analizi obeh fenomenov moramo zaradi obravnave akterjevega potenciala pri spreminjanju struktur in vplivanju nanje pod drobnogled vzeti tudi britanska raziskovalca Anthonyja Giddensa in Margaret Archer. Giddens je s svojo teorijo strukturacije poskušal omenjeno dvojnost teoretično preseči z izpostavljanjem dualnosti struktur in njihovega (re)produciranja.[28] Pri tem večkrat – tako kot Bourdieu – poudari refleksivno sposobnost akterja, čigar praktično in teoretično znanje izredno vpliva na celoten proces, saj naj bi ta svoj položaj v družbi pogosto zavestno analiziral. Po njegovi teoriji je obstoj strukture tesno povezan z delovanjem akterja, saj brez njegovega delovanja v praksi ne more obstajati.[29] Takšno delovanje naj bi bilo v prvi vrsti povezano z akterjevim diskurzivnim in praktičnim znanjem, ki ju Giddens razdeli glede na posameznikovo (ne)refleksivno delovanje v določeni situaciji. Praktično znanje se izkazuje kot ravnanje, ki je tiho v delovanju, a močno prisotno v številnih vsakodnevnih situacijah in navadah, ki jih jemljemo za samoumevne, pri njih pa ne prihaja do vsakokratne ponovne refleksije.[30] Drugi vidik predstavlja diskurzivno racionalno znanje, ki ga akter uporabi takrat, ko potrebuje refleksivno utemeljitev svojih dejanj, pri čemer se opira na redno spremljanje svojih motivov in delovanja.[31] Razlika med obema je predvsem v situacijskih kontekstih, saj praktično znanje omogoča vsakodnevno delovanje, ki pogosto temelji na ponotranjenih vzorcih obnašanja v družbi, medtem ko ima diskurzivno znanje posameznika moč spoprijemanja z novimi nepredvidljivimi situacijami in iskanjem rešitev, prek katerih se spreminjajo tudi izvorna strukturna razmerja.[32]

Vidik strukture sestavljajo pravila, ki so za posameznika omogočujoča ali omejujoča, in resursi, ki jih Giddens razdeli na alokacijske in avtoritativne, povezani pa so predvsem s posedovanjem določenih (materialnih) dobrin ali zmožnostjo vpliva na druge člane družbe.[33] V praksi to pomeni, da sta posameznikovo delovanje in struktura medsebojno soodvisna in neločljiva, kar nazorno prikaže na primeru jezika, istočasno delujočega kot (omejujoča) struktura, ki se z govorjenjem po določenih pravilih reproducira, a se sčasoma, zaradi aktivnosti akterjev in posedovanja obeh oblik virov, tudi spreminja. Strukture torej po Giddensovi teoriji niso nekaj zunanjega in deterministično vplivnega od zgoraj navzdol, temveč predvsem rezultat delovanja in reprodukcije posameznih članov družbe na mikroravni.

Teorija strukturacije še zdaleč ni popolna, zaradi česar je bila deležna številnih kritik, istočasno pa se je na primeru kompleksnosti migracij in nacionalizma tudi nadgrajevala. Teorijo so številni raziskovalci še naprej razvijali, družbene strukture pa

28 Anthony Giddens je svojo teorijo strukturacije razvijal v različnih strokovnih delih: Anthony Giddens, *New Rules of Sociological Method* (London: Hutchinson, 1976). Anthony Giddens, *Studies in Social and Political Theory* (New York: Basic Books, 1977). Anthony Giddens, *Central Problems in Social Theory* (London: MacMillan, 1979). Anthony Giddens, *The Constitution of Society: Outline of the Theory of Structuration* (Cambridge: Polity Press, 1984).

29 Haralambos in Holborn, *Sociologija*, 911.

30 Peter Stankovič, »Giddensova teorija strukturacije: zagate teoretskega eklekticizma,« *Teorija in praksa* 37, št. 3 (2000): 459, pridobljeno 11. 6. 2025, http://dk.fdv.uni-lj.si/db/pdfs/tip20003stankovic.pdf. Jones in Bradbury, Introducing social theory, pogl. 7.

31 Jones in Bradbury, *Introducing Social Theory*, pogl. 7.

32 Stankovič, »Giddensova teorija strukturacije«, 459.

33 Giddens, *The Constitution of Society*, 33.

delili na zunanje strukturne sloje, ki so jih razdelili na zgornje oddaljene strukturne sloje (velike zgodovinske sile, zunanje pogoje in globalne družbene spremembe, kot so npr. politični in gospodarski sistemi, vojna, lakota)[34] in bližnje strukturne sloje (kontekstualno in področno specifične omejitve, kot so denimo pravila, zakoni, politike, organizacijski okviri ipd.).[35] Poleg zunanjih so h kompleksni analizi družbenega sveta dodali tudi koncept t. i. notranjih struktur, pri čemer so pogosto uporabili že omenjeno Bourdieujevo teorijo habitusa in koncept konjunkcijsko specifičnih internih struktur. Konjunkcijsko specifične interne strukture predstavljajo tisti vidik, ki ga najlažje razložimo s situacijsko priučenostjo oziroma procesom spoznavanja, načini ponotranjenega razmišljanja in delovanja v danem kontekstu.[36]

Kompleksnost takih modelov (raz)ločevanja med strukturo in delovanjem ima seveda tako prednosti kot slabosti, a je pri preučevanju (vsakdanjega in banalnega) nacionalizma izredno pomembna, saj se s pomočjo omenjenih kategorij lažje razume vpliv ideologije nacionalizma na različnih ravneh družbe. Če razvejanost strukturnih slojev konkretiziramo na primeru bivanja migrantov v tujem okolju, slednje ni samo produkt zunanjih makrostruktur ali akterjevega delovanja, temveč časovno in prostorsko specifičen kontekst obojega.[37] Koncept namreč predvideva pomembno soodvisnost akterjevega delovanja na podlagi socialno-kulturnih virov, ti pa so vedno odvisni od vpliva različnih struktur, ki podpirajo to delovanje.[38] V praksi bi to pomenilo, da je akterjevo delovanje v vsakdanjem življenju tesno prepleteno z ekonomskimi dinamikami, migracijskimi politikami, kulturnimi vzorci ipd., ki sčasoma vplivajo na njegovo orientacijo v družbi, medtem ko se ta spreminja glede na njegov ekonomski status, kulturni in družbeni kapital ter politični položaj v državi bivanja.[39] Kot že omenjeno, strukturacija pri tem predvideva tudi obraten proces transformativne narave teh struktur, ki se zaradi akterja na dolgi rok spreminjajo.[40]

Prav omenjeni zapletenost in prepletenost strukture in delovanja pri nekaterih sociologih nista bili najbolje sprejeti. Kritika Giddensa je največkrat povezana prav s prevelikim poudarkom na akterjevi zmožnosti spreminjanja strukture, čemur je najbolj nasprotovala britanska sociologinja Margaret Archer.[41] Ideje, da lahko akter v vsaki

34 Rob Stones, *Structuration Theory* (Basingstoke: Palgrave Macmillan, 2005). Orla McGarry, »Knowing 'how to go on': structuration theory as an analytical prism in studies of intercultural engagement,« *Journal of Ethnic and Migration Studies* [42], št. 12 (2016): 2071, 2072, pridobljeno 13. 6. 2025, https://doi.org/10.1080/1369183X.2016.1148593.

35 Karen O'Reilly, »Structuration, practice theory, etnography and migration: bringing it all together,« *IMI Working paper series* 61 (2012): 7–9, pridobljeno 15. 6. 2025, https://ora.ox.ac.uk/objects/uuid:f7ffb7f9-d6d0-4601-95e1-156da3c714a3.

36 Ibid.

37 Ewa Morawska, »International migration: its various mechanisms and different theories that try to explain it,« *IMER/MIM* (2007), 13, pridobljeno 16. 6. 2025, https://www.diva-portal.org/smash/get/diva2:1409965/FULLTEXT01.pdf.

38 McGarry, »Knowing 'how to go on',« 2071, 2072.

39 Morawska, »International migration,« [13].

40 Ibid.

41 Margaret Archer, *Realist Social Theory: The Morphogenetic Approach* (Cambridge: Cambridge University Press, 1995). Margaret Archer, »Morphogenesis versus structuration: on combining structure and action,« *British Journal of Sociology*, 33, št. 4 (1982): 455–83, pridobljeno 18. 6. 2025, https://doi.org/10.2307/589357.

situaciji ravna drugače in spreminja strukture zgolj s spreminjanjem lastnega vedenja, so lahko problematične predvsem z vidika nezadostne razlage načinov, kako strukture ovirajo akterje pri njihovem delovanju.[42] Ker Giddens pri analizi družbene dinamike nikakor ni mogel spregledati nekaterih eksternih strukturnih lastnosti družbe, ki se s konceptom prepletenosti niso mogle prepričljivo skladati, je zato izoblikoval nedorečen teoretski kompromis, ki pri njegovih kritikih še zdaleč ni ostal spregledan.[43] Analitični dualizem Margaret Archer je ločnico med strukturo in delovanjem vsekakor postavil nekoliko ostreje. Čeprav njene teorije morfogeneze nikakor ne moremo klasificirati kot funkcionalistične, pa predpostavlja zgolj omejeno zmožnost akterjev pri spreminjanju struktur, saj naj bi bile nekatere strukturne značilnosti družbe izven njihovega nadzora.[44] Strukture tako avtorica razdeli na družbene (npr. materialni viri) in kulturne (npr. znanje), te pa posedujejo nastajajoče lastnosti (angl. *emergent properties*), ki jih ni mogoče zreducirati na posamezne člene družbe, a imajo vzročno moč in so relativno dolgotrajne.[45] Kot take so strukture lahko avtonomne in samostojne, saj temeljijo na nekaterih strukturnih predpogojih, vendar so v nekaterih razmerah do določene mere še vedno spremenljive. Pri analitičnem dualizmu je Archer razlikovala med načini, kako je družba preoblikovana ali reproducirana, s čimer je poskušala uvesti metodologijo, ki bi na analitični ravni strukture in delovanje razločevala, hkrati pa dopuščala dvosmerno možnost spremembe.[46] Pri tem je pod drobnogled še posebej vzela refleksivnost akterja, ki kljub omejenemu vplivu na strukture konstantno analizira in ocenjuje svoj položaj v razmerju do naravnega, praktičnega in družbenega reda ter v njihovem okviru na podlagi svojih izkušenj izbira situacijsko pravo pot.[47]

Naštete teorije, na trenutke prepletajoče se ali izključujoče, so za sociološko analiziranje nacionalizma ključne, saj je prav vpliv njegove ideološke (strukturne) narave na raven akterja izredno pomemben za razumevanje njegove idejne moči. Pri tem je osredotočanje na en vidik lahko težavno, saj predpostavljanje enosmernega razumevanja nacionalnih idej od zgoraj navzdol zanemarja njegovo dojemanje, konstruiranje in reproduciranje na ravni posameznika. Narodi in nacije[48] namreč še zdaleč niso samo produkt večjih strukturnih sil, temveč tudi rezultat praktičnega delovanja ljudi v vsakdanjem življenju.[49] Idejo o prenosu raziskovalnega fokusa z višje na nižjo raven zasledimo že pri Michaelu Billigu, pionirju ideje o banalnem nacionalizmu, ki

42 Archer, »Morphostasis versus structuration«, 459, 461, 462.

43 Stankovič, »Giddensova teorija strukturcije,« 463.

44 Haralambos in Holborn, *Sociologija*, 914.

45 Jones in Bradbury, *Introducing Social Theory*, pogl. 7.

46 Ibid.

47 Archer, »Realism and the problem of agency,« *Journal of Critical Realism* 5, št. 1 (2015): 16, 17, pridobljeno 17. 6. 2025, https://doi.org/10.1558/aleth.v5i1.11.

48 Narod in nacija sta v tem primeru skupaj omenjena zaradi neskladnosti izraza »nation« v angleškem in slovenskem jeziku. Čeprav se v slovenščini besedi pomensko razlikujeta, pa je za razumevanje nacionalizma kot »ideološkega gibanja, ki si prizadeva za pridobitev ali ohranitev avtonomije, enotnosti in identitete obstoječega ali potencialnega naroda« (Smith, 1989, v Mandelc, 2011, 24), v kontekstu pričujočega pregleda literature pomembno predvsem osredotočanje na njun družbeni pomen in ne toliko na konceptualno (etnično ali politično) dimenzijo. – Več o kompleksnosti izrazoslovja pri razumevanju nacije, naroda in nacionalizma v Mandelc, *Na mejah nacije*, 13–27.

49 Jon E. Fox in Cynthia Miller-Idriss, »Everyday nationhood,« 554.

je pod drobnogled vzel banalne oblike njegove reprodukcije na ravni posameznika.[50] Prav zato je smiselno, da mu pri analizi damo prednost pred kasnejšimi deli, ki so z njegovim polemizirala ter ga vsebinsko in tematsko nadgradila.

Vsakdanji in banalni nacionalizem med strukturo in delovanjem

Knjiga *Banalni nacionalizem* (1995) Michaela Billiga predstavlja velik doprinos k razumevanju nacionalizma kot strukturno oblikovane ideologije in njenega vpliva na nezavedno delovanje posameznika v prostoru in času. Dotedanjim strategijam preučevanja nacionalizma Billig pripiše pretirano poudarjanje vpliva strukturno pogojenih sil, povezanih s političnim delovanjem, pogosto nasilnimi poskusi doseganja nacionalne suverenosti in posledično emocionalnim odzivom množic.[51] Takšni pristopi naj bi zato vodili do nenamernega zanemarjanja vprašanj o tem, kako se nacionalne kategorije ohranjajo in vzdržujejo v ustaljenih nacionalnih državnih okvirih, saj tudi v (stabilnih) državah, kjer t. i. vroči nacionalizem ni tako močan, nacionalna identiteta ostaja ena izmed pomembnejših družbenih in osebnih opredelitev. Kot v svojem delu navaja Billig, nacionalne identitete ne smemo dojemati samo kot notranjega psihološkega stanja ali individualne samodefinicije,[52] temveč kot »način življenja, ki se dnevno živi v svetu nacionalnih držav«.[53] To je tudi eden izmed glavnih razlogov, zakaj prebivalci nacionalne identitete ne pozabimo v obdobjih, ko nas nanje ne opominjajo posebne priložnosti, kot so denimo državni prazniki ali športne prireditve. Billig se osredotoča na banalne označevalce nacionalne identitete, ki bistveno prispevajo k vsakodnevni reprodukciji nacionalnih kategorij, saj sčasoma postanejo samoumevni. Kot vzorci, prakse in navade so vgrajeni v materialno in socialno okolje posameznika, ki jih sprejema zaradi izpolnjevanja določenih psiholoških potreb in osebnega prilagajanja socialnemu okolju, ki te kategorije ohranja resnične in nujne.[54]

Billig je pri konstrukciji nacionalnosti še posebno pozoren na diskurzivne prakse, ki nacionalno identiteto ohranjajo prek prisotnosti nacionalnih simbolov (npr. izobešanja zastav, motivov na kovancih ipd.) ali političnih govorov in množičnih medijev, kjer se s pomočjo manjših besednih zvez (npr. »mi«, »naš«, »nam«, »tukaj«) ohranja predstava o zaokroženi (narodni) skupnosti.[55] Iz tega lahko zaznamo nekatere podobnosti s prej analizirano Bourdieujevo teorijo habitusa ali Giddensovim praktičnim znanjem, saj narodnost v tem primeru deluje kot nezavedna osnova posameznikovih

50 Billig, *Banal Nationalism*. Andrew Thompson, »Nations, national identities and human agency: putting people back into nations,« *The Sociological Review* 49, št. 1 (2001): 28, pridobljeno 23. 6. 2025, https://doi.org/10.1111/1467-954X.00242.

51 Billig, *Banal Nationalism*.

52 Ibid., 69.

53 Ibid., 68.

54 Jonathan Hearn, »National identity: banal, personal and embedded,« *Nations and Nationalism* 13, št. 4 (2007): 660, pridobljeno 21. 6. 2025, https://doi.org/10.1111/j.1469-8129.2007.00303.x.

55 Billig, *Banal Nationalism*, 105.

izbir in delovanj,[56] ki jih pridobimo med socializacijo. Vzdrževalci nacionalne ideologije na banalni ravni so torej samoumevne prakse, ki idejo ohranjajo prisotno na neopazne, nevidne in nezaznavne načine. Takšna ponotranjenost pa je mogoča samo, če nacionalnost postane del človekovega habitusa, ki je sestavljen tako iz spominjanja kot tudi pozabljanja elementov vsakdana, kar posamezniku omogoča funkcioniranje v ponavljajočih se dnevnih rutinah.[57]

Seveda takšna prisotnost ide(ologi)je ne bi bila mogoča brez strukturnega vpliva elitnih nacionalnih diskurzov. Prav zato je kritika Billiga zelo podobna kritiki Bourdieuja, ki mu je bilo očitano, da se je pri svojem delu pretirano približal strukturalnim pristopom in neustrezno naslovil proces akterjevega spreminjanja ali reprodukcije struktur družbenega življenja.[58] To naj bi bila tudi glavna pomanjkljivost teorije o banalnem nacionalizmu, ki večinoma obravnava strukturne vplive nacionalizma na vsakodnevno raven posameznika, ob tem pa zanemarja njegov doprinos.[59] Kot poudarja Antonsich, akterjevo vlogo v *Banalnem nacionalizmu* ovira predvsem njegova nerefleksivnost pri ideološki reprodukciji, saj Billig nacionalizem razume kot strategijo in pogoje, ki posameznika ovirajo, pri tem pa predpostavlja »nerealistično predstavo o enotni in homogeni nacionalni publiki«.[60]

Da prebivalci neke države nikakor niso homogena gmota, ki brez premisleka absorbira nacionalni diskurz, temveč prej heterogena zmes različno dojemajočih in razmišljujočih posameznikov z lastnim mnenjem, nakazujejo tudi nekatera dela, ki so se z vplivi nacionalizma ukvarjala nekoliko kasneje.[61] Čeprav kritično obravnavana, je bila Billigova ideja razširjena in dopolnjena z različnih raziskovalnih perspektiv in v različnih geografskih okvirih. Pri tem se je od njega s svojim delom *National identity, popular culture and everday life* (2002) še najmanj oddaljil Tim Edensor. Knjiga ponuja temeljit vpogled v reprodukcijo nacionalne identitete na različnih družbenih ravneh, ki jih avtor predstavi kot matrico, ki posamezniku omogoča širok razpon razpoložljivih resursov za nacionalno identitetno orientacijo. Pri svoji analizi se osredotoči na prostorsko dimenzijo nacionalizma, performativne prakse na skupinski (proslave, popularne oblike državnih praznovanj, šport, karnevali) ali individualni ravni (nerefleksivne vsakodnevne prakse, oblikovane skozi nacionalni narativ), prispevek materialne kulture h konstituiranju nacionalne identitete, njeno reprezentacijo v popularni kulturi in razpršeno dojemanje simbolov (britanske) nacionalne identitete med prebivalstvom.[62]

56 Bourdieu, *Outline of a Theory of Practice*, 166, cit. po: Fox in Miller Idriss, »Everday nationhood,« 544.

57 Billig, *Banal Nationalism*, 42.

58 Jones in Bradbury, *Introducing Social Theory*, pogl. 7.

59 Več o tem: Thompson, »Nations, national identities,« 28. Hearn, »National identity,« 660–62. Skey, »The national in everyday life,« 335–38. Marco Antonsich, »The 'everyday' of banal nationalism – ordinary people's views on Italy and Italian,« *Political Geography* 54, (2016): 3, 34, pridobljeno 7. 6. 2025, https://doi.org/10.1016/j.polgeo.2015.07.006.

60 Antonsich, »The 'everyday' of banal nationalism,« 33.

61 Tim Edensor, *National Identity, Popular Culture and Everyday Life* (Oxford: Berg, 2002). Rogers Brubaker et al., *Nationalist Politics and Everyday Ethnicity in a Transylvanian Town* (Princeton: Princeton University Press, 2006). Michael Skey, *National Belonging and Everyday Life: The Significance of Nationhood in an Uncertain World* (Basingstoke: Palgrave Macmillan, 2011).

62 Edensor, *National Identity*, Predgovor (VII–VIII).

Rutinsko izražanje zavednih in nezavednih oblik nacionalne performativnosti, pri analizi katerih se teoretsko nasloni na Bourdiejevo teorijo habitusa in Goffmanovo teorijo družbenih vlog[63], opiše kot »skupne navade, ki krepijo čustvene in kognitivne vezi, utrjujejo občutek skupnega delovanja in doxe, ki tvorijo habitus, vključno s pridobljenimi veščinami, ki zmanjšujejo nepotrebno razmišljanje vsakič, ko je treba sprejeti odločitev«.[64] Čeprav zavestno reproduciranje družbenih struktur zaradi ontološke varnosti in temeljne želje po določeni stopnji družbene predvidljivosti poudari tudi Giddens,[65] pa je prav zaradi tega lažje razumljiva tudi kritika Edensorjevega dela, ki do neke mere ohranja analizo prevladujočega diskurza in strukturnih sil pri konstruiranju nacionalne identitete v vsakdanjem življenju, pri čemer pa nekoliko zanemari dimenzijo etnične, rasne in verske raznolikosti družbe.[66] Kljub temu da Edensor akterjeve vloge nikakor ne predstavi samo kot enosmernega sprejemanja in reprodukcije strukturno pogojenega nacionalnega diskurza, pa v smislu prenosa vpliva od struktur na akterja ni tako uspešen, zaradi česar se njegovo delo umešča na presečišče banalnega in vsakdanjega nacionalizma.

Nekoliko bolj se bistvu vsakdanjega nacionalizma približa Michael Skey, ki v knjigi *National Belonging and Everyday Life: The Significance of Nationhood in Uncertain World* (2011) obravnava pomembnost nacionalne identitete in pripadnosti na primeru belske etnične večine v Veliki Britaniji.[67] Delo, ki temelji na avtorjevi doktorski disertaciji, pod vprašaj postavi nacionalizem kot preprosto vprašanje potrošnje, saj ga analizira na podlagi skupinskih pogovorov o nacionalni identiteti, njeni razlagi in razumevanju med splošnim prebivalstvom. S tem se tematsko odmakne od preproste Billigove banalnosti nacionalizma in njegove nezavedne reprodukcije. Avtor namreč išče odgovore na vprašanje, kaj pomeni biti Britanec, s čimer poskuša prodreti v različne vidike nacionalnega (samo)zavedanja, ki presegajo njegovo preprosto samoumevnost.[68] Razlika v primerjavi z deli Billiga in Edensorja je predvsem osredotočanje na tiste vidike nacionalizma, ki niso neposredno povezani z njegovimi preprostimi razlagami in splošno sprejetimi simboli, temveč zahtevajo nekoliko natančnejši premislek in pojasnilo. S takim pristopom je Skey osvetlil vidik izkušnje nacionalnosti skozi oči akterjev, ki nacionalne kategorije ohranjajo z vsakodnevnim delovanjem. V delu se tako meša vidik (nezavednega) banalnega nacionalizma, ki ga poskuša avtor preseči z odkrivanjem zavednega razumevanja nacionalne identitete. Medsebojna prepletenost se izkazuje predvsem v njunem (so)oblikovanju, kar močno spominja na Giddensovo tezo o dualnosti strukture in delovanja, pri čemer se oba pola definirata drug skozi drugega.[69] Samoumevnost nacionalnosti in njenega izražanja lahko torej pripišemo predvsem področju banalnega nacionalizma, ki obravnava njene najočitnejše strukturno

63 Erving Goffman, *Predstavljanje sebe v vsakdanjem življenju* (Ljubljana: Studia Humanitatis, 2014).
64 Edensor, *National Identity*, 90.
65 Peter Stankovič, »Giddensova teorija strukturacije,« 465, 466. Haralambos in Holborn, *Sociologija*, 912.
66 Jonathan Hearn in Marco Antonsich, »Theoretical and methodological,« 11, 12.
67 Skey, *National Belonging and Everyday Life*, 6–8.
68 Skey, *National Belonging and Everyday Life*.
69 Stankovič, »Giddensova teorija strukturacije,« 458.

pogojene vidike, močno prisotne tudi na vsakdanji ravni. Vendar pa po drugi strani zanemarja vse tiste vidike, ki ostajajo nevid(e)ni, nesliš(a)ni in nezaznavni v vsakodnevnih opravilih oziroma delovanju navadnih ljudi. Čeprav so tudi takšne prakse pod vplivom strukturnih predpogojev, jih s svojim tihim delovanjem istočasno ohranjajo ali spreminjajo, zaradi česar sta banalni in vsakdanji nacionalizem v samem bistvu dve strani istega kovanca.[70]

Na podoben način so k preučevanju nacionalizma na vsakdanji ravni pristopili tudi avtorji knjige *Nationalist politics and everday ethnicity in a Transylvanian town* (2006), ki so svojo analizo usmerili v razlago narodnosti in etničnosti med prebivalci narodnostno mešanega romunskega mesta Cluj. Delo je osredotočeno na analitično razumevanje kompleksnega zgodovinskega sobivanja madžarsko-romunskega prebivalstva »tako od spodaj navzgor kot od zgoraj navzdol, tako v mikroanalitični kot tudi makroanalitični perspektivi«.[71] Pri tem je zelo pomembna preusmeritev raziskovalnega fokusa etničnosti, ki jo poskušajo raziskati neodvisno od etničnih kategorij kot osnovnih enot analize. Z drugimi besedami to pomeni, da so s svojo obsežno raziskavo, ki je zajemala osemletno etnografsko raziskovanje z izvedbo intervjujev, analizirali vsakodnevne vidike izražanja etničnosti oziroma nacionalnosti skozi pridobljene (banalne) prakse in zavestne strategije posameznikov, ki so etnične kategorije prilagajali praktičnim potrebam in institucionalnim kontekstom.[72] Delo je zanimivo predvsem zaradi poskusa razumevanja vsakodnevnih praks, ki niso prepoznane kot etnično zaznamovane, a zaradi vtkanosti v vsakodnevne kontekste samoumevno ohranjajo razlike v etničnih kategorijah. Pri tem poudarijo spontano pojavnost etničnosti v vsakodnevnih interakcijah med prijatelji, sosedi, sodelavci ter v različnih institucionalnih in organizacijskih okvirih, s čimer se zelo približajo bistveni ideji vsakdanjega nacionalizma, ki temelji na pomembni vlogi akterjev pri dojemanju nacionalnih kategorij in njihovi uporabi v vsakodnevnih kontekstih za njihove praktične namene.[73] Ker nacionalne kategorije interakcij in samozavedanja posameznikov ne prežemajo konstantno, je vprašanje o tem, kdaj se take tendence pojavljajo, ključno pri razumevanju pojavnosti nacionalizma.[74] To odpira številne teoretične premisleke o kontekstualnem delovanju akterja, ki je lahko povezano z njegovo analizo zunanjih okoliščin in situacijskim praktičnim znanjem, pa tudi s pomembnim vidikom njegovega nastopanja in ustvarjanja vtisa v javnosti.[75]

S teoretičnega vidika je za obravnavo obeh pristopov izredno pomembno tudi delo sociologa Siniše Malaševića, ki v knjigi z naslovom *Grounded nationalisms* opiše kompleksnost prisotnosti nacionalizma na različnih ravneh družbe.[76] Pri tem poudari tri ključne vidike, ki obsegajo njegovo ideološko prodornost, vpletenost v družbene

70 Jon E. Fox in Maarten Van Ginderachter, »Everyday nationalism's evidence problem,« 7.
71 Brubaker et al., *Nationalist Politics and Everyday Ethnicity,* 14.
72 Ibid, 272, 273 in 297.
73 Fox, »Banal nationalism in everyday life,« 864.
74 Fox in Miller-Idriss, »Everyday nationhood,« 540.
75 Goffman, *Predstavljanje sebe v vsakdanjem življenju*.
76 Siniša Malašević, *Grounded Nationalisms* (Cambridge: Cambridge University Press, 2019).

organizacijske okvire in prisotnost v vsakdanjih medčloveških odnosih, ki krojijo vsakdanje življenje posameznikov. Takšna prisotnost po celotni družbeni vertikali, obsegajoča makronarative (banalni nacionalizem) in mikrokontekstualno vsakdanjo reprodukcijo (vsakdanji nacionalizem), prispeva k njegovi ukoreninjenosti v postmoderne družbe. Prav zato Malašević opozarja, da nacionalizma ne bi smeli dojemati kot ideološkega relikta preteklosti, temveč kot globoko ukoreninjen in razvijajoč se družbeni fenomen, ki tudi v sodobnem globaliziranem času ostaja pomembna družbena sila.[77]

Na presečišču do sedaj obravnavanih del, ki k preučevanju nacionalizma na vsakdanji ravni pristopajo z različnih zornih kotov, so se oblikovale tudi številne kasnejše raziskave. Glede na to, kakšno pomembnost so pri tem pripisovale elitnemu nacionalnemu diskurzu oziroma akterjevemu prispevku pri njegovem oblikovanju, ohranjanju in spreminjanju, jih lahko prav z vidika razmerja med strukturo in delovanjem razporedimo na premico med vsakdanjim in banalnim nacionalizmom. Kljub temu da se nekatera dela jasno ukvarjajo s političnim diskurzom in njegovo konstrukcijo nacionalne identitete[78] ali uporabo (naravnih) resursov za nacionalne mobilizacijske procese,[79] pa se večina del umešča v prostor med obema pristopoma. O dinamičnem razumevanju razlike med obema pričajo tako dela, ki z Billigovo teorijo jasno polemizirajo,[80] kot tudi odgovori na takšne kritike in njihova pojasnila.[81] K omenjenim študijam lahko prištejemo še članke, ki se z obema pristopoma ukvarjajo s teoretičnega in metodološkega vidika.[82] Ob tem so izredno pomembna tudi dela, ki obravnavajo narodnost v vsakdanjih kontekstih potrošniških navad[83] ali performativnih praks,[84] kar odstira pomembne plasti razumevanja situacij, v katerih se nacionalizem izraža in utrjuje v vsakdanjem življenju.

Nikakor seveda ne čudi, da so omenjene raziskovalne perspektive močno vplivale tudi na historiografske pristope k preučevanju nacionalizma. Pomen nacionalnih

77 Ibid.

78 Rudolf De Cillia, Martin Reisigl in Ruth Wodak, »The discoursive construction of national identities,« *Discourse & Society* 10, št. 2 (1999), pridobljeno 16. 6. 2025, https://doi.org/10.1177/0957926599010002002.

79 Nalie Koch in Tom Perrault, »Resource nationalism,« *Progress in Human Geography* 43, št. 4 (2018), pridobljeno 27. 6. 2025, https://doi.org/10.1177/0309132518781497.

80 Skey, »The national in everyday life.« Fox in Miller-Idriss, »Everyday nationhood.« Fox, »Banal nationalism in everyday life.«

81 Michael Billig, »Reflecting on a critical engagement with banal nationalism – reply to Skey,« *The Sociological Review* 57, št. 2 (2009), pridobljeno 7. 6. 2025, https://doi.org/10.1111/j.1467-954X.2009.01837.x. Sophie Duchesne, »Who's afraid of banal nationalism?«

82 Thompson, »Nations, national identities.« Knott, »Everyday nationalism.« Goode in Stroup, »Everyday nationalism.« Eric Kaufmann, »Complexity and nationalism,« *Nations and Nationalism* 23, št. 1 (2017), pridobljeno 26. 6. 2025, https://doi.org/10.1111/nana.12270. Hearn in Antonsich, »Theoretical and methodological.« Fox in Ginderachter, »Everday nationalism's evidence problem.«

83 Robert J. Foster, *Materializing the Nation: Commodities, Consumption, and Media in Papua New Guinea* (Bloomington: Indiana University Press, 2002). Ronald Ranta in Atsuko Ichijo, *Food, National Identity and Nationalism: From Everyday to Global Politics* (Basingstoke: Palgrave Macmillan, 2016). Helen Andersson, »Recontextualizing Swedish nationalism for commercial purposes: a multimodal analysis of a milk marketing event,« *Critical Discourse Studies* 16, št. 5 (2019), pridobljeno 6. 6. 2025, https://doi.org/10.1080/17405904.2019.1637761.

84 Thomas Eriksen, »Formal and informal nationalism,« *Ethnic and Racial Studies* 16, št. 1 (1993), pridobljeno 3. 7. 2025, https://doi.org/10.1080/01419870.1993.9993770. Kristin Surak, *Making Tea, Making Japan: Cultural Nationalism in Practice* (Redwood City: Stanford University Press, 2012).

kategorij pri osebnih identitetnih opredelitvah, ki so v daljši časovni perspektivi pripomogle k praktičnemu oblikovanju nacionalnih držav, je področje, ki so ga zgodovinarji večkrat natančneje obravnavali. Pri tem procesu je posebno raziskovalno pozornost dobila predvsem dilema med vplivom elitno oblikovanih nacionalnih diskurzov (in njihovega uveljavljanja v družbi) ter nacionalno brezbrižnostjo (*national indifference*) navadnih ljudi v vsakdanjem življenju.[85] Kot namreč v svojem delu *Kidnapped Souls – National Indifference and the Battle for Children in the Bohemian Lands* na primeru Češke poudarja zgodovinarka Tara Zahra, naj bi brezbrižnost, protislovja in oportunizem do nacionalnih kategorij pod vprašaj postavili v zgodovini uveljavljeno trditev, da je samo homogena nacionalna država lahko zagotavljala trajno demokracijo, mir in blaginjo.[86] Čeprav so pri historičnih obravnavah omenjenega fenomena zaradi svoje specifične nadnacionalne zgodovine največkrat v ospredju države (vzhodne) srednje Evrope,[87] pa jih s primeri iz drugih delov stare celine natančneje obravnava zbornik *National Indifference and the History of Nationalism in Modern Europe*.[88] Zgodovinske raziskave osebnih privrženosti ali brezbrižnosti do nacionalnih idej in kategorij so za razumevanje razmerja med vplivom banalnega in vsakdanjega nacionalizma v sodobnem času zelo pomembne predvsem zaradi metodološkega pristopa, ki se v tem primeru ne osredotoča samo na analize nacionalnih (elitnih) političnih diskurzov, temveč predvsem na podrobnosti vsakdanjih življenj navadnih ljudi.[89]

Med deli, ki takšen metodološki pristop pojasnijo z najrazličnejših zornih kotov, velja omeniti predvsem članek z naslovom *Everyday nationhood* (2008), v katerem avtorja obravnavata teoretične in metodološke vidike preučevanja vsakdanjega nacionalizma, ki poskuša zajeti tiste vidike narodnosti, ki jih strukturalni pristopi ne morejo.[90] Ker »narodi niso samo produkt strukturnih sil, temveč so hkrati praktični dosežek navadnih ljudi, ki v vsakdanjem življenju počnejo vsakdanje dejavnosti«,[91] se avtorja osredotočata predvsem na vidike (vsakdanje interakcije, nacionalne izbire, pomen nacionalnih simbolov in potrošniške prakse), s katerimi ljudje »v različnih kontekstih svojega vsakdanjega življenja izražajo in uresničujejo (ter ignorirajo in zavračajo) narodnost in nacionalizem«.[92] Pomen akterjevega doprinosa pri vsakdanjem soustvarjanju

85 Tara Zahra, *Kidnapped Souls: National Indifference and the Battle for Children in the Bohemian Lands, 1900–1948* (Ithaca: Cornell University Pres, 2011). Maarten Van Ginderachter, »How to gauge banal nationalism and national indifference in the past: proletarian tweets in Belgium's belle epoque,« *Nations and Nationalism* 24, št. 3 (2018), pridobljeno 16. 6. 2025, doi: 10.1111/nana.12420.

86 Zahra, *Kiddnaped Souls*, X.

87 Pri analizah privrženosti in zavračanja nacionalnih idej na vsakdanji ravni v specifičnih lokalnih ali regionalnih kontekstih tudi slovenski prostor ni izjema: Pieter Judson, *Guardians of the Nation: Activists on the Language Frontiers of Imperial Austria* (Cambridge: Harvard University Press, 2006). Marko Zajc, »Josip Jurčič's tradition in Muljava: The boundaries of localism and nationalism,« *Prispevki za novejšo zgodovino* 53, št. 2 (2013), pridobljeno 3. 10. 2025, https://ojs.inz.si/pnz/article/view/65/66. Jernej Kosi, »'Yugoslavia has nothing. Yugoslavia has no bread. But Hungary gives us bread': Access to food and (dis)loyalty in a 'redeemed' Yugoslav borderland,« *Austrian History Yearbook* 55 (2024), pridobljeno 3. 10. 2025, https://doi.org/10.1017/S0067237824000055.

88 Maarten Van Ginderachter in Jon E. Fox, ur., *National Indifference and the History of Nationalism in Modern Europe* (London, New York: Routledge University Press, 2019).

89 Van Ginderachter in Fox, *National Indifference*, 1.

90 Fox in Miller-Idriss, »Everyday nationhood.«

91 Ibid., 554.

92 Ibid., 537.

nacionalne družbene realnosti so raziskovalci do sedaj obravnavali na različne načine, pri čemer v svojih študijah poudarjajo perspektivo posameznikov znotraj in zunaj prevladujočih nacional(istič)nih diskurzov.[93] V zvezi s tem sta še posebno zanimiva članka, ki obravnavata svojevrstno dojemanje in dekodiranje nacionalnih kategorij z migrantske perspektive[94] ali kompleksno osebno dojemanje kulturnih prvin multikulturalizma kot pomembnega dela družbene identitete.[95]

Pri preučevanju (vsakdanjega) vpliva nacionalnih kategorij na osebne in družbene dimenzije nacionalne identitete v razmerju med makro- (struktura) in mikroravnijo (delovanje) obeh spektrov družbene realnosti nikakor ne moremo popolnoma ločiti. Pomembno raziskovalno perspektivo ponudi članek *National identity: banal, personal and embedded* (2007), ki obravnava specifičnost izražanja in ohranjanja nacionalnih kategorij v organizacijskih kontekstih.[96] Članek obravnava koncept nacionalne identitete v razmerju med osebno in družbeno ravnijo, posamezni vidiki pa so razloženi na primeru raziskave socialno-organizacijskega okolja in njegovega vpliva na zaposlene ob združitvi škotske in angleške banke. Z analizo različnih vidikov (banalne in osebne) nacionalne identitete poskuša avtor razumeti povezavo med osebnostjo in vplivi vmesnih družbenih struktur ter organizacijskih kontekstov na posameznikovo razumevanje samega sebe.[97] Ker organizacijski konteksti predstavljajo vmesno stopnjo med makrostrukturnimi vplivi in delovanjem na mikroravni, se s tem odpre vprašanje vpliva vmesnih struktur, ki so jih kot pomembne prepoznali tudi nekateri raziskovalci teorije strukturacije.[98] Razširitev osnovne Giddensove ideje o prepletenosti strukture in delovanja z uvajanjem novih podstruktur tako ostaja dobra teoretična osnova za preučevanje banalnega, še bolj pa vsakdanjega nacionalizma, saj se kljub konceptualni osredotočenosti na akterja tega ne da popolnoma ločiti od družbeno-organizacijskih kontekstov.[99]

93 Marco Antonsich, »The 'everyday' of banal nationalism.« Cynthia Miller-Idriss, »Everyday understanding of citizenship in Germany,« *Citizenship Studies* 10, št. 5 (2006), pridobljeno 3. 6. 2025, https://doi.org/10.1080/13621020600954978. Jon E. Fox, »The edges of the nation: a research agenda for uncovering the taken-for-granted foundations of everyday nationhood,« *Nations and Nationalism* 23, št. 1 (2017), pridobljeno 8. 6. 2025, https://doi.org/10.1111/nana.12269. Michael Skey, »'There are times when I feel like a bit of an alien': middling migrants and the national order of things,« *Nations and Nationalism* 24, št. 3 (2018), pridobljeno 12. 6. 2025, doi: 10.1111/nana.12422. Edensor in Sumartojo, »Geographies of everyday nationhood.«

94 Skey, »There are times.«

95 Tim Edensor in Shanti Sumartojo, »Geographies of everyday nationhood.«

96 Jonathan Hearn, »National identity: banal, personal and embedded,« *Nations and Nationalism* 13, št. 4 (2007), pridobljeno 15. 6. 2025, https://doi.org/10.1111/j.1469-8129.2007.00303.x.

97 Hearn, »National identity,« 671.

98 Rob Stones, *Structuration Theory* (Basingstoke: Palgrave Macmillan, 2005). McGarry, »Knowing how to go on.« O'Reilly, »Structuration, practice theory, ethnography and migration.«

99 Hearn in Antonsich, »Theoretical and methodological,« 2 in 6, 7.

Soodvisnost strukture in delovanja ter pomen vmesnih posrednikov

Pregled temeljnih del o banalnih in vsakdanjih vidikih nacionalizma dokazuje, da koncepti naroda, nacionalizma in nacionalne identitete zagotovo niso fenomeni, ki bi se v zavest ljudi usidrali samo prek strukturne prisile, denimo elitnih nacionalnih diskurzov, ki jih posredujejo politični govori, medijskih ali historičnih reprezentacij, javno izobešenih zastav ali nacionalnih praznikov. Gre namreč za relativno kompleksen proces identificiranja posameznika z neko ide(ologi)jo, ki močno presega enosmernost. Kot je namreč poudaril Hobsbawm, niso niti »uradne ideologije držav in gibanj vodilo za to, kar mislijo celo najbolj lojalni državljani«.[100] Na tem primeru se s sociološkega vidika vnovič poraja vprašanje o razmerju med močjo strukturnih sil, delovanjem akterja in prepletenostjo obeh vidikov pri oblikovanju nekega družbenega fenomena.

Konceptualna razlika z uveljavljenimi raziskovalnimi paradigmami je na podlagi tega razmerja vplivala na številna znanstvena nestrinjanja. Poskusi razumevanja kompleksnega nacionalnega (samo)zavedanja izven prevladujočega nacionalnega diskurza so pri nekaterih raziskovalcih naleteli na ostro kritiko. Kot poudarja Anthony Smith, so »navadni ljudje in njihove aktivnosti vedno umeščeni v zgodovinski kontekst«,[101] preučevanje nacionalnosti v vsakdanjem življenju pa naj bi se selektivno osredotočalo na situacijsko specifičnost oziroma mikroanalitične vidike in s tem ločevalo »vsakdanjo narodnost« od »historične narodnosti«.[102] Smith ob tem doda, da lahko pretirano osredotočanje na to, kako, kdaj in na kakšen način se narodnost izkazuje v vsakdanjem življenju, pripelje do zanemarjanja vzročnih in družbeno-historičnih aspektov nacionalizma.[103] Z drugimi besedami: pretirano preučevanje nacionalizma in poskusi njegovega razumevanja skozi prizmo akterja naj ne bi bili ustrezni zaradi strukturnih vplivov, ki jih raziskovalci ne smejo zanemariti.

Vendar pa ostra delitev med uveljavljenimi pristopi preučevanja nacionalizma in analizo njegovih banalnih in vsakdanjih oblik še zdaleč ni preprosta. Čeprav se oba pristopa osredotočata na specifične vsakodnevne kontekste, ju le stežka jasno kategoriziramo in popolnoma zanemarimo prepletenost obeh sociooloških perspektiv. Zaradi preučevanja elitnega (makro)diskurza, njegovega vertikalnega pronicanja navzdol in reprodukcije na stopnji posameznika nekateri avtorji banalni nacionalizem sicer uvrščajo bližje strukturni ravni,[104] saj samoumevnost nacionalnih kategorij deluje predvsem v sferi posameznikovega nezavednega.[105] Pri njegovi obravnavi se avtorji pogosto približajo teoriji (in terminologiji), ki jo je pri zbliževanju strukture in

100 Hobsbawm, *Nations and Nationalism*, 10, cit. po: Thompson, »Complexity and nationalism,« 8.

101 Anthony Smith, »The limits of everyday nationhood,« Ethnicities 8, št. 4 (2008): 566, pridobljeno 13. 6. 2025, https://doi.org/10.1177/1468796808080040102.

102 Ibid.

103 Ibid.

104 Hearn in Antonsich, »Theoretical and methodological,« 3.

105 Billig, *Banal Nationalism*, 38 in 144.

delovanja uporabil že Bourdieu, saj nezavedno reprodukcijo nacionalne ideje označujejo z besedami »nacionalni habitus«, »nacionalna doxa« ali »nacionalni kapital«.[106] To nakazuje na večinoma pridobljene prakse, ki jih v svojih življenjih nezavedno ponotranjimo in reproduciramo, pri čemer imajo jasen vpliv strukturne razmere, ki smo jim izpostavljeni. Toda pri analizi nacionalizma na osebni ravni posameznika je takšno (raz)ločevanje nekoliko kompleksnejše, saj se ta lahko izraža kot osrednji predmet in namen delovanja, lahko delovanje samo uokvirja ali pa ni bistvo dejanja, a deluje kot nevidna predpostavka v njegovem ozadju.[107]

Že iz Smithove kritike lahko razberemo, da vpliv struktur ni zanemarljiv niti na polju vsakdanjega nacionalizma, ki v ospredje še bolj postavi akterja in njegovo delovanje.[108] Uveljavljene socio-historične paradigme sicer niso popolnoma zanemarjale vloge akterja, a so predpostavljale samoumevnost vrednot elitnega nacionalnega diskurza, prek katerega je determinirano tudi akterjevo delovanje.[109] Prav poudarek na akterjevi sposobnosti osmišljanja in preoblikovanja nacionalnih kategorij pa je za razumevanje pojavnosti nacionalizma ključno, saj jih posamezniki – kljub nezavedni vključenosti vanje – uporabljajo za razlago, umeščenost in istovetenje samih sebe v odnosu do družbe.[110] Narodi v tem primeru niso samo produkt diskurza elit in njihovih upravnih mehanizmov, temveč nastajajoča lastnost (*emergent property*) kompleksnega obnašanja in interakcije na nižji ravni oziroma delovanja posameznih delov sistema, ki sledijo nekaterim osnovnim smernicam.[111] Takšen spontani red, ki nastane iz omenjene kompleksnosti, ima zaradi kroženja idej in praks večjo moč kot pa centralno koordiniran red, kjer je idejna moč skoncentrirana med elitnimi načrtovalci. Prav to je tudi eden od bistvenih razlogov za samoohranjanje idejnih kategorij nacionalizma med sicer heterogenim prebivalstvom. Različnost (nacionalnih) identitetnih perspektiv posameznih članov družbe se z elitnim nacionalnim diskurzom povezuje prek posameznih individualnih percepcij nacionalnih kategorij, ki so z makronarativom zvezane v delujočo celoto in se v praksi izkazujejo kot koherentna podstat neke družbe.[112] Pri tem je treba poudariti, da posamezni »osebni nacionalizmi« še zdaleč niso samo manifestacije družbene strukture in elitnega diskurza, temveč predvsem osebno osmišljanje nacionalnih kategorij, kar utrjuje pomembnost nacionalizma v vsakdanjem delovanju posameznika v družbi.[113] Z drugimi besedami to pomeni, da je bistvo preučevanja vsakdanjih vidikov nacionalizma predvsem v analizi različnosti posameznih osebnih interpretacij nacionalnih kategorij, saj »niti dva

106 Fox in Van Ginderachter, »Everyday nationalism's evidence problem,« 3.

107 Hearn in Antonsich, »Theoretical and methodological,« 5.

108 Smith, »The limits of everyday nationhood.«

109 Thompson, »Nations, national identities,« 19–22.

110 Andrew Thompson, »Nations, national identities,« 20. Jonathan Hearn, »National identity,« 663, 664. Kaufmann, »Complexity and nationalism,« 19, 20.

111 Kaufmann, »Complexity and nationalism,« 13 in 19.

112 Hearn, »National identity,« 663, 664. Kaufmann, »Complexity and nationalism,« 19–21.

113 Anthony Cohen, »Personal nationalism: a Scottish view of some rites, rights and wrongs,« *American Ethnologist* 23, št. 4 (1996): 802, pridobljeno 7. 7. 2025, https://doi.org/10.1525/ae.1996.23.4.02a00070, cit. po: Hearn, »National identity,« 663. Robert Mann in Steve Fenton, »The personal contexts of national sentiments,« *Journal of Ethnic and Migration Studies* 35, št. 4 (2009): 520, pridobljeno 10. 7. 2025, https://doi.org/10.1080/13691830902764882.

posameznika nista umeščena v popolnoma enak etnični, geografski, spolni ali psihološki prostor. Na narod gledata z edinstvenih zornih kotov, kar ima za posledico različne nacionalne identitete.«[114]

Kljub poudarku na razumevanju posameznikovega nacionalnega izkustva pa preučevanje vsakdanjega nacionalizma nikakor ne bi smelo zanemariti družbenega in organizacijskega konteksta, v katerem nastaja. To je bistveno pri razlagi, kako lahko nacionalne kategorije oblikujejo družbeno življenje.[115] Preučevanje vsakdanjega nacionalizma zato ne more biti ločeno od vmesnih oziroma posrednih organizacijskih struktur med makro- in mikroravnijo (družina, delovna mesta, izobraževalne ustanove ipd.), saj so to največkrat ključni spodbujevalci človekovega delovanja.[116] Pri analizi nacionalizma torej ni pomemben samo elitni diskurz ali njegova kontekstualna uporaba v dnevni interakciji posameznikov, temveč tudi posredni mediatorji, kot so organizacijski konteksti, v katerih posameznik vsakodnevno (so)deluje. Takšne družbene povezave lahko bolj ali manj uspešno opolnomočijo posameznike in njihovo istovetenje s širšimi identitetnimi kategorijami, zaradi česar je močnejša tudi njihova privrženost nacionalnim idejam ali ciljem.[117]

Pri razumevanju te kompleksnosti nam lahko pomaga prav Giddensova teorija strukturacije, ki so jo nekateri raziskovalci razvijali v smeri preučevanja medsebojnega delovanja mikro-, makro- in vmesnih strukturnih slojev.[118] Z njimi lahko lažje pojasnimo akterjevo umeščenost v družbeni kontekst, ki neposredno oblikuje posameznikovo življenje ter mobilizira njegove interese in identiteto.[119] Strukturacijski celostni pristop k preučevanju družbe omogoča preseganje razumevanja nacionalizma v obliki dveh ločenih polov (osebnega ali strukturnega) in osredotočanje na njegovo reprodukcijo v vrsti odnosov, ko posamezniki osebno relevantnost širših kognitivnih (nacionalnih) kategorij črpajo iz različnih oblik družbene organiziranosti.[120]

V primeru banalnega in vsakdanjega nacionalizma, ki nacionalnost preučujeta z vsakodnevne perspektive posameznika, omogoča razumevanje strukturnega vpliva na banalne vidike vsakdana, ki so pogosto povezani z akterjevo praktično zavestjo in nerefleksivnim obnašanjem, ter razumevanje refleksivne uporabe nacionalnih kategorij, ki so posledica posameznikove diskurzivne racionalne zavesti.[121] Pri oblikovanju širših teoretskih razlag nacionalizma je takšen pristop sinteze strukture in delovanja uporaben zaradi možnosti raziskovanja relevantnih življenjskih praks v povezavi s širšimi strukturnimi konfiguracijami naroda in nacionalizma ter ohranjanja njegove vsakodnevne pojavnosti in pomembnosti v daljšem časovnem obdobju.[122]

114 Kaufmann, »Complexity and nationalism,« 20.
115 Hearn in Antonsich, »Theoretical and methodological,« 2 in 6.
116 Ibid., 6. Hearn, »National identity,« 667, 670, 671.
117 Hearn, »National identity,«, 667, 670–72. Kaufmann, »Complexity and nationalism,« 18.
118 Rob Stones, *Structuration Theory* (Basingstoke: Palgrave Macmillan, 2005). Orla McGarry, »Knowing 'how to go on',« 2070–75. Karen O'Reilly, »Structuration, practice theory, ethnography and migration,« 7–9.
119 Hearn in Antonsich, »Theoretical and methodological,« 6.
120 Hearn, »National identity,« 671.
121 Jones in Bradbury, *Introducing Social Theory,* pogl. 7.
122 Thompson, »Nations, national identities,« 24. Goode in Stroup, »Everyday nationalism: constructivism for the masses,« 724, 725.

Ker je nacionalizem v družbi izrazito kompleksen fenomen, ki pogosto ni povezan samo z elitnim nacionalnim narativom, temveč nastaja kot stranski produkt različnih dejanj posameznikov in z njimi povezanih procesov[123], teorija strukturacije nudi pomembna orodja in bogato konceptualno besedišče, s katerimi lahko zajamemo njegovo celotno kompleksnost.[124]

Sklep

V članku obravnavam koncepta vsakdanjega in banalnega nacionalizma, ki predstavljata teoretski odmik od klasičnih makropristopov preučevanja nacionalizma, saj se ukvarjata s pronicanjem, reprodukcijo, razumevanjem in uporabo nacionalnih kategorij na ravni akterja. Medtem ko se banalni osredotoča na načine prodiranja ideologije nacionalizma z makroravni na mikroraven, vsakdanji nacionalizem obravnava akterjevo zavedno in nezavedno uporabo nacionalnih kategorij v praktične vsakodnevne namene. S člankom želim poudariti, da se tudi v primeru preučevanja vsakdanjega in banalnega nacionalizma pojavlja klasična sociološka dilema pri iskanju pravega pristopa, ki bi zajel celotno kompleksnost dinamike v razmerju med strukturnimi vplivi (makroraven) in vsakodnevnim delovanjem posameznikov (mikroraven). S pregledom temeljnih del prikažem kompleksno interakcijo in povezanost obeh spektrov družbene realnosti in poudarke, ki so jih pri svojem preučevanju uporabili različni raziskovalci. Z obravnavanimi primeri in njihovo sociološko kontekstualizacijo dokazujem pomembnost sociooloških pristopov, ki poskušajo zajeti povezanost med strukturnimi vplivi in delovanjem posameznih članov družbe. Pri tem še posebej poudarim teorijo strukturacije, ki omogoča določitev vmesnih strukturnih vplivov, in konceptualno soodvisnost akterja in strukture, s čimer je pojasnjevanje kompleksnosti horizontalnih in vertikalnih vplivov nacionalizma nekoliko lažje ter razumljivejše. Kljub temu da pristop še zdaleč ni popoln[125] in končnega zadovoljujočega odgovora tudi v povezavi z nacionalizmom ne more ponuditi, pa je to tudi pozitivna plat, saj od raziskovalcev zahteva, da si pri analitičnem preučevanju takih fenomenov še naprej prizadevajo za celosten pogled z vidika obeh perspektiv. Pri tem lahko teorija strukturacije omogoči konceptualni okvir, sintezo posameznih vidikov preučevanja ter identifikacijo vzajemnega vpliva strukture in delovanja na ohranjanje nacionalnih kategorij v kompleksni postmoderni družbi.

123 Eric Kaufmann, »Complexity and nationalism,«, 9, 10.

124 McGarry, »Knowing 'how to go on',« 2070.

125 Margaret Archer, »Morphogenesis versus structuration«. Peter Stankovič, »Giddensova teorija strukturacije,« 463472.

Zahvala

Članek je nastal v okviru raziskovalnega programa P5-0070 Narodna in kulturna identiteta slovenskega izseljenstva v kontekstu raziskovanja migracij, ki ga financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS) iz državnega proračuna.

Viri in literatura

Andersson, Helen. »Recontextualizing Swedish nationalism for commercial purposes: a multimodal analysis of a milk marketing event.« *Critical Discourse Studies* 16, št. 5 (2019): 583–603. Pridobljeno 6. 6. 2025. https://doi.org/10.1080/17405904.2019.1637761.

Antonsich, Marco. »The 'everyday' of banal nationalism – Ordinary people's views on Italy and Italian.« *Political Geography* 54 (2016): 32–42. Pridobljeno 3. 6. 2025. https://doi.org/10.1016/j.polgeo.2015.07.006.

Archer, Margaret. »Morphogenesis versus structuration: on combining structure and action.« *British Journal of Sociology* 33, št. 4 (1982): 455–83. Pridobljeno 18. 6. 2025. https://doi.org/10.2307/589357.

Archer, Margaret. »Realism and the problem of agency.« *Journal of Critical Realism (Alethia)* 5, št. 1 (2015): 11–20. Pridobljeno 17. 6. 2025. https://doi.org/10.1558/aleth.v5i1.11.

Archer, Margaret. *Realist Social Theory: The Morphogenetic Approach*. Cambridge: Cambridge University Press, 1995.

Billig, Michael. »Reflecting on a critical engagement with banal nationalism – reply to Skey.« *The Sociological Review* 57, št. 2 (2009): 347–52. Pridobljeno 7. 6. 2025. https://doi.org/10.1111/j.1467-954X.2009.01837.x.

Billig, Michael. *Banal Nationalism*. London: SAGE Publications, 1995.

Bourdieu, Pierre. *Outline of a Theory of Practice*. Cambridge: Cambridge University Press, 1977.

Brubaker, Rogers, Liana Grancea, Jon Fox in Margit Feishmidt. *Nationalist Politics and Everyday Ethnicity in a Transylvanian Town.* Princeton: Princeton University Press, 2006.

Cohen, Anthony. »Personal nationalism: a Scottish view of some rites, rights and wrongs.« *American Ethnologist* 23, št. 4 (1996): 802. Pridobljeno 7. 7. 2025. https://doi.org/10.1525/ae.1996.23.4.02a00070.

De Cillia, Rudolf, Martin Reisigl in Ruth Wodak. »The discursive construction of national identities.« *Discourse & Society* 10, št. 2 (1999): 149–73. Pridobljeno 16. 6. 2025. https://doi.org/10.1177/0957926599010002002.

Duchesne, Sophie. »Who's afraid of banal nationalism?.« *Nations and Nationalism* 24, št. 4 (2018): 841–56. Pridobljeno 9. 6. 2025. https://doi.org/10.1111/nana.12457.

Edensor, Tim in Shanti Sumartojo. »Geographies of everyday nationhood: experiencing multiculturalism in Melbourne.« *Nations and Nationalism* 24, št. 3 (2018): 553–78. Pridobljeno 12. 6. 2025. https://doi.org/10.1111/nana.12421.

Edensor, Tim. *National Identity, Popular Culture and Everyday Life*. Oxford: Berg, 2002.

Eriksen, H. Thomas. »Formal and informal nationalism.« *Ethnic and Racial Studies* 16, št. 1 (1993): 1–25. Pridobljeno 3. 7. 2025. https://doi.org/10.1080/01419870.1993.9993770.

Foster, J. Robert. *Materializing the Nation: Commodities, Consumption, and Media in Papua New Guinea*. Bloomington: Indiana University Press, 2002.

Fox, E. Jon in Cynthia Miller-Idriss. »Everyday nationhood.« *Ethnicities* 8, št. 4 (2018): 536–76. Pridobljeno 5. 6. 2025. https://doi.org/10.1177/1468796808088925.

Fox, E. Jon in Cynthia Miller-Idriss. »The »here and now« of everyday nationhood.« *Ethnicities* 8, št. 4 (2018): 573–76. Pridobljeno 11. 6. 2025. https://doi.org/10.1177/1468796808008004010 3.

Fox, E. Jon in Maarten Van Ginderachter. »Everday nationalism's evidence problem.« *Nations and Nationalism* 24, št. 3 (2018): 546–52. Pridobljeno 5. 6. 2025. doi: https://doi.org/10.1111/nana.12418.

Fox, E. Jon. »Banal nationalism in everyday life.« *Nations and Nationalism* 24, št. 4 (2018): 862–66. Pridobljeno 7. 6. 2025. doi: 10.1111/nana.12458.

Fox, E. Jon. »The edges of the nation: a research agenda for uncovering the taken-for-granted foundations of everyday nationhood.« *Nations and Nationalism* 23, št. 1 (2017): 26–47. Pridobljeno 8. 6. 2025. https://doi.org/10.1111/nana.12269.

Giddens, Anthony. *Central Problems in Social Theory*. London: MacMillan, 1979.

Giddens, Anthony. *New Rules of Sociological Method*. London: Hutchinson, 1976.

Giddens, Anthony. *Studies in Social and Political Theory*. New York: Basic Books, 1977.

Giddens, Anthony. *The Constitution of Society: Outline of the Theory of Structuration*. Cambridge: Polity Press, 1984.

Goffman, Erving. *Predstavljanje sebe v vsakdanjem življenju*. Ljubljana: Studia humanitatis, 2014.

Goode, J. Paul in R. David Stroup. »Everday nationalism: constructivism for the masses.« *Social Science Quarterly* 96, št. 3 (2015): 717–39. Pridobljeno 8. 6. 2025. https://doi.org/10.1111/ssqu.12188.

Haralambos, Michael in Martin Holborn. *Sociologija: teme in pogledi*. Ljubljana: Državna založba Slovenije, 1999.

Hearn, Jonathan in Marco Antonsich. »Theoretical and methodological considerations for the study of banal and everyday nationalism.« *Nations and Nationalism* 24, št. 3 (2018): 594–605. Pridobljeno 7. 6. 2025. https://doi.org/10.1111/nana.12419.

Hearn, Jonathan. »National identity: banal, personal and embedded.« *Nations and Nationalism* 13, št. 4 (2007): 657–74. Pridobljeno 21. 6. 2025. https://doi.org/10.1111/j.1469-8129.2007.00303.x.

Hobsbawm, J. Eric. *Nations and Nationalism since 1780: Programme, Myth, Reality*. Cambridge: Cambridge University Press, 1992.

Ichijo, Atsuko in Ronald Ranta. *Food, National Identity and Nationalism: From Everyday to Global Politics*. Basingstoke: Palgrave Macmillan, 2016.

Jones, Pip in Liz Bradbury. *Introducing Social Theory* (3rd edition). Cambridge: Polity Press. Pridobljeno 10. 6. 2025. https://research-ebsco-com.nukweb.nuk.uni-lj.si/c/rsg6t3/search/details/ecieo34n3v?db=nlebk.

Judson, M. Pieter. *Guardians of the Nation – Activists on the Language Frontiers of Imperial Austria*. Cambridge MA: Harvard University Press, 2006.

Kaufmann, Eric. »Complexity and nationalism.« *Nations and Nationalism* 23, št. 1 (2017): 6–25. Pridobljeno 26. 6. 2025. https://doi.org/10.1111/nana.12270.

Knott, Eleanor. »Everyday nationalism. A review of the literature.« *Studies on National Movements (SNM)* 3 (2015). Pridobljeno 8. 6. 2025. https://test.snm.nise.eu/index.php/studies/article/view/0308s.

Koch, Natalie in Tom Perreault. »Resource nationalism.« *Progress in Human Geography* 43, št. 4 (2019): 611–31. Pridobljeno 27. 6. 2025. https://doi.org/10.1177/0309132518781497.

Kosi, Jernej. »'Yugoslavia has nothing. Yugoslavia has no bread. But Hungary gives us bread': Access to food and (dis)loyalty in a 'redeemed' Yugoslav borderland.« *Austrian History Yearbook* 55 (2024): 283–97. Pridobljeno 3. 10. 2025. https://doi.org/10.1017/S0067237824000055.

Malašević, Siniša. *Grounded Nationalisms*. Cambridge: Cambridge University Press, 2019

Mandelc, Damjan. *Na mejah nacije: teorije in prakse nacionalizma*. Ljubljana: Filozofska fakulteta, 2011.

Mann, Robin in Steve Fenton. »The personal contexts of national sentiments.« *Journal of Ethnic and Migration Studies* 35, št. 4 (2009): 517–34. Pridobljeno 10. 7. 2025. https://doi.org/10.1080/13691830902764882.

McGarry, Orla. »Knowing 'how to go on': structuration theory as an analytical prism in studies of intercultural engagement.« *Journal of Ethnic and Migration Studies* 42, št. 12 (2016): 2067–85. Pridobljeno 13. 6. 2025. https://doi.org/10.1080/1369183X.2016.1148593.

Miller-Idriss, Cynthia. »Everday understanding of citizenship in Germany.« *Citizenship Studies* 10, št. 5 (2006): 541–70. Pridobljeno 3. 6. 2025. https://doi.org/10.1080/13621020600954978.

Morawska, Eva. »International migration: its various mechanisms and different theories that try to explain.« *Willy Brandt series of working papers IMER/MIM* (2007). Pridobljeno 16. 6. 2025. https://www.diva-portal.org/smash/get/diva2:1409965/FULLTEXT01.pdf.

National Indifference and the History of Nationalism in Modern Europe, uredila Maarten Van Ginderachter in Jon E. Fox. London, New York: Routledge, 2019.

O'Reilly, Karen. »Structuration, practice theory, ethnography and migration: bringing it all together.« *IMI Working paper series* 61 (2012). Pridobljeno 15. 6. 2025. https://ora.ox.ac.uk/objects/uuid:f7ffb7f9-d6d0-4601-95e1-156da3c714a3/files/sbz60cw729.

Reicher, Stephen in Nick Hopkins. *Self and Nation: Categorization, Contestation and Mobilization*. Thousand Oakes: SAGE, 2001.

Skey, Michael. »The national in everyday life: A critical engagement with Michael Billig's thesis of Banal nationalism.« *The Sociological Review* 57, št. 2 (2009): 331–46. Pridobljeno 7. 6. 2025. https://doi.org/10.1111/j.1467-954X.2009.01832.x.

Skey, Michael. »'There are times when I feel like a bit of an alien': middling migrants and the national order of things.« *Nations and Nationalism* 24, št. 3 (2018): 606–23. Pridobljeno 12. 6. 2025. 10.1111/nana.12422.

Skey, Michael. *National Belonging and Everyday Life: The Significance of Nationhood in an Uncertain World*. Basingstoke: Palgrave Macmillan, 2011.

Smith, Anthony. »The limits of everday nationhood.« *Ethnicities* 8, št. 4 (2018): 563–73. Pridobljeno 8. 6. 2025. https://doi.org/10.1177/1468796808008004102.

Smith, Anthony. *National Identity*. London: Penguin, 1991.

Stankovič, Peter. »Giddensova teorija strukturacije: zagate teoretskega eklekticizma.« *Teorija in praksa* 37, št. 3 (2000): 455–74. Pridobljeno 11. 6. 2025. http://dk.fdv.uni-lj.si/db/pdfs/tip20003stankovic.pdf.

Stones, Rob. *Structuration Theory.* Basingstoke: Palgrave Macmillan, 2005.

Surak, Kristin. *Making Tea, Making Japan: Cultural Nationalism in Practice*. Redwood City: Stanford University Press, 2012.

Študije o etnonacionalizmu, uredil Rudi Rizman. Ljubljana: KRT, 1991.

Thompson, Andrew. »Nations, national identities and human agency: putting people back into nations.« *The Sociological Review* 49, št. 1 (2001): 18–32. Pridobljeno 23. 6. 2025. https://doi.org/10.1111/1467-954X.00242.

Van Ginderachter, Maarten. »How to guage banal nationalism and national indifference in the past: proletarian tweets in Belgium's belle epoque.« *Nations and Nationalism* 24, št. 3 (2018): 579–593. Pridobljeno 16. 6. 2025. https://doi.org/10.1111/nana.12420.

Whitmeyer, M. Joseph. »Elites and popular nationalism.« *British Journal of Sociology* 53, št. 3 (2002): 321–41. Pridobljeno 12. 7. 2025. https://doi.org/10.1080/0007131022000000536.

Zahra, Tara. *National Indifference and the Battle for Children in the Bohemian Lands, 1900–1948.* Ithaca: Cornell University Press, 2011

Zajc, Marko. »Josip Jurčič's tradition in Muljava: The boundaries of localism and nationalism.« *Prispevki za novejšo zgodovino* 53, št. 2 (2013): 23–36. Pridobljeno 3. 10. 2025. https://ojs.inz.si/pnz/article/view/65/66.

Nik Obid

EVERYDAY AND BANAL NATIONALISM BETWEEN STRUCTURE AND AGENCY

SUMMARY

The article focuses on presenting the concepts of banal and everyday nationalism, which mark a conceptual shift from the classic approaches of studying nationalism. It aims to connect both approaches to the sociological theories that transcend the classic theoretical division between structural approaches and theories of social action. This division also influenced the study of nationalism, as researchers typically emphasised its top-down structural and ideological influence, often overlooking the role of the individual in understanding and (re)shaping it. While classic approaches generally focus on the structural nature of nationalism and its elite-driven discourse, the everyday and banal nationalism approaches mainly examine its impact and role at the level of individual members of society. The essence of banal nationalism is to understand how the ideology of nationalism spreads from the structural to the individual level, while everyday nationalism focuses on how actors interpret, use, and transform national categories in daily life contexts.

The first part of the contribution thus focuses on presenting sociological theories that study the relationship, complexity, and interconnectedness of structure and agency. It highlights the theories of Pierre Bourdieu, Anthony Giddens, and Margaret Archer, who all attempted to connect the influence of structure and agency on social dynamics in various ways. It then connects these theories to the fundamental works on banal and everyday nationalism, providing basic insights into the research methods and aspects of their study from both historical and contemporary perspectives. In the final chapter, the contribution emphasises the complexity of studying nationalism at the everyday level, due to the interplay of structural influences and the individual's conscious and unconscious use of national categories in daily life. It highlights the strengths and weaknesses of structuration theory, which aims to go beyond understanding society through two separate poles, enabling the synthesis of individual study aspects, the inclusion of intermediate structural layers, and the identification of the reciprocal influence of structure and agency on maintaining the relevance of national categories within a complex postmodern society.

1.02 DOI: https://doi.org/10.51663/pnz.65.3.13

Klemen Kocjančič*

From Camp Followers to Leaders: A Historical Evolution of the Role of Women in the Military

IZVLEČEK

OD SPREMLJEVALK TABOROV DO VODITELJIC: ZGODOVINSKI RAZVOJ VLOGE ŽENSK V OBOROŽENIH SILAH

Članek predstavlja zgodovinski razvoj vloge žensk v oboroženih silah, ki so v tisočletjih človeškega obstoja v oboroženih silah opravljale različne naloge. Sprva so se priključevale oziroma so bile priključene vojskam na pohodu, kjer so spremljale vojake in predvsem izvajale podporne naloge. V starem in srednjem veku se je vloga žensk bolj malo spremenila, kljub temu pa so se občasno pojavile ženske vojaške voditeljice. Šele v novem veku se je vloga žensk pričela spreminjati: sprva so pridobile formalno vlogo v vojaško-zdravstvenem sistemu, nato pa so začele prevzemati tudi bojno-podporne vloge. Med prvo in drugo svetovno vojno so ženske postale pomemben člen vojaške industrije in organizacije, vključno z neposrednim sodelovanjem v spopadih in oblikovanjem popolnoma ženskih bojnih enot. A šele po drugi svetovni vojni so pričele prevzemati tudi vodstvene funkcije v oboroženih silah.

Ključne besede: oborožene sile, ženske v oboroženih silah, vojaška zgodovina, resolucija Varnostnega sveta Združenih narodov 1325, delavska zgodovina

* **PhD, Research Associate, University of Ljubljana, Faculty of Social Sciences, Kardeljeva ploščad 5, SI-1000 Ljubljana; Ministry of Defence of the Republic of Slovenia, Undersecretary, Vojkova cesta 55, SI-1000 Ljubljana, klemenkocjancic@gmail.com**

ABSTRACT

The article traces the historical development of women's roles in the armed forces, emphasising their participation in various military tasks throughout human history. Originally, women were attached to armies on the march, accompanying soldiers and mainly performing support roles. In Antiquity and the Middle Ages, women's roles in the military remained relatively unchanged, despite the occasional emergence of female military leaders. It was only in modern times that this began to shift. Initially, women were assigned formal roles within the military medical system, while later, they also took on support roles in combat. During World Wars I and II, women became an essential part of the military industry and organisation, and they started to participate directly in combat operations. The first exclusively female combat units were established. However, it was not until after World War II that women started to take on leadership roles within the armed forces.

Keywords: armed forces, women in the armed forces, military history, UNSC Resolution 1325, history of labour

Introduction

Women's roles in the military have never been static, but they have generally been overlooked historically. This article contends that the broader development of this role is neither straightforward nor unavoidable. The inclusion of women has been influenced by the evolving nature of warfare, various social and legal changes, and, in recent decades, institutional decisions within the armed forces. By examining this progression – "from camp followers to leaders" – the article demonstrates how women's participation has grown in scope (from informal support to command positions), depth (from spontaneous contributions to formalised military careers), and significance (from auxiliary roles to a professional identity).

In Antiquity, women were seen as a supplementary part of the military: the so-called "camp followers", consisting of wives, traders, laundresses, nurses, and entertainers, who accompanied armies during their campaigns. Historically, their efforts were structurally invisible yet essential, particularly in provisioning, care, and morale. Such an unofficial status of women in the military was common in the era of pre-industrial warfare. Some women also joined the military ranks, often disguised as men. In the early modern era, women officially took on certain roles in the military, following the specialisation and unification of education and the professionalisation of some jobs. Initially, women were included to provide nursing and medical support, aligning with the professionalisation of military health services, where women played prominent roles,[1] though usually without parity with their male counterparts or clear paths to command.

1 For example, see Sharon S. Dittmar et al., "Images and Sensations of War: A Common History of Military Nursing," *Health Care for Women International* 17, No. 1 (1996): 69–80, https://doi.org/10.1080/07399339609516221.

Full mobilisation of entire societies during the world wars transformed both the demand for and the perception of women's military capabilities. From taking on various jobs in the military industry to being included in auxiliary and combat-support branches (communications, intelligence, logistics, air defence), women eventually also formed all-female combat units.[2] During World War II, Soviet women, in particular, served as snipers, pilots, and partisans, while resistance movements across occupied Europe relied heavily on their contributions.[3] After World War II, women were once again largely excluded from combat duties, though their roles in combat support were gradually expanded and formalised. The social changes in the second half of the 20th century led many armed forces to open most occupational specialisations to women and to begin, though again unevenly, addressing the previously limited or even non-existent promotion of women and filling leadership positions.[4]

A key geopolitical milestone in this development was United Nations Security Council Resolution 1325 (2000), which articulated the Women, Peace and Security (WPS) agenda. The UNSCR 1325 redefined women not only as victims needing protection but also as vital participants in peace processes and security organisations overall.[5] This included the military, which developed national action plans, enhanced gender advisor roles,[6] set integration standards in peace operations and military services, specialities, or branches,[7] and started training on gender perspectives in planning and rules of engagement.[8] The Resolution also had a broader social influence, as seen

2 For example, see Jeremy A. Crang, *Sisters in Arms: Women in the British Armed Forces during the Second World War* (Cambridge: Cambridge University Press, 2020). Beate Fieseler, M. Michaela Hampf, and Jutta Schwarzkopf, "Gendering combat: Military women's status in Britain, the United States, and the Soviet Union during the Second World War," *Women's Studies International Forum* 47 (2014): 115–26, https://doi.org/10.1016/j.wsif.2014.06.011. Ursula von Gersdorff, Frauen im Kriegsdienst, 1914–1945 (Stuttgart: Deutsche Verlags-Anstalt, 1969).

3 Kristal L. M. Alfonso, Femme Fatale: *An Examination of the Role of Women in Combat and the Policy Implications for Future American Military Operations* (Maxwell Air Force Base: Air University Press, 2009), 7–19. Anna Krylova, *Soviet Women in Combat: A History of Violence on the Eastern Front* (Cambridge, New York: Cambridge University Press, 2010). Rochelle Nowaki, Nachthexen: Soviet Female Pilots in WWII, Hohonu 13 (2015): 56–62. Ingrid Strobl, *Partisanas: Women in the Armed Resistance to Fascism and German Occupation (1936–1945)* (Edinburgh, West Virginia: AK Press, 2008).

4 Sandra Carson Stanley and Mady Wechsler Segal, "Military Women in NATO: An Update," *Armed Forces and Society* 14, No. 4 (1988): 559–85.

5 For example, see Marius-Emanuel Caragea, "Modern Challenges to Military Management. UN Security Council Resolution 1325 'Women, Peace and Security'," *Management & Marketing* 21, No. 2 (2023): 312–27. Jane Derbyshire, "An Analysis and Critique of the UNSCR 1325 Resolution – What are Recommendations for Future Opportunities?," *Sodobni vojaški izzivi* 18, No. 3 (2016): 83–93, https://doi.org/10.33179/BSV.99.SVI.11.CMC.18.3.7.

6 For example, see Megan Bastick and Claire Duncanson, "Agents of Change? Gender Advisors in NATO Militaries," *International Peacekeeping* (2018): 1–24, https://doi.org/10.1080/13533312.2018.1492876.

7 For example, see Pablo Castillo Diaz, "Military Women in Peacekeeping Missions and the Politics of UN Security Council Resolution 1325," *Sodobni vojaški izzivi* 18, No. 3 (2016): 23–34, https://doi.org/10.33179/BSV.99.SVI.11.CMC.18.3.3. Nadja Fulan Štante, "Strenghts and Weaknesses of Women's Religious Peace-Building (in Slovenia)," *Annales* 30, No. 3 (2020): 343–54, https://doi.org/10.19233/ASHS.2020.21. Jovanka Šaranović, Brankica Potkonjak-Lukić, and Tatjana Višacki, "Achievements and Perspectives of the Implementation of UNSCR 1325 in the Ministry of Defence and the Serbian Armed Forces," *Sodobni vojaški izzivi* 18, No. 3 (2016): 65–81, https://doi.org/10.33179/BSV.99.SVI.11.CMC.18.3.6. Suzana Tkavc, "Some of the Best Practices in Gender Perspective and the Implementation of UNSCR 1325 in the 25 Years of Slovenian Armed Forces," *Sodobni vojaški izzivi* 18, No. 3 (2016): 45–63, https://doi.org/10.33179/BSV.99.SVI.11.CMC.18.3.5.

8 For example, see Jamie Leonheart, "Gender Perspectives for Operational Effectiveness – An Opportunity for U.S. Forces Japan and the Japan Self Defense Forces," *NIDS Commentary,* No. 333 (2024).

in military scholarships for women and various related gender issues. These scholarships highlight aspects such as the effectiveness and cohesion of military organisations in which women served;[9] the organisational culture related to women in the military;[10] the career advancement of women in the armed forces;[11] and civil-military aspects, especially regarding the role of politics.[12]

Although these scholarships initially concentrated solely on women, they later expanded to include research on gender roles[13] and integration in specific operational units,[14] the effects of military service on health,[15] the roles of veterans,[16] military

9 For example, see Uzi Ben-Shalom, Eyal Lewin, and Shimrit Engel, "Organizational Processes and Gender Integration in Operational Military Units: An Israel Defense Forces Case Study," *Gender, Work & Organization* 26, No. 9 (2019): 1289–1303, https://doi.org/10.1111/gwao.12348. Robert Egnell, Petter Hojem, and Hannes Berts, *Gender, Military Effectiveness, and Organizational Change: The Swedish Model* (Houndsmills, New York: Palgrave Macmillan, 2014). Mady Wechler Segal et al., "The Role of Leadership and Peer Behaviors in the Performance and Well-Being of Women in Combat: Historical Perspectives, Unit Integration, and Family Issues," Military Medicine 181 (2016): 1–28, https://doi.org/10.7205/MILMED-D-15-00342.

10 For example, see Melissa T. Brown, "'A Woman in the Army Is Still a Woman': Representations of Women in US Military Recruiting Advertisements for the All-Volunteer Force," *Journal of Women, Politics & Policy* 33 (2012): 151–75. Nadja Furlan Štante, "Ženske v oboroženih silah: Med nasiljem in ranljivostjo," S*odobni vojaški izzivi* 18, No. 3 (2016): 95–105, https://doi.org/10.33179/BSV.99.SVI.11.CMC.18.3.8.

11 For example, see J. Norman Baldwin, "Female Promotions in Male-Dominant Organizations: The Case of the United States Military," *The Journal of Politics* 58, No. 4 (1996): 1184–97. Nani Kusmiyati and Hady Efendy, "The Leadership of Women in Military on Military Organization," *International Journal of Human Resource Studies* 7, No. 4 (2017): 165–74.

12 For example, see Bradford Booth, William W. Falk, David R. Segal, and Mady Wechsler Segal, "The Impact of Military Presence in Local Labor Markets on the Employment of Women," *Gender & Society* 14, No. 2 (2000): 318–32. Joan Chandler, Lyn Bryant, and Tracey Bunyard, "Women in Military Occupations," *Work, Employment & Society* 9, No. 1 (1995): 123–35.

13 For example, see Cati Connell, A Few Good Gays: *The Gendered Compromises behind Military Inclusion* (Oakland: University of California Press, 2023). Máel Embser-Herbert and Bree Fram, eds., *With Honor and Integrity: Transgender Troops in Their Own Words* (New York: New York University Press, 2021). Pavel Vuk and Saša Galičič, "Socialna diverziteta v luči inkluzivnosti istospolno usmerjenih pripadnic in pripadnikov v slovenski vojski," *Teorija in praksa* 59, No. 2 (2022): 568–88, 596, 597. Pavel Vuk, "The Slovenian Armed Forces Faces the Challenge of Inclusion of Their Homosexual Members," *Journal of Homosexuality* 71, No. 5 (2024): 1231–52, https://doi.org/10.1080/00918369.2023.2169088.

14 For example, see Frank Gasca, Ryan Voneida, and Ken Goedecke, "Unique Capabilities of Women in *Special Operations Forces,*" *Special Operations Journal* 1, No. 2 (2015): 105–11, https://doi.org/10.1080/23296151.2015.1070613. Karmen Poklukar and Pavel Vuk, "Vključevanje žensk v specialne sile," *Sodobni vojaški izzivi* 22, No. 4 (2020): 85–105, https://doi.org/10.33179/BSV.99.SVI.11.CMC.22.4.5.

15 For example, see Morgan K. Anderson et al., "Effect of Mandatory Unit and Individual Physical Training on Fitness in Military Men and Women," *American Journal of Health Promotion* (2016), 1–10, https://doi.org/10.1177/0890117116666977. Beveryl P. Bergman and Simon A. St J. Miller, "Equal Opportunities, Equal Risks? Overuse Injuries in Female Military Recruits," *Journal of Publich Health Medicine* 23, No. 1 (2001): 35–39. Carissa van den Berk Clark, Jennifer Chang, Jessica Servery, and Jeffrey D. Quinlan, "Women's Health and the Military," *Primary Care: Clinics in Office Practice* 45, No. 4 (2018): 677–86, https://doi.org/10.1016/j.pop.2018.07.006.

16 For example, see Julia Baumann, Charlotte Williamson, and Dominic Murphy, "Exploring the Impact of Gender-Specific Challenges during and after Military Service on Female UK Veterans," *Journal of Military, Veteran and Family Health* 8, No. 2 (2022): 72–81, https://doi.org/10.3138/jmvfh-2021-0065. Valerija Bernik, "Veteranke druge svetovne vojne," *Sodobni vojaški izzivi* 19, No. 2 (2017): 71–87, https://doi.org/10.33179/BSV.99.SVI.11.CMC.19.2.5. Alison S. Fell, *Women as Veterans in Britain and France after the First World War* (Cambridge: Cambridge University Press, 2018).

families,[17] and more. Such research is often linked with other social determinants, such as age, race, and education.[18] Additionally, a new area of study has emerged concerning sexuality: issues like sexual violence in armed conflicts,[19] sexual violence within the military,[20] military prostitution,[21] and related topics.

This article's contribution is twofold. Firstly, it provides a synthetic historical narrative that connects the logistics-heavy yet unofficial roles of camp followers to leadership positions in modern militaries, emphasising institutional learning. Secondly, it promotes the institutional-process model that links external influences (such as total war, legal mandates like UNSCR 1325, and technological advances) to internal reforms (including occupational access, training standards, and evaluation and promotion rules) and, ultimately, leads to women attaining leadership positions in the military. Methodologically, by analysing the literature discussing the various roles of women in armed forces, the article fulfils both descriptive and explanatory objectives regarding the long-term transformation of women's roles in the military throughout history, particularly in the contemporary era.

Followers of Military Camps

In ancient times, women's roles in the military could be divided into two categories: camp followers who accompanied military units during campaigns, or residents based in or near (semi)permanent military installations.

The perception of women in military matters in the Ancient Greek world stemmed from the division between public and private life, with military affairs regarded as part

17 For example, see Donabelle C. Hess, "Military Family Readiness: The Importance of Building Familial Resilience and Increasing Family Well-being Through Military Community Support and Services," *Sodobni vojaški izzivi* 22, No. 2 (2020): 89–99, https://doi.org/10.33179/BSV.99.SVI.11.CMC.22.2.5. Ljubica Jelušič, Julija Jelušič Južnič, and Jelena Juvan, "The Relevance of Military Families for Military Organizations and Military Sociology," *Sodobni vojaški izzivi* 22, No. 2 (2020): 51–67, https://doi.org/10.33179/BSV.99.SVI.11.CMC.22.2.3. Jelena Juvan, "Usklajevanje delovnih in družinskih obveznosti v vojaški organizaciji," *Socialno delo* 48, No. 4 (2009): 227–34. Kairi Kasearu et al, "Military Families in Estonia, Slovenia and Sweden: Similarities and Differences," *Sodobni vojaški izzivi* 22, No. 2 (2020): 69–87, https://doi.org/10.33179/BSV.99.SVI.11.CMC.22.2.4. Janja Vuga Beršnak and Bojana Lobe, "Socioecological Model of a Military Family's Health and Well-being: Inside a Slovenian Military Family," *Armed Forces and Society* 50, No. 1 (2024): 224–52, https://doi.org/10.1177/0095327X221115679.

18 For example, see Bradford Booth, and David R. Segal, "Bringing the Soldiers Back In: Implications of Inclusion of Military Personnel Market Research on Race, Class, and Gender," *Race, Gender & Class* 12, No. 1 (2005): 34–57. Sandra Bolzenius, "Asserting Citizenship: Black Women in the Women's Army Corps (WAC)," *International Journal of Military History and Historiography* 39 (2019): 208–231.

19 For example, see Sabine Hirschauer, *The Securitization of Rape: Women, War and Sexual Violence* (Houndmills, New York: Palgrave Macmillan, 2014). Inger Skjelsbaek, "Sexual Violence and War: Mapping Out a Complex Relationship," E*uropean Journal of International Relations* 7, No. 2 (2001): 211–37.

20 For example, see Vicki J. Magley et al., "The Impact of Sexual Harassment on *Military Personnel*: Is It the Same for Men and Women?," *Military Psychology* 11, No. 3 (1999): 283–302, https://doi.org/10.1207/s15327876mp1103_5. John B. Pryor, "The Psychological Impact on Sexual Harassment on Women in the U.S. Military," *Basic and Applied Social Psychology* 17, No. 4 (1995): 581–603, https://doi.org/10.1207/s15324834basp1704_9.

21 For example, see Hata Ikuhiko, *Comfort Women and Sex in the Battle Zone* (Lanham, Boulder, New York, London: Hamilton Books, 2018). Erik Ropers, "Representation of Gendered Violence in Manga: The Case of Enforced Military Prostitution," *Japanese Studies* 31, No. 2 (2011): 249–66.

of the public life, which was the domain of men. Naturally, women were indeed present during military sieges of cities. For example, during the siege of the Greek city of Gela in Sicily in 405 BC, women and children "actively helped the defenders, particularly by taking part in restoring damaged sections of the town walls". The women's role in military activities at home could include cooking, weaving, and fulfilling other basic needs of soldiers, as well as participating in the military "industry" by manufacturing ammunition (spears, arrows) and armour. Later, women started to follow soldiers during their campaigns. Thucydides writes that during the Peloponnesian War, "roughly one woman was assigned to prepare food (and presumably to take care of other non-combatant tasks) for every four men," thus women made up one-fifth of the campaign personnel. Ancient Greek sources also report that women cared for wounded warriors and were involved in disinformation operations. For example, during the siege of Sinope around 379/78 BC, "the townswomen clad themselves in dummy armor and joined their men on the walls, making the defenders look more numerous than they actually were". The involvement of ancient Greek women was a result of the total war concept, where the entire community was "involved in the military affairs of the community".[22]

The ancient Roman military had a complex relationship with sutlers, who could be slaves or free people. They followed the legions or lived near or inside military camps due to familiar connections or economic reasons, including peddlers, merchants (also known as sutlers), prostitutes, artisans, prophets and diviners, and foragers for food and firewood, among others. During peacetime, the presence of women was not an issue. However, during combat operations or in cases of poor discipline, female sutlers could be expelled from the vicinity of military units. Generally, sutlers offered logistical support but could also be a burden. Senior military officers had the privilege of having their families live with them in military camps.[23] Civilian supporters, especially merchants, played a crucial role in connecting legionnaires with the local community and beyond.[24] Roman legionnaires sometimes encountered female warriors when fighting against foreign tribes. One famous example was Boudica, a queen who led a failed revolt by the British Iceni tribe against Rome in AD 61 or 60.[25] Interestingly, women fought alongside and against men as gladiatrixes.[26]

22 Jorit Wintjes, "'Keep the Women Out of the Camp!' Women and Military Institutions in the Classical World," in Barton C. Hacker and Margaret Vining, eds., *A Companion to Women's Military History* (Leiden, Boston: Brill, 2012), 21–30.

23 Penelope M. Allison, "Mapping for Gender. Interpreting Artefact Distribution Inside 1st- and 2nd-Century A.D. Forts in Roman Germany," *Archaeological Dialogues* 13, No. 1 (2006): 1–20, https://doi.org/10.10.17/S1380203806211851. Chiara Cenati and Peter Kruschwitz, "Poetic Baggage: Representations of Camp Followers in the Latin Verse Inscriptions," *Electrum* 31 (2024): 153–83, https://doi.org/10.4467/20800909EL.24.012.19162.

24 Ben Kolbeck, "A Foot in Both Camps: The Civilian Suppliers of the Army in Roman Britain," *Theoretical Roman Archaeology Journal* 1, No. 1 (2018): 1–19, https://doi.org/https://doi.org/10.16995/traj.355.

25 Valentine J. Belfiglio, "Women and the Ancient Roman Army," *Journal of Clinical Research and Case Studies* 1, No. 1 (2023): 2. Graham Webster, Boudica: *The British Revolt against Rome AD 60* (London: B. T. Batsford, 1993), 46.

26 Stephan Brunet, "Women with Swords: Female Gladiators in the Roman World," in Paul Cristesen and Donald G. Kyle, eds., *A Companion to Sport and Spectacle in Greek and Roman Antiquity* (Chichester: Wiley Blackwell, 2016), 478–91.

Similarly, medieval armies included women as well. The Viking raiding parties that invaded Britain also had female camp followers, who could either be free or slaves.[27] Female warriors, known as shieldmaidens, and mythical figures like Valkyries were also known to participate in raids. It is possible that some female warriors enjoyed high status, as shown by their burial arrangements.[28] Reports from the Rus' court noted that each warrior in the King's retinue had two slave girls.[29] Byzantine sources also mentioned the active participation of female warriors during the battles of Dorostolon (also known as Silistra) in 971, when Byzantine troops laid siege to Dorostolon, a Kievan Rus' fortress. After the fortress fell, the Byzantines "found women lying among the fallen, equipped like men; women who had fought against the Romans [e.g., Byzantines] together with the men".[30]

Early Modern Militaries and Women

Women were also a common sight in the early modern armies. As before, they cooked, cleaned, laundered, and provided nursing care to soldiers. When stationed in camps or settlements, they were responsible for housing, food, and other supplies. Their presence was necessary as they eased additional burdens on soldiers caused by logistical issues. During this period, women as queens or regents supported or led their armies in military campaigns. For example, Isabella I of Castile (1451–1504) actively supported the army of her husband, Ferdinand II of Aragon (1452–1516), while it was on the battlefield. She organised "a collection of supplies, the hiring of mercenaries, and the establishment of field hospitals" and visited the troops. Sometimes, in her husband's absence, she also commanded the soldiers. Similarly, Queen Elizabeth I of England (1533–1603) was involved in military administration and propaganda efforts. Even noble women would supplement or take over their husbands' roles in military administration.[31]

It was rare for women of humble (peasant) origins to join military campaigns and attain leadership positions. One example was Joan of Arc (1412?–1431). During the Hundred Years' War, in March 1429, she persuaded the French dauphin Charles to let her join his army in an effort to rescue the besieged town of Orleans. The victorious Battle of Orleans led to a new campaign aimed at capturing the Loire.

27 Dawn M. Hadley et al., "The Winter Camp of the Viking Great Army, AD 872–3, Torksey, Lincolnshire," *The Antiquaries Journal* 96 (2016): 54, https://doi.org/10.1017/s0003581516000718.

28 Jóhanna Katrín Friðriksdóttir, *Valkyrie: The Women of the Viking World* (London, New York: Bloomsbury Academic, 2020), 7–19. Neil Price et al, "Viking Warrior Women? Reassessing Birk Chamber Grave Bj.581," *Antiquity* 93, No. 367 (2019): 192–94, https://doi.org/10.15184/aqy.2018.258.

29 Ben Raffield, Neil Price, and Mark Collard, "Polygyny, Concubinage, and the Social Lives of Women in Viking-age Scandinavia," *Viking and Medieval Scandinavia* 13 (2017), 190, https://doi.org/10.1484/J.VMS.5.114355.

30 John Skylitzes, *A Synopsis of Byzantine History, 811–1057* (Cambridge: Cambridge University Press, 2010), 290.

31 Mary Elizabeth Ailes, "Camp Followers, Sutlers, and Soldiers' Wives: Women in Early Modern Armies (c. 1450–1650)," in Barton C. Hacker and Margaret Vining, eds., *A Companion to Women's Military History* (Leiden, Boston: Brill, 2012), 61–71.

Later, in September 1429, Joan took part in the unsuccessful attack on Paris and in the campaign for Compiègne in April–May 1430. On 23 May 1430, she was captured and sold to the English. She was then tried and burned as a heretic (for wearing male clothing) on 30 May 1431.[32]

The beginning of the 17th century marked a shift in Europe from aggregate contract armies to state commission armies, resulting in better professional logistical support and reducing the need for female camp followers. However, garrison communities persisted, along with women who followed various mercenary bands. Especially while attached to these bands, women began to take on non-traditional roles, such as "the custodians of the books and the money in small business", and managing plunder. During sieges, women assisted with more physically demanding tasks, such as "binding fascines, filling ditches, digging pits and mounting cannon in difficult places." With the professionalisation of armies, (unmarried) women, particularly prostitutes, became unwelcome and were even banned from military units. Soldiers' wives remained and performed traditional roles as laundresses, seamstresses, and nurses.[33]

Formalisation of Women's Roles

In the late 18th century, women were still present as camp followers, often as "soldiers' wives or consorts who more than earned their keep with foraging, cooking, laundry, needlework, and nursing." However, the Napoleonic Wars brought about many changes. The adoption of professional armies with extended military service resulted in smaller standing forces compared with the earlier mass conscripted armies. As a result, the presence of women within the military was greatly diminished, although they still supported soldiers as laundresses and canteen managers, particularly in permanent establishments. Another major factor contributing to the reduced number of women accompanying military units was industrialisation, which created more opportunities for women in factories and cities, while combat became more manoeuvrable and fast-paced. Whereas, in the past, camp followers were predominantly women of lower social standing, by the mid-19th century, more noble ladies, involved in advancing medical science and services, began appearing on the battlefield.[34]

During the Russo-Turkish War in Crimea, where the British forces supported Russia in the fight against the Ottoman Empire, public attention quickly turned to the terrible battlefield conditions and the inadequate medical services provided to the

32 Deborah A. Fraioli, *Joan of Arc and the Hundred Years War* (Westport Connecticut, London: Greenwood Press, 2005), 97–101.

33 John A. Lynn II, "Essential Women, Necessary Wives, and Exemplary Soldiers: The Military Reality and Cultural Representation of Women's Military Participation (1600–1815)," in Barton C. Hacker and Margaret Vining, eds., *A Companion to Women's Military History* (Leiden, Boston: Brill, 2012), 94–113.

34 Barton C. Hacker, "Reformers, Nurses, and Ladies in Uniform: The Changing Status of Military Women (c. 1815–c. 1914)," in Barton C. Hacker and Margaret Vining, eds., *A Companion to Women's Military History* (Leiden, Boston: Brill, 2012), 137–41. Jan Kilián, "A Soldier and a Townsman during the Thirty Years' War. Coexistence – Confrontation – Cooperation," *Przeglad Zachodniopomorski* 63, No. 4 (2019): 51, 52, https://doi.org/10.18276/pz.2019.4-02.

sick and wounded soldiers. The British Secretary of War asked Florence Nightingale (1820–1910) to organise a volunteer group of female nurses to improve these conditions. She was accompanied by 38 women, a priest, and a courier. With their help, she laid the foundations for modern military nursing. Nightingale's team improved sanitary conditions in military hospitals, particularly addressing outbreaks of infectious diseases, and began ordering essential supplies to improve medical treatment. While her companions focused on nursing, Nightingale was mainly involved in health management, education, and administration. Interestingly, battlefield conditions affected both sides to reach the same conclusion; only some days after Nightingale's arrival in Crimea, on the Russian side, Grand Duchess Helena Pavlovna (1806–1873) formed the Community of the Cross of the Sisters Caring for the Wounded and Sick Warriors, the Russian equivalent of Nightingale's group.[35]

The World Wars

During World War I, women's roles in military contexts expanded significantly, reflecting the unprecedented mobilisation of entire societies. Along with the changing structure of militaries and the expanding bureaucracy of military administration, the need for essential tasks such as nursing, medical support, communications, and logistical support provided new opportunities for women. Apart from working directly within the armed forces, even more women replaced male workers in related industries, especially munitions production.[36] Following Nightingale's example, volunteer organisations such as the Voluntary Aid Detachments in Britain and the American Red Cross successfully organised and deployed tens of thousands of women to frontline aid stations and hospitals. Many female physicians worked behind the frontlines, caring for sick and wounded soldiers. In France, female radiology assistants managed stationary and mobile X-ray units. Some countries granted temporary officer ranks to female physicians. Although there were cases of women joining the military in disguise, some served openly as women. Such rare instances were accepted in Russia, where individual women were initially allowed to serve unofficially in medical, reconnaissance, cavalry, infantry, and artillery units. Then, in 1917, a Women's Battalion of Death was formed; while later, several other all-female military units were established. Military intelligence services also began employing women as clerks, historians, translators, cryptographers, couriers, and intelligence agents. Another important development was the creation of numerous paramilitary and volunteer organisations, enabling women to contribute to the war effort from the home front. While World War I encouraged

35 Fatima Jasim Mohammed Ali, "The Russo-Turkish War in Crimea and Nightingale's Role in It (1854–1855)," *World Journal of Advanced Research and Reviews* 18, No. 2 (2023), https://doi.org/10.30574/wjarr.2023.18.2.0746. T. S: Sorokina, "Russian Nursing in the Crimean War," *Journal of the Royal College of Physicians of London* 29, No. 1 (1995): 57–63. Ugurgul Tunc, "Lessons from the Crimean War: How Hospitals were Transformed by Florence Nightingale and Others," *Infectious Diseases & Clinical Microbiology* 1, No. 2 (2019): 110–18, https://doi.org/10.36519/idcm.2019.19020.

36 Urška Strle, "K razumevanju ženskega dela v veliki vojni," *Prispevki za novejšo zgodovino* 55, No. 2 (2015): 103–25.

the professionalisation of certain roles, especially in military medicine, the traditional structures continued to hinder women's advancement and limited their command opportunities.[37] Interestingly, in the Kingdom of Serbia, Milunka Savić (1892–1973) joined the Serbian Army in disguise already during the First Balkan War in 1912. Her gender was revealed after she was treated for wounds sustained in combat. Due to her skills, she was permitted to remain in military service. She continued serving throughout World War I and became the most decorated female soldier of that war.[38]

Most newly formed (para)military organisations in which women participated during World War I were disbanded afterwards. Typically, women left their military positions and jobs in the military industry, although some continued serving in the traditional military occupations like nursing and administration. When World War II broke out, the story of women's participation in the military effort was repeated, but on a much larger and more diverse scale. In the United Kingdom, all-female auxiliary military organisations were established, such as the Women's Auxiliary Air Force, the Auxiliary Territorial Service, and the Women's Royal Naval Service. Over 600,000 women joined these organisations, serving in various roles: cooks, clerks, orderlies, equipment assistants, drivers, balloon fabric workers, storewomen, writers, communication workers, and more. Some women even joined mixed anti-aircraft batteries and intelligence units. They played such a vital role in the war effort that after the conflict and demobilisation, these female branches became permanent (Women's Royal Air Force; Women's Royal Army Corps; Women's Royal Naval Service), offering women opportunities for military careers beyond nursing.[39]

After the United States entered the war, they followed the British example by establishing female branches such as the Women's Army Corps, Women's Reserve of the Naval Reserve, Women's Reserve of the Coast Guard Reserve, and Marine Corps Women's Reserve, as well as female organisations supporting the war effort (like Women Airforce Service Pilots). These women performed similar duties to their British counterparts but faced additional challenges due to racial segregation.[40] In the Soviet Union, they also followed the example from World War I. Soviet women served

37 Kimberly Jensen, "Volunteers, Auxiliaries, and Women's Mobilization: The First World War and Beyond (1914–1939)," in Barton C. Hacker and Margaret Vining, eds., *A Companion to Women's Military History* (Leiden, Boston: Brill, 2012), 189–223. Janet Lee, "Sisterhood at the Front: Friendship, Comradeship, and the Feminine Appropriation of Military Heroism Among World War I First Aid Nursing Yeomanry (FANY)," *Women's Studies International Forum* 31 (2008): 16–29. Margaret Vining and Barton C. Hacker, "From Camp Follower to Lady in Uniform: Women, Social Class and Military Institutions before 1920," *Contemporary European History* 10, No. 3 (2001): 353–73, https://doi.org/10.1017/S0960777301003022.

38 Vidoje D. Golubović, *Dobrovoljka Milunka Savić: Srpska heroina* (Beograd: Udruženje ratnih dobrovoljaca 1912–1918, njihovih potomaka i poštovalaca, 2013), 10ff.

39 Crang, *Sisters in Arms*, 25f.. Helen Fry, *Women in Intelligence: The Hidden History of Two World Wars* (Yale: Yale University Press, 2023). Gerard J. de Groot, "Combatants or Non-Combatants? Women in Mixed Anti-Aircraft Batteries during the Second World War," *RUSI Journal* (1995): 65–70.

40 Bolzenius, "Asserting Citizenship," 210ff, Fieseler, Hampf and Schwarzkopf, "Gendering combat," 119–22. Laurie Schrivener, "U.S. Military Women in World War II: The SPAR, WAC, WAVEs, WASP, and Women Marines in U.S. Government Publications," *Journal of Governmental Information* 26, No. 4 (1999): 361–83. Margaret Vining, "Women Join the Armed Forces: The Transformation of Women's Military Work in World War II and After (1939–1947)," in Barton C. Hacker and Margaret Vining, eds., *A Companion to Women's Military History* (Leiden, Boston: Brill, 2012), 254.

as nurses, participated in labour and logistical roles, operated anti-aircraft guns, and served on the frontlines as snipers, fighter and bomber pilots, scouts, riflewomen, partisans, among others. All-female military units were also formed.[41] Even the British princess, the future Queen Elizabeth II (1926–2022), joined the Auxiliary Territorial Service in 1944 when she turned 18 and was trained as a mechanic and driver.[42]

The resistance movements fighting the Axis forces also accepted women. In these groups, women were not limited to traditional roles such as logistics, nursing, and communication, but also served as fighters and political commissars, and operated within the intelligence services. Some even rose to the rank of commanders.[43]

The story of women in the Third Reich was similar: although officially barred from military service, women worked in the military industry, as nurses, communication operators, clerks, air-defence personnel, and so on.[44]

The Integration of Women into the Military

While postwar demobilisation decreased women's military presence in most countries, the war highlighted their operational effectiveness across various domains, shaping future integration policies and the development of gender roles in armed forces.

The United States Women's Army Corps (WAC), established during World War II, continued operating after the war ended. Female soldiers took part in the Korean and Vietnam Wars. Influenced by the women's rights movement and the professionalisation of armed forces, a political debate arose about the role, function, and future of the WAC. In the final years of these Corps, they began to shut down some of its subordinate units, until in October 1978, the WAC was abolished. With this act, the United States Army addressed accusations of discrimination against women and subsequently fully integrated female soldiers into the military organisation.[45]

Other Western militaries also adopted similar solutions, transforming their forces into fully professional armed forces. Although compulsory military service persisted for men, women were excluded from it. For instance, in the former Yugoslavia, military service was mandatory only for men, while women had the option to voluntarily join the territorial defence forces, which were subordinated to the Yugoslav People's Army

41 Roger D. Markwick and Euridice Charon Cardona, *Soviet Women on the Frontline in the Second World War* (Houndmills, New York: Palgrave Macmillan, 2012), 20ff.

42 Vikki Hawkins, *A Princess at War: Queen Elizabeth II During World War II. The National WWII Museum*, 2021, accessed on 5 September 2025, https://www.nationalww2museum.org/war/articles/queen-elizabeth-ii-during-world-war-ii.

43 Valerija Bernik, "Ženske v slovenski partizanski vojski (1941–1945)," in Ljubica Jelušič and Mojca Pešec, eds., *Seksizem v vojaški uniformi* (Ljubljana: Obramboslovni raziskovalni center, Generalštab Slovenske vojske, 2002), 106–26.

44 Gersdorf, *Frauen im Kriegsdienst*, 27ff.

45 Bettie J. Morden, *The Women's Army Corps, 1945–1978* (Washington, D.C.: Center of Military History, 1990), 10–397.

(YPA). Between July 1983 and July 1985, the YPA even conducted a trial military training for women.[46]

The only Western country with compulsory military service for both men and women is Israel, mainly because of its smaller population compared to its hostile neighbours. While during the Israeli War of Independence, women took part in combat roles, later, in 1952, their military careers were largely restricted to educational and administrative positions. Another legal change implemented in 2000 opened 90% of positions in the Israeli Defence Forces to women as well, including combat roles (light infantry, search and rescue, etc.), and women even accounted for more than half (56%) of junior officer ranks.[47] In 2000, the Caracal Battalion was established as a mixed-gender operational infantry battalion.[48]

Another Western country, Norway, initially (1957–1978) allowed women to serve in its reserve forces, while in the event of war, they would replace men in administration, communication, or health services. In 1976, women could join Norway's regular forces, and by 1985, they had access to all military positions, including combat roles. As recently as 2015, mandatory military service was extended to women, although it is not universal like in Israel, as not all draftees are called to serve.[49]

In Western countries where most positions have been opened to women, the next step towards full integration involves including women in combat roles, similar to Israel, including the most specialised units like special forces.[50] Norway pioneered this effort when, in 2014, an all-female platoon of conscripts was trained as special forces. Since then, the *Jegertroppen* has specialised in urban special reconnaissance and close-quarters combat.[51]

Lately, participation in recent armed conflicts (e.g., the Global War on Terror) has led to female military members unintentionally being exposed to or involved in combat, often when their unit or base was ambushed or attacked.[52] On this occasion,

46 Maja Garb, "Ženske na služenju vojaškega roka v JLA," in Ljubica Jelušič and Mojca Pešec, eds., *Seksizem v vojaški uniformi* (Ljubljana: Obramboslovni raziskovalni center, Generalštab Slovenske vojske, 2002), 128–31.

47 Ephrat Huss and Julie Cwikel, "Women's Stress in Compulsory Army Service in Israel: A Gendered Perspective," *Work* 50, No. 1 (2015): 38, https://doi.org/10.3233/WOR-141930. Tair Karazi-Presler, Orna Sasson-Levy, and Edna Lomsky-Feder, "Gender, Emotions Management, and Power in Organizations: The Case of Israeli Women Junior Military Officers," *Sex Roles* 78 (2018): 573–86, https://doi.org/10.1007/s11199-017-0810-7.

48 Luke Carroll, "Raising a Female-centric Infantry Battalion: Do We Have the Nerve?," *Australian Army Journal* 11, No. 1 (2014): 40.

49 Sanna Strand, *The 'Scandinavian Model' of Military Conscription: A Formula for Democratic Defence Forces in 21st Century Europe?* (Vienna: Austrian Institute for International Affairs, 2021), 7–11, https://www.oiip.ac.at/cms/media/policy-analysis-scandinavian-model-of-military-conscription.pdf.

50 Anne Fieldhouse and T. J. O'Leary, "Integrating Women into Combat Roles: Comparing the UK Armed Forces and Israeli Defense Forces to Understand where Lessons can be Learnt," *BMJ Military Health* 169, No. 1 (2023): 78–80, https://doi.org/10.1136/bmjmilitary-2020-001500. Gasca, Voneida, and Goedecke, "Unique Capabilities." Poklukar and Vuk, "Vključevanje žensk."

51 Ingunn Helene Landsend Monsen, *Female Integration in Jordan's Special Forces – an Empirical Analysis of the Project's Content and Value for Norway and Jordan* (Oslo: Norwegian Defence Research Establishment, 2025), 25, 26.

52 Thomas A. Bruscino, "Palm Sunday Ambush, 20 March 2005", in William G. Robertson, ed., *In Contact! Case Studies from the Long War: Volume I* (Fort Leavenworth, Kansas: Combat Studies Institute Press, 2003) 59–82. Amy E. Street, Dawne Vogt, and Lissa Dutra, "A New Generation of Women Veterans: Stressors Faced by Women Deployed to Iraq and Afghanistan," *Clinical Psychology Review* 29 (2009): 686, https://doi.org/10.1016/j.cpr.2009.08.007.

a new type of women-focused military unit was established: the so-called female cultural support teams. In these teams, female military personnel were tasked with establishing contact with local women to help improve civil-military relations with the local population.[53]

Military Leaders

Another significant aspect of the full integration of women concerns leadership roles. For instance, in the United States, in 1920, women nurses were granted officer ranks but did not enjoy the same privileges and rights as their male counterparts. In 1942, women could hold leadership positions but only within the WAC. In 1967, restrictions on promoting women were lifted, and the following year, the first woman attained the highest enlisted rank of command sergeant major. By 1970, the first two women – Anna Mae Hays (1920–2018), Chief of the Army Nurse Corps, and Elizabeth P. Hoisington (1918–2007), Director of the Women's Army Corps – were promoted to the rank of general (brigadier general). At that point, women were permitted to command men, except in combat units. The first woman granted a combat command was Captain Linda Bray (1960), in 1989, who commanded a company during the invasion of Panama. In 2008, the first woman achieved the highest military rank during peacetime: Ann E. Dunwoody (1953) was promoted to the rank of general (OF-9) and simultaneously took command of a major United States Army unit. She was the first woman to do so. As recently as 2021, the first woman became a commander of a geographic combatant command when Laura J. Richardson (1963) assumed leadership of the United States Southern Command.[54] To date, no woman has been appointed to command a military branch or the entire armed forces of the United States.

In the Slovenian Armed Forces, female officers commonly held command of lower-level military units, up to the size of a company, while higher positions were generally inaccessible to women. Simultaneously, research indicated a preference for appointing male commanders to limited command roles.[55] However, in Slovenia, Alenka Ermenc (1963) made history within NATO militaries as the first woman to assume the highest military position, serving as the Chief of the General Staff of the Slovenian Armed Forces. In 2006, Ermenc became the first female to command a battalion (the 5th Intelligence-Reconnaissance Battalion). She was also the first woman

53 Naomi Head, "'Women Helping Women': Deploying Gender in US Counterinsurgency Wars in Iraq and Afghanistan," *Security Dialogue* 55, No. 2 (2023), https://doi.org/10.1177/09670106231203839. Rosellen Roche et al., "The Unseen Patriot: Female Cultural Support Team Members and Combat Definition," *Journal of Veteran Studies* 7, No. 1 (2021): 271–79, https://doi.org/10.21061/jvs.v7i1.285.

54 Army Women's Foundation, *Army Women In History*, accessed on 16 August 2025, https://www.awfdn.org/army-women-in-history/.

55 Liliana Brožič and Mojca Pešec, "Ženske v oboroženih silah – primer Slovenske vojske," *Teorija in praksa* 54, No. 1 (2017): 123–25. Pavel Vuk and Ela Tonin Mali, "Pripadnice Slovenske vojske na poveljniških dolžnostih na mednarodnih operacijah in misijah," *Teorija in praksa* 57, No. 3 (2020): 736–38.

promoted to the rank of brigadier (2011), followed by the rank of major general (2018). In the same year, she was appointed Deputy and subsequently Chief of the General Staff of the Slovenian Armed Forces.[56]

Contemporary Warrioresses

Outside the regular militaries, in the 21st century, women participated in non-state (para)military organisations. One such organisation is the Women's Protection Units (YPJ), an all-female militia that is the counterpart of the Kurdish-majority People's Defence Units (YPG), involved in the Syrian Civil War. The first mixed-gender units within the YPG that fought against the Syrian regime and Islamist rebel groups were formed around 2011, while the YPJ was officially established in April 2013. Although primarily composed of Kurdish women, women of other ethnicities also joined their ranks and fought alongside the YPG in many operations and battles. The YPJ established Women's Military Academies in each of the three cantons in the Rojava region, which provided comprehensive training for women.[57]

The latest armed conflict in Europe, the Russian invasion of Ukraine (2014–), once again shows that wars are the natural outcome of women's evolving roles in the military, as personnel requirements demand greater inclusion of the population. In 2014, nearly 50,000 women served in the Armed Forces of Ukraine, with 16,500 of them directly in military units. Most were medical and communications specialists, accountants, clerks, and cooks. However, after the Russian invasion, their numbers grew by 40%, and by January 2024, over 62,000 women made up about 7.3% of all personnel. Currently, 45,500 women serve in military units, with more than four thousand deployed on the frontlines. This increase in the ranks was also a result of the 2015 mobilisation, which included women aged between 20 and 50. Next year, women will be permitted to take on certain combat roles. In 2018, a law was enacted allowing women to participate in all military positions, including combat roles. Presently, women serve as drivers, tank and armoured vehicle crews, reconnaissance units, machine gunners, snipers, and unmanned aerial vehicle operators, among other roles. In 2021, the first woman, Tetiana Ostashchenko, was promoted to the rank of brigadier general and appointed commander of the Medical Forces of the Armed Forces of Ukraine.[58]

56 Andreja Rakovec, Ermenc, Alenka, *Slovenska biografija* (Ljubljana: ZRC SAZU, 2013), accessed on 14 August 2025, http://www.slovenska-biografija.si/oseba/sbi1024520/#novi-slovenski-biografski-leksikon.

57 Valetina Dean, "Kurdish Female Fighters: The Western Depiction of YPJ Combatants in Rojava," *Globalism*, No. 1 (2019): 11, https://doi.org/10.12893/gjcpi.2019.1.7.

58 Daryna Hrysiuk, *How Many Women are Defending Ukraine Against Russia's Invasion?*, 26 March 2024, accessed on 14 August 2025, https://war.ukraine.ua/articles/how-many-women-are-defending-ukraine-against-russia-s-invasion. Jessica Trisko Darden, "Ukrainian Wartime Policy and the Construction of Women's Combatant Status," *Women's Studies International Forum* 96 (2023), https://doi.org/10.1016/j.wsif.2022.102665.

Fewer women served or are serving on the Russian side. While the Russian Armed Forces permit women to serve, they are excluded from "certain military positions, considered harmful to their reproductive abilities". According to the Russian Ministry of Defence, in March 2023, approximately 1,100 women participated in combat operations against Ukraine, making up less than half a percent of all Russian military personnel. Later that year, the Russian Ministry of Defence increased the recruitment of women, including women in prisons. Such recruitment efforts were also undertaken by Russian private military companies, which seek women for combat roles.[59]

Conclusion

The historical development of women's roles in the military shows both continuity and change. From the often-invisible camp followers who provided essential logistical and emotional support to armies in Antiquity, through the gradual professionalisation of medical and auxiliary services in modern times, to the visible leadership of women in today's armed forces, the story of women in the military illustrates a progressive evolution of the military organisation as an institution influenced by social, political, and operational needs. The periods of intense armed conflict, especially the two world wars, repeatedly demonstrated that women are not only capable but vital in combat support and, at times, also in frontline roles. However, the demobilisations following these conflicts reveal how inclusion was often viewed as a temporary measure, emphasising the ongoing gender bias within military organisations.

The adoption of UNSCR 1325 represented a turning point, embedding women's participation in security as a normative principle and catalysing reforms in recruitment, training, and promotion. Its influence reinforced the notion that women are not just passive beneficiaries of protection but active contributors to effectiveness, legitimacy, and leadership within armed forces. The more recent experiences of Iraq, Afghanistan, and particularly Ukraine since 2014 demonstrate how contemporary conflicts have once again accelerated women's integration, both in combat and command structures.

Ultimately, the long evolution "from camp followers to leaders" affirms that women's military contributions are not mistakes but have been essential to the history and future of warfare. Recognising and institutionalising these roles is not just a matter of equality but of operational necessity. The challenge ahead is not in proving women's capability – which history has repeatedly confirmed – but in ensuring that the structures of promotion, command, and culture genuinely reflect that reality.

59 Egle E. Murauskaite, *Russian Women in the Face of War Against Ukraine, Foreign Policy Research Institute*, 26 March 2024, accessed on 14 August 2025, https://www.fpri.org/article/2024/03/russian-women-in-the-face-of-war-against-ukraine.

Sources and Literature

Literature

Ailes, Mary Elizabeth. "Camp Followers, Sutlers, and Soldiers' Wives: Women in Early Modern Armies (c. 1450– 1650)." In *A Companion to Women's Military History*, edited by Barton C. Hacker and Margaret Vining, 61–71. Leiden and Boston: Brill, 2012.

Alfonso, Kristal L. M. *Femme Fatale: An Examination of the Role of Women in Combat and the Policy Implications for Future American Military Operations*. Maxwell AFB: Air University Press, 2009.

Ali, Fatima Jasim Mohammed. "The Russo-Turkish War in Crimea and Nightingale's Role in It (1854–1855)." *World Journal of Advanced Research and Reviews* 18, No. 2 (2023). https://doi.org/10.30574/wjarr.2023.18.2.0746.

Allison, Penelope M. "Mapping for Gender: Interpreting Artefact Distribution inside 1st- and 2nd-Century A.D. Forts in Roman Germany." *Archaeological Dialogues* 13, No. 1 (2006): 1–20. https://doi.org/10.10.17/S1380203806211851.

Anderson, Morgan K., et al. "Effect of Mandatory Unit and Individual Physical Training on Fitness in Military Men and Women." *American Journal of Health Promotion* 30, No. 2 (2016): 100–10. https://doi.org/10.1177/0890117116666977.

Baldwin, J. Norman. "Female Promotions in Male-Dominant Organizations: The Case of the United States Military." *Journal of Politics* 58, No. 4 (1996): 1184–97.

Bastick, Megan, and Claire Duncanson. "Agents of Change? Gender Advisors in NATO Militaries." *International Peacekeeping* 25, No. 3 (2018): 345–68. https://doi.org/10.1080/13533312.2018.1492876.

Baumann, Julia, Charlotte Williamson, and Dominic Murphy. "Exploring the Impact of Gender-Specific Challenges during and after Military Service on Female UK Veterans." *Journal of Military, Veteran and Family Health* 8, No. 2 (2022): 72–81. https://doi.org/10.3138/jmvfh-2021-0065.

Belfiglio, Valentine J. "Women and the Ancient Roman Army." *Journal of Clinical Research and Case Studies* 1, No. 1 (2023): 2–5.

Ben-Shalom, Uzi, Eyal Lewin, and Shimrit Engel. "Organizational Processes and Gender Integration in Operational Military Units: An Israel Defense Forces Case Study." *Gender, Work & Organization* 26, No. 9 (2019): 1289–1303. https://doi.org/10.1111/gwao.12348.

Bergman, Beveryl P., and Simon A. St J. Miller. "Equal Opportunities, Equal Risks? Overuse Injuries in Female Military Recruits." *Journal of Public Health Medicine* 23, No. 1 (2001): 35–39.

Bernik, Valerija. "Veteranke druge svetovne vojne." *Sodobni vojaški izzivi* 19, No. 2 (2017): 71–87. https://doi.org/10.33179/BSV.99.SVI.11.CMC.19.2.5.

Bernik, Valerija. "Ženske v slovenski partizanski vojski (1941–1945)." In *Seksizem v vojaški uniformi*, edited by Ljubica Jelušič and Mojca Pešec, 106–26. Ljubljana: Obramboslovni raziskovalni center, Generalštab Slovenske vojske, 2002.

Bolzenius, Sandra. "Asserting Citizenship: Black Women in the Women's Army Corps (WAC)." *International Journal of Military History and Historiography* 39 (2019): 208–31.

Booth, Bradford, and David R. Segal. "Bringing the Soldiers Back In Implications of Inclusion of Military Personnel Market Research on Race, Class, and Gender." *Race, Gender & Class* 12, No. 1 (2005): 34–57.

Booth, Bradford, William W. Falk, David R. Segal, and Mady Wechsler Segal. "The Impact of Military Presence in Local Labor Markets on the Employment of Women." *Gender & Society* 14, No. 2 (2000): 318–32.

Brown, Melissa T. "'A Woman in the Army Is Still a Woman': Representations of Women in US Military Recruiting Advertisements for the All-Volunteer Force." *Journal of Women, Politics & Policy* 33, No. 2 (2012): 151–75.

Brožič, Liliana, and Mojca Pešec. "Ženske v oboroženih silah – primer Slovenske vojske." *Teorija in praksa* 54, No. 1 (2017): 112–28.

Brunet, Stephan. "Women with swords: Female gladiators in the Roman world." In *A Companion to Sport and Spectacle in Greek and Roman Antiquity,* edited by Paul Cristesen and Donald G. Kyle, 478–91. Chichester: Wiley Blackwell, 2016.

Bruscino, Thomas A. "Palm Sunday Ambush, 20 March 2005." In *In Contact! Case Studies from the Long War: Volume I*, edited by William G. Robertson, 59–82. Fort Leavenworth, KS: Combat Studies Institute Press, 2003.

Caragea, Marius-Emanuel. "Modern Challenges to Military Management: UN Security Council Resolution 1325 'Women, Peace and Security.'" *Management & Marketing* 21, No. 2 (2023): 312–27.

Caroll, Luke. "Raising a Female-Centric Infantry Battalion: Do We Have the Nerve?." *Australian Army Journal* 11, No. 1 (2014): 40–55.

Carson Stanley, Sandra, and Mady Wechsler Segal. "Military Women in NATO: An Update," *Armed Forces and Society* 14, No. 4 (1988): 559–85.

Castillo Diaz, Pablo. "Military Women in Peacekeeping Missions and the Politics of UN Security Council Resolution 1325", *Sodobni vojaški izzivi* 18, No. 3 (2016): 23–34. https://doi.org/10.33179/BSV.99.SVI.11.CMC.18.3.3.

Cenati, Chiara, and Peter Kruschwitz. "Poetic Baggage: Representations of Camp Followers in the Latin Verse Inscriptions." *Electrum* 31 (2024): 153–83. https://doi.org/10.4467/20800909EL.24.012.19162.

Chandler, Joan, Lyn Bryant, and Tracey Bunyard. "Women in Military Occupations." *Work, Employment & Society* 9, No. 1 (1995): 123–35.

Connell, Cati. *A Few Good Gays: The Gendered Compromises behind Military Inclusion*. Oakland: University of California Press, 2023.

Crang, Jeremy A. *Sisters in Arms: Women in the British Armed Forces during the Second World War*. Cambridge: Cambridge University Press, 2020.

de Groot, Gerard J. "Combatants or Non-Combatants? Women in Mixed Anti-Aircraft Batteries during the Second World War." *RUSI Journal* 140, No. 6 (1995): 65–70.

Dean, Valetina. "Kurdish Female Fighters: The Western Depiction of YPJ Combatants in Rojava." *Globalism*, No. 1 (2019). https://doi.org/10.12893/gjcpi.2019.1.7.

Derbyshire, Jane. "An Analysis and Critique of the UNSCR 1325 Resolution – What are Recommendations for Future Opportunities?." *Sodobni vojaški izzivi* 18, No. 3 (2016): 83–93. https://doi.org/10.33179/BSV.99.SVI.11.CMC.18.3.7.

Dittmar, Sharon S., et al. "Images and Sensations of War: A Common History of Military Nursing." *Health Care for Women International* 17, No. 1 (1996): 69–80. https://doi.org/10.1080/07399339609516221.

Egnell, Robert, Petter Hojem, and Hannes Berts. *Gender, Military Effectiveness, and Organizational Change: The Swedish Model*. Houndsmills, New York: Palgrave Macmillan, 2014.

Fell, Alison S. *Women as Veterans in Britain and France after the First World War*. Cambridge: Cambridge University Press, 2018.

Fieldhouse, Anne, and T. J. O'Leary. "Integrating Women into Combat Roles: Comparing the UK Armed Forces and Israeli Defense Forces to Understand where Lessons can be Learnt." *BMJ Military Health* 169, No. 1 (2023): 78–80. https://doi.org/10.1136/bmjmilitary-2020-001500.

Fieseler, Beate, M. Michaela Hampf, and Jutta Schwarzkopf. "Gendering Combat: Military Women's Status in Britain, the United States, and the Soviet Union during the Second World War." *Women's Studies International Forum* 47 (2014): 115–26. https://doi.org/10.1016/j.wsif.2014.06.011.

Fraioli, Deborah A. *Joan of Arc and the Hundred Years War*. Westport Connecticut, London: Greenwood Press, 2005.

Friðriksdóttir, Jóhanna Katrín. *Valkyrie: The Women of the Viking World*. London, New York: Bloomsbury Academic, 2020.

Fry, Helen. *Women in Intelligence: The Hidden History of Two World Wars*. New Haven: Yale University Press, 2023.

Fulan Štante, Nadja. "Strenghts and Weaknesses of Women's Religious Peace-Building (in Slovenia)." *Annales* 30, No. 3 (2020): 343–54. https://doi.org/10.19233/ASHS.2020.21.

Furlan Štante, Nadja. "Ženske v oboroženih silah: Med nasiljem in ranljivostjo." *Sodobni vojaški izzivi* 18, No. 3 (2016): 95–105. https://doi.org/10.33179/BSV.99.SVI.11.CMC.18.3.8.

Garb, Maja. "Ženske na služenju vojaškega roka v JLA." In *Seksizem v vojaški uniformi*, edited by Ljubica Jelušič and Mojca Pešec, 127–36. Ljubljana: Obramboslovni raziskovalni center, Generalštab Slovenske vojske, 2002.

Gasca, Frank, Ryan Voneida, and Ken Goedecke. "Unique Capabilities of Women in Special Operations Forces." *Special Operations Journal* 1, No. 2 (2015): 105–11. https://doi.org/10.1080/23296151.2015.1070613.

Gersdorff, Ursula von. *Frauen im Kriegsdienst, 1914–1945*. Stuttgart: Deutsche Verlags-Anstalt, 1969.

Golubović, Vidoje D. *Dobrovoljka Milunka Savić: Srpska heroina*. Beograd: Udruženje ratnih dobrovoljaca 1912–1918, njihovih potomaka i poštovalaca, 2013.

Hacker, Barton C. "Reformers, Nurses, and Ladies in Uniform: The Changing Status of Military Women (c. 1815–c. 1914)." In *A Companion to Women's Military History*, edited by Barton C. Hacker and Margaret Vining, 137–41. Leiden and Boston: Brill, 2012.

Hadley, Dawn M., et al. "The Winter Camp of the Viking Great Army, AD 872–3, Torksey, Lincolnshire." *The Antiquaries Journal* 96 (2016): 23–67. https://doi.org/10.1017/S0003581516000718.

Head, Naomi. "'Women Helping Women': Deploying Gender in US Counterinsurgency Wars in Iraq and Afghanistan." *Security Dialogue* 55, No. 2 (2024): 153–69. https://doi.org/10.1177/09670106231203839.

Hess, Donabelle C. "Military Family Readiness: The Importance of Building Familial Resilience and Increasing Family Well-Being through Military Community Support and Services." *Sodobni vojaški izzivi* 22, No. 2 (2020): 89–99. https://doi.org/10.33179/BSV.99.SVI.11.CMC.22.2.5.

Hirschauer, Sabine. *The Securitization of Rape: Women, War and Sexual Violence*. Houndmills, New York: Palgrave Macmillan, 2014.

Huss, Ephrat, and Julie Cwikel. "Women's Stress in Compulsory Army Service in Israel: A Gendered Perspective." *Work* 50, No. 1 (2015): 37–48. https://doi.org/10.3233/WOR-141930.

Ikuhiko, Hata. *Comfort Women and Sex in the Battle Zone*. Lanham, Boulder, New York, and London: Hamilton Books, 2018.

Jelušič, Ljubica, Julija Jelušič Južnič, and Jelena Juvan. "The Relevance of Military Families for Military Organizations and Military Sociology." *Sodobni vojaški izzivi* 22, No. 2 (2020): 51–67. https://doi.org/10.33179/BSV.99.SVI.11.CMC.22.2.3.

Jensen, Kimberly. "Volunteers, Auxiliaries, and Women's Mobilization: The First World War and Beyond (1914–1939)." *A Companion to Women's Military History*, edited by Barton C. Hacker and Margaret Vining, 189–223. Leiden, Boston: Brill, 2012.

Juvan, Jelena. "Usklajevanje delovnih in družinskih obveznosti v vojaški organizaciji." *Socialno delo* 48, No. 4 (2009): 227–34.

Karazi-Presler, Tair, Orna Sasson-Levy, and Edna Lomsky-Feder. "Gender, Emotions Management, and Power in Organizations: The Case of Israeli Women Junior Military Officers." *Sex Roles* 78, No. 7–8 (2018): 573–86. https://doi.org/10.1007/s11199-017-0810-7.

Kasearu, Kairi, et al. "Military Families in Estonia, Slovenia and Sweden: Similarities and Differences." *Sodobni vojaški izzivi* 22, No. 2 (2020): 69–87. https://doi.org/10.33179/BSV.99.SVI.11.CMC.22.2.4.

Kilián, Jan. "A Soldier and a Townsman during the Thirty Years' War. Coexistence – Confrontation – Cooperation." *Przeglad Zachodniopomorski* 63, No. 4 (2019): 40–60. https://doi.org/10.18276/pz.2019.4-02.

Kolbeck, Ben. "A Foot in Both Camps: The Civilian Suppliers of the Army in Roman Britain." *Theoretical Roman Archaeology Journal* 1, No. 1 (2018): 1–19. https://doi.org/10.16995/traj.355.

Krylova, Anna. *Soviet Women in Combat: A History of Violence on the Eastern Front*. Cambridge and New York: Cambridge University Press, 2010.

Kusmiyati, Nani, and Hady Efendy. "The Leadership of Women in Military on Military Organization." *International Journal of Human Resource Studies* 7, No. 4 (2017): 165–74.

Landsend Monsen, Ingunn Helene. *Female integration in Jordan's Special Forces – an empirical analysis of the project's content and value for Norway and Jordan*. Oslo: Norwegian Defence Research Establishment, 2025.

Lee, Janet. "Sisterhood at the Front: Friendship, Comradeship, and the Feminine Appropriation of Military Heroism among World War I First Aid Nursing Yeomanry (FANY)." *Women's Studies International Forum* 31, No. 1 (2008): 16–29.

Leonheart, Jamie. "Gender Perspectives for Operational Effectiveness: An Opportunity for U.S.Forces Japan and the Japan Self Defense Forces." *NIDS Commentary*, No. 333 (2024).

Lynn II, John A. "Essential Women, Necessary Wives, and Exemplary Soldiers: The Military Reality and Cultural Representation of Women's Military Participation (1600–1815)." In *A Companion to Women's Military History*, edited by Barton C. Hacker and Margaret Vining, 93–136. Leiden and Boston: Brill, 2012.

Magley, Vicki J., et al. "The Impact of Sexual Harassment on Military Personnel: Is It the Same for Men and Women?" *Military Psychology* 11, No. 3 (1999): 283–302. https://doi.org/10.1207/s15327876mp1103_5.

Markwick, Roger D., and Euridice Charon Cardona. *Soviet Women on the Frontline in the Second World War*. Houndmills and New York: Palgrave Macmillan, 2012.

Martsenyuk, Tamara. "Women's Participation in Defending Ukraine in Russia's War." *Kyiv-Mohyla Law and Politics Journal* 8 (2022): 43–59.

Morden, Bettie J. *The Women's Army Corps, 1945–1978*. Washington, DC: Center of Military History, 1990.

Nowaki, Rochelle. "Nachthexen: Soviet Female Pilots in WWII." *Hohonu* 13 (2015): 56–62.

Poklukar, Karmen, and Pavel Vuk. "Vključevanje žensk v specialne sile." *Sodobni vojaški izzivi* 22, No. 4 (2020): 85–105. https://doi.org/10.33179/BSV.99.SVI.11.CMC.22.4.5.

Price, Neil, et al. "Viking Warrior Women? Reassessing Birk Chamber Grave Bj.581." *Antiquity* 93, No. 367 (2019): 192–94. https://doi.org/10.15184/aqy.2018.258.

Pryor, John B. "The Psychological Impact of Sexual Harassment on Women in the U.S. Military." *Basic and Applied Social Psychology* 17, No. 4 (1995): 581–603. https://doi.org/10.1207/s15324834basp1704_9.

Raffield, Ben, Neil Price, and Mark Collard. "Polygyny, Concubinage, and the Social Lives of Women in Viking-Age Scandinavia." *Viking and Medieval Scandinavia* 13 (2017): 189–209. https://doi.org/10.1484/J.VMS.5.114355.

Roche, Rosellen, et al. "The Unseen Patriot: Female Cultural Support Team Members and Combat Definition." *Journal of Veterans Studies* 7, No. 1 (2021): 271–79. https://doi.org/10.21061/jvs.v7i1.285.

Ropers, Erik. "Representation of Gendered Violence in Manga: The Case of Enforced Military Prostitution." *Japanese Studies* 31, No. 2 (2011): 249–66.

Schrivener, Laurie. "U.S. Military Women in World War II: The SPAR, WAC, WAVEs, WASP, and Women Marines in U.S. Government Publications." *Journal of Governmental Information* 26, No. 4 (1999): 361–83.

Skjelsbaek, Inger. "Sexual Violence and War: Mapping Out a Complex Relationship." *European Journal of International Relations* 7, No. 2 (2001): 211–37.

Skylitzes, John. *A Synopsis of Byzantine History, 811–1057*. Cambridge: Cambridge University Press, 2010.

Sorokina, T. S. "Russian Nursing in the Crimean War." *Journal of the Royal College of Physicians of London* 29, No. 1 (1995): 57–63.

Street, Amy E., Dawne Vogt, and Lissa Dutra. "A New Generation of Women Veterans: Stressors Faced by Women Deployed to Iraq and Afghanistan." *Clinical Psychology Review* 29 (2009): 685–94. https://doi.org/10.1016/j.cpr.2009.08.007.

Strle, Urška. "K razumevanju ženskega dela v veliki vojni." *Prispevki za novejšo zgodovino* 55, No. 2 (2015): 103–25.

Strobl, Ingrid. *Partisanas: Women in the Armed Resistance to Fascism and German Occupation (1936–1945)*. Edinburgh and Oakland, CA: AK Press, 2008.

Šaranović, Jovanka, Brankica Potkonjak-Lukić, and Tatjana Višacki. "Achievements and Perspectives of the Implementation of UNSCR 1325 in the Ministry of Defence and the Serbian Armed Forces." *Sodobni vojaški izzivi* 18, No. 3 (2016): 65–81. https://doi.org/10.33179/BSV.99.SVI.11.CMC.18.3.6.

Tkavc, Suzana. "Some of the Best Practices in Gender Perspective and the Implementation of UNSCR 1325 in the 25 Years of Slovenian Armed Forces." *Sodobni vojaški* izzivi 18, No. 3 (2016): 45–63. https://doi.org/10.33179/BSV.99.SVI.11.CMC.18.3.5.

Trisko Darden, Jessica. "Ukrainian Wartime Policy and the Construction of Women's Combatant Status." *Women's Studies International Forum* 96 (2023). https://doi.org/10.1016/j.wsif.2022.102665.

Tunc, Ugurgul. "Lessons from the Crimean War: How Hospitals Were Transformed by Florence Nightingale and Others." *Infectious Diseases & Clinical Microbiology* 1, No. 2 (2019): 110–18. https://doi.org/10.36519/idcm.2019.19020.

Valetina, Dean, "Kurdish Female Fighters: The Western Depiction of YPJ Combatants in Rojava." *Globalism*, No. 1 (2019): 1–29. https://doi.org/10.12893/gjcpi.2019.1.7.

van den Berk Clark, Carissa, Jennifer Chang, Jessica Servery, and Jeffrey D. Quinlan. "Women's Health and the Military." *Primary Care: Clinics in Office Practice* 45, No. 4 (2018): 677–86. https://doi.org/10.1016/j.pop.2018.07.006.

Vining, Margaret. "Women Join the Armed Forces: The Transformation of Women's Military Work in World War II and After (1939–1947)." In *A Companion to Women's Military History*, edited by Barton C. Hacker and Margaret Vining, 233–89. Leiden, Boston: Brill, 2012.

Vining, Margaret, and Barton C. Hacker. "From Camp Follower to Lady in Uniform: Women, Social Class and Military Institutions before 1920." *Contemporary European History* 10, No. 3 (2001): 353–73. https://doi.org/10.1017/S0960777301003022.

Vuga Beršnak, Janja, and Bojana Lobe. "Socioecological Model of a Military Family's Health and Well-being: Inside a Slovenian Military Family." *Armed Forces and Society* 50, No. 1 (2024): 224–52. https://doi.org/10.1177/0095327X221115679.

Vuk, Pavel. "The Slovenian Armed Forces Faces the Challenge of Inclusion of Their Homosexual Members." *Journal of Homosexuality* 71, No. 5 (2024): 1231–52. https://doi.org/10.1080/00918369.2023.2169088.

Vuk, Pavel, and Ela Tonin Mali. "Pripadnice Slovenske vojske na poveljniških dolžnostih na mednarodnih operacijah in misijah." *Teorija in praksa* 57, No. 3 (2020): 731–50.

Vuk, Pavel, and Saša Galičič. "Socialna diverziteta v luči inkluzivnosti istospolno usmerjenih pripadnic in pripadnikov v slovenski vojski." *Teorija in praksa* 59, No. 2 (2022): 568–88.

Webster, Graham. *Boudica: The British Revolt against Rome AD 60*. London: B. T. Batsford, 1993.

Wechler Segal, Mady, et al. "The Role of Leadership and Peer Behaviors in the Performance and Well-Being of Women in Combat: Historical Perspectives, Unit Integration, and Family Issues." *Military Medicine* 181 (2016): 1–28. https://doi.org/10.7205/MILMED-D-15-00342.

Wintjes, Jorit. "'Keep the Women Out of the Camp!' Women and Military Institutions in the Classical World." In *A Companion to Women's Military History*, edited by Barton C. Hacker and Margaret Vining, 21–30. Leiden and Boston: Brill, 2012.

With Honor and Integrity: Transgender Troops in Their Own Words, edited by Mael Embser-Herbert and Bree Fram. New York: New York University Press, 2021.

Online sources

Army Women's Foundation. "Army Women in History." Accessed August 16, 2025. https://www.awfdn.org/army-women-in-history/.

Hawkins, Vikki. "A Princess At War: Queen Elizabeth II During World War II." The National WWII Museum, 2021. Accessed September 5, 2025. https://www.nationalww2museum.org/war/articles/queen-elizabeth-ii-during-world-war-ii.

Hrysiuk, Daryna. "How Many Women Are Defending Ukraine against Russia's Invasion?" March 26, 2024. Accessed August 14, 2025.https://war.ukraine.ua/articles/how-many-women-are-defending-ukraine-against-russia-s-invasion/.

Murauskaite, Egle E. "Russian Women in the Face of War against Ukraine." Foreign Policy Research Institute, March 26, 2024. Accessed August 14, 2025. https://www.fpri.org/article/2024/03/russian-women-in-the-face-of-war-against-ukraine.

Rakovec, Andreja. "Ermenc, Alenka (1963–)." *Slovenska biografija.* Ljubljana: ZRC SAZU, 2013. Accessed August 14, 2025. https://www.slovenska-biografija.si/oseba/sbi1024520/#novi-slovenski-biografski-leksikon.

Strand, Sanna. The 'Scandinavian model' of military conscription: A formula for democratic defence forces in 21st century Europe? Vienna: Austrian Institute for International Affairs, 2021, Accessed August 21. 2025. https://www.oiip.ac.at/cms/media/policy-analysis-scandinavian-model-of-military-conscription.pdf.

Klemen Kocjančič

OD SPREMLJEVALK TABOROV DO VODITELJIC: ZGODOVINSKI RAZVOJ VLOGE ŽENSK V OBOROŽENIH SILAH

POVZETEK

Članek obravnava dolgotrajen in zapleten razvoj vloge žensk v oboroženih silah. Od antike do zgodnjega novega veka so ženske nastopale predvsem kot spremljevalke taborov, kjer so opravljale logistične, negovalne in moralne naloge, večinoma brez formalnega priznanja. V novem veku so se njihove naloge začele formalizirati, zlasti na področju zdravstva in oskrbe, kar ponazarjajo pionirke, kot je Florence Nightingale. Prelom sta pomenili prva in druga svetovna vojna, ko so ženske vstopile v vojaško industrijo, pomožne enote, obveščevalne službe in celo sodelovale v boju, s čimer so dokazale svojo usposobljenost in nujnost v vojaški službi. Čeprav je sledila povojna demobilizacija, so kasnejše družbene spremembe prinesle postopno integracijo, okrepljeno z resolucijo VS ZN 1325, ki je ženske prepoznala kot aktivne udeleženke varnosti. Sodobni oboroženi konflikti, zlasti trenutni v Ukrajini, kažejo na njihov pomen v vlogi zdravstvenega osebja, ostrostrelk, poveljnic in generalk. Razvoj dokazuje, da ženske niso obroben, temveč bistveni del preteklosti, sedanjosti in prihodnosti vojaških sil.

1.01 DOI: https://doi.org/10.51663/pnz.65.3.14

Beti Žerovc*

Cultural and Historical Overview of the Life of the Painter Heinrich Wettach (1858–1929), II. The Artist's Engagement in Ljubljana Social Life and Societies and His Final Years in Carinthia

IZVLEČEK

KULTURNOZGODOVINSKI ORIS ŽIVLJENJA SLIKARJA HEINRICHA WETTACHA (1858–1929), II. SLIKARJEVA DRUŽBENA IN DRUŠTVENA VPETOST V LJUBLJANI TER NJEGOVA ZADNJA LETA NA KOROŠKEM

Članek obravnava družbeno vpetost slikarja Heinricha Wettacha (1858–1929), ki je deloval v Ljubljani od leta 1885 do konca prve svetovne vojne, ter ga s tem konkretneje usidra tudi v kulturni milje kranjske prestolnice na prelomu stoletja. Pretresa njegovo povezanost z društvi in organizacijami, s katerimi je bil najmočneje povezan, in opredeljuje naravo njegovega sodelovanja z njimi. Zadnji del članka se posveča neprostovoljni odselitvi slikarja in njegove družine iz Ljubljane ter njegovim zadnjim letom na Koroškem.

Ključne besede: Heinrich Wettach, Ljubljana okoli 1900, nemško čuteči kranjski kulturni krog, slovensko slikarstvo 19. stoletja, avstrijsko slikarstvo 19. stoletja, Filharmonično društvo, Društvo Kazina

* **PhD, Associate Professor, University of Ljubljana, Faculty of Arts, Department of Art History, Aškerčeva 2, SI-1000 Ljubljana, beti.zerovc@ff.uni-lj.si**

ABSTRACT

The article elaborates on the extensive social commitments of the painter Heinrich Wettach (1858–1929), who lived and worked in Ljubljana from 1885 until the end of World War I, to position him firmly within the cultural context of the Carniolan capital at the turn of the century. It explores his connections with societies and organisations he was dedicated to and defines the nature of his collaboration with them. The last part of the article explores the involuntary move of the painter and his family from Ljubljana and his final years in Carinthia.

Keywords: Heinrich Wettach, Ljubljana around 1900, German-leaning cultural circle of Carniola, Slovenian 19th-century painting, Austrian 19th-century painting, Philharmonic Society, Kazina Society

Engagement in Social Life and Societies

In the 19th century, individuals were largely engaged in public life through activities provided by different associations and societies. Heinrich Wettach, for example, belonged to a relatively large social structure which we can reconstruct, with a reasonable level of certainty, using newspaper sources. The connected article in the previous issue of this journal mentions his involvement with the Philharmonic Society, which was quite possibly among the key factors that influenced his decision to remain in Ljubljana. The activities of this venerable Ljubljana music society and Wettach's appearances are recorded in the society's annual reports that were published between 1863 and 1918. Wettach immortalised their long-time editor, the Philharmonic Society's inexhaustible director and physician Friedrich Keesbacher (1831–1901) with a three-quarter length portrait in 1892.[1] The portrait now hangs in the office of the director of the Slovenian Philharmonic, while the main hall is still adorned by Wettach's personifications of the four symphonic movements.

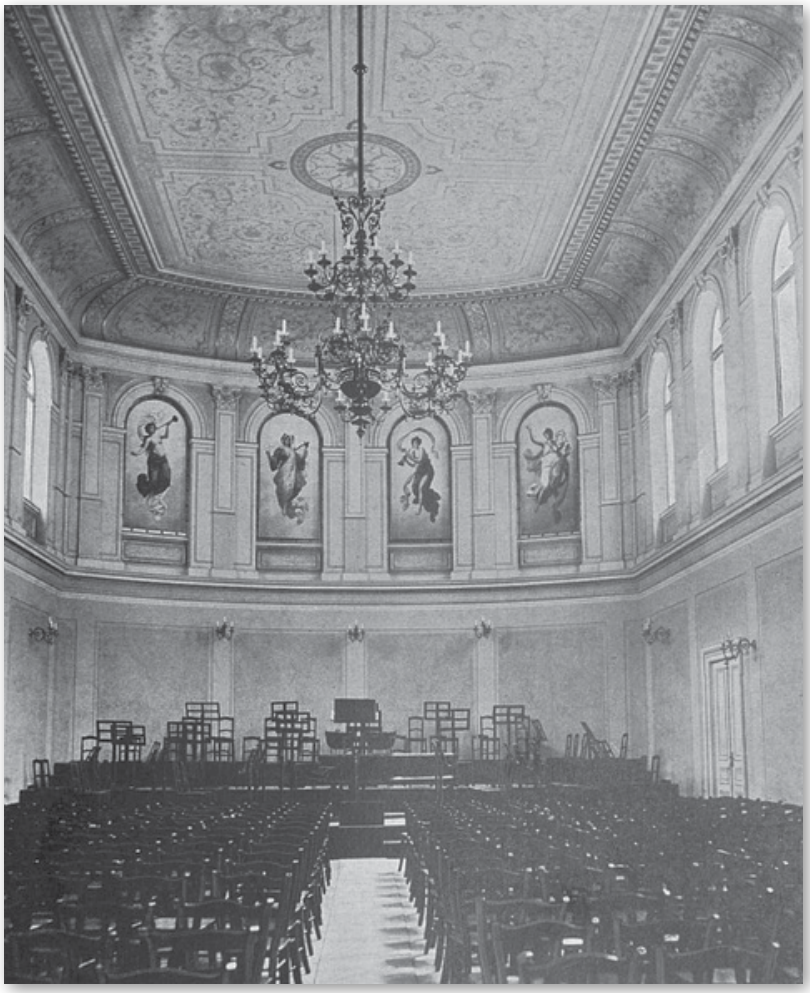
Sources portray Wettach as an enthusiastic musician who, since his arrival in Ljubljana in 1885, sang in the male choir of the Philharmonic Society.[2] In 1887, he was already a violinist in its orchestra and soon became a regular member of Gerstner's string quartet, in which he played both the violin and viola.[3] The quartet was very well known and popular in Ljubljana, news and reviews of their work can be regularly found in *Laibacher Zeitung*, sometimes even several times per season. Wettach also appears as a pianist, accompanying female vocal performances.

1 "Philharmonische Gesellschaft," *Laibacher Zeitung*, CXI/8, 12 January 1892, 67.

2 Friedrich Keesbacher, ed., *Jahres-Bericht der philharmonischen Gesellscahft in Laibach: für die Zeit vom 1. Oktober 1885 bis 30. September 1886* (Laibach: Verlag der philharmonischen Gesellschaft, 1886), 35.

3 Beti Žerovc, "Cultural and Historical Overview of the Life of the Painter Heinrich Wettach (1858–1929), I. The Painter's Beginnings and Settling in Ljubljana," *Prispevki za novejšo zgodovino* 65, No. 2 (2025): 98–117. See fn 21 and 22.

Figure 1: Heinrich Wettach, Paintings with the personifications of four symphonic movements in the original Great Hall of the Ljubljana Philharmonic, photo



Source: Emil Bock, *Die philharmonische Gesellschaft in Laibach*, Laibach 1902, n. pag.

The Philharmonic society was, at the time, by far the finest Ljubljana cultural institution, and its orchestra was, although not professional, recognised for an enviably high level of quality at the turn of the century. Considering all this, Wettach must have been an excellent musician.[4] In the 1890s, he started taking over tasks in the Philharmonic Society management.

In 1894, he was the custodian for instruments and from 1902 onwards, an

4 He also performed and participated in organising the celebrations of the Philharmonic Society's bicentennial in the Anniversary season 1901/1902, which was one of the most resonant music events of the season in Austria. As a musician in one of the most venerable music societies in Central Europe, he performed and met with very important people in the Austrian music world. – Primož Kuret, "Jubilejna koncertna sezona 1901–1902 Ljubljanske filharmonične družbe," *Muzikološki zbornik* 19 (1983), 41–50.

archivist.[5] The society rewarded him for his loyal collaboration and outstanding services to the institution with an honorary membership in 1919.[6]

From 1896 onwards, Wettach is also listed in membership records of the Ljubljana social club Kazina. In his biography, this coincides with the time when, following his 1895 wedding, he firmly decided to put down his roots in Ljubljana. Kazina was a typical bourgeois social association created in the early 1800s, and since the 1830s onwards provided entertainment for Ljubljana's middle class in the Kazina building on the corner of Congress Square.[7] The society provided several services for its members: reading rooms, smoking rooms, a pool room, a ballroom, a coffee house, a restaurant and other facilities for gatherings and offered entertainment for the Ljubljana elite. They organised dances, parties, recitals, lectures, charity events etc. In the second half of the 19th century, a large number of Liberal Party members also belonged to Kazina, but as decades passed and national tensions increased, Kazina gained a reputation as a declaratively German association and a key social and cultural hub for German-leaning Carniolans.[8] As a member Wettach participated in different constellations. It is clear from his undertaking time-consuming creations of decorations for venues, sets and costume designs and even tableaux vivants for numerous social and charity events that he was wholeheartedly committed to the association's activities.[9]

At the beginning of the 20th century, Kazina also developed its own programme for visual arts and in a series of exhibitions invited various renowned Central European art associations to exhibit there. Under the leadership of Ottomar Bamberg, they set up a special six-member committee to prepare the exhibitions and Heinrich Wettach was one of its members.[10]

5 Sara Železnik, *Repertoarne smernice Filharmonične družbe v Ljubljani: Katalogi muzikalij Filharmonične družbe* (Ljubljana: Znanstvena založba Filozofske fakultete, 2014), 34–37.

6 Primož Kuret, *Ljubljanska filharmonična družba 1794–1919: Kronika ljubljanskega glasbenega življenja v stoletju meščanov in revolucij* (Ljubljana: Nova revija, 2005), 445. Wettach worked devotedly for the Philharmonic Society even during World War I. For this reason, the article in the *Laibacher Zeitung* described him, together with Hans Gerstner and Viktor Ranth, as "one of the three founding pillars on whom the pursuit of music rests in the Philharmonic Society in the time of war." – "Das fünfte Gesellschaftskonzert der Philharmonischen Gesellschaft," *Laibacher Zeitung* CXXXVI/79, 6 April 1917, 515.

7 Miha Valant, "Ljubljansko društvo Kazina in združenja za likovno umetnost na Kranjskem med leti 1848 in 1918" (doctoral dissertation, Ljubljana: Faculty of Arts, 2023), 1–30. The overview of the Kazina Society membership reveals that in Wettach's time, many Carniolans at the peak of their careers in the fields of finance, industry, commerce, politics, as well as arts belonged to it, for example musicians Hans Gerstner and Josef Zöhrer, printer Ottomar Bamberg, merchant Josef Luckmann, members of the families Kosler, Galle etc. Members included the managers of the Carniola Savings Bank. See fn 47.

8 A concise summary of the history of the Kazina Society from its establishment to its demise after World War II was presented by Marko Zajc, "Kazina skozi čas," in Aleš Gabrič, ed., *Zgodovinopisje v zrcalu zgodovine: 50 let inštituta za novejšo zgodovino* (Ljubljana: Inštitut za novejšo zgodovino, 2009), 127–38. On the division of Carniolans into Slovenian-leaning and German-leaning in the field of culture and the role of Kazina and other visual arts associations in the growing conflict at the end of the 19th century, see Valant, "Ljubljansko društvo Kazina", 189–91 et passim.

9 "Wohltätigkeist-Concert," *Laibacher Zeitung* CXV/90, 20 April 1896, 733. "Wohltätigkeist-Vorsetellungen," Laibacher Zeitung CXV/95, 25 April 1896, 776. "Chrysantemen-Fest," *Laibacher Zeitung* CXX/256, 7 November 1901, 2119. "Alpines Fest," *Laibacher Zeitung* CXXII/16, 21 January 1903, 127. "Ein Rendezvous in der Unterwelt," *Laibacher Zeitung* CXXV/46, 26 February 1906, 403. *The Feast of Chrysanthemums and the Alpine Feast* were Ljubljana Schulverein events, which, like many other associations, organised larger events in the Kazina Palace.

10 Miha Valant and Beti Žerovc, "Društva za likovno umetnost na Kranjskem v obdobju od 1848 do 1918," *Likovne besede*, No. 113 (2019): 10.

Figure 2: Heinrich Wettach, Flight to Egypt and Saint Elizabeth of Hungary, 1910, oil on canvas, 560 x 220 cm. Originally at the old Ljubljana Hospice in Zaloška Street, now at the parish church of the Stična Abbey.

Source: Branko Petauer

Alongside foreign and other Austrian authors, the local Carniolan artists or organisations participated in these exhibitions, among them Hans Klein, Elsa Kastl, Frida Weiß, the Carniolan Association for Art Weaving and of course Wettach, who exhibited landscapes, particularly mountain motifs.[11]

Many painting and family trips to the Carniolan and Carinthian Alps testify to his love for the mountains, trips where they were occasionally joined by other families, for example, the family of the already mentioned musician Hans Gerstner.[12] If Wettach was not a member of the German-Austrian Mountaineering Association, he at least moved in its circles. The Association contacted him twice with commissions, and in 1892, he participated in their exhibition in the Philharmonic Society building.[13]

All this indicates that the societies and activities that we mention were often connected, and their members merged and mixed.[14] If, within the activities and events at the Kazina or the Philharmonic Society, Wettach collaborated with individuals who differed vastly in their political orientation, particularly within the spectrum of the Austrian centralist and German-leaning Carniolans, he also joined some of Ljubljana's more markedly German-oriented societies.[15] Ever since his arrival in Ljubljana, he collaborated, first as a choirmaster and later also as an official, with the German gymnastics society Turnverein, which was at odds with various Slovenian associations, for example, the South Sokol gymnastic society.[16]

Besides their enthusiasm for art and culture, the Wettachs both expressed great interest in schooling and education; they were openly partial to associations with uncontestably "German" character, such as *Deutscher Schulverein* and *Südmark*.

11 For more about this and the importance of the Kazina for his painting school, see Beti Žerovc and Miha Valant, "The Artistic Formation of the Painter Heinrich Wettach (1858–1929) and His Educational Work," [forthcoming]. The question that remains open is his potential membership of art associations in Austria at the time. Since the end of the 1870s Carniola did not have one yet newspapers show that Wettach appeared in exhibitions of the Styrian art association from Graz, first in Ljubljana in 1889 and then in Graz in 1896. – "Gemälde-Ausstellung," *Laibacher Wochenblatt*, 444, 9 February 1889, n. pag. "Die Frühjahres-Ausstellung des steiermärkischen Kunstvereienes," *Grazer Tagblatt*, VI/124, 5 May 1896, 2.

12 Jernej Weiss, *Hans Gerstner. (1851–1939): Življenje za glasbo* (Maribor: Litera and Pedagoška fakulteta, 2010), 145, 146.

13 "Die Eröffnung der Triglavhütte ober dem Kotthale am 31. Juli 1887," *Laibacher Zeitung* CVI/174, 3 August 1887, 1444. "Die Section 'Krain' des deutschen und österreichischen Alpenvereines," *Laibacher Wochenblatt*, 451, 30 March 1889, n. pag. "Section 'Krain' des Alpenvereines," *Laibacher Zeitung* CXII/53, 5 March 1892, 438. More about this society in Peter Mikša, "'Da je Triglav ostal v slovenskih rokah, je največ moja zasluga.' Jakob Aljaž in njegovo planinsko delovanje v Triglavskem pogorju." *Zgodovinski časopis* 69, No. 1/2 (2015): 113–16. Marija Mojca Peternel, *"Ljubiteljem kranjskih Alp!": Kranjska podružnica Nemškega in avstrijskega planinskega društva* (Ljubljana: Založba Univerze, 2023).

14 Valant, "Ljubljansko društvo Kazina," 18–25.

15 Even within Kazina, Wettach belonged to the Green Island group *(Grüne Insel)* that was more pro-German. –"Zum Tode Heinrich Wettachs," *Freie Stimmen* XLIX/238, 15 October 1929, 6.

16 "Die Anastasius-Grün-Feier in Laibach," *Deutsche Wacht*, XI/45, 6 June 1886, 3. "Der Familienabend des laibacher deutschen Turnvereines," *Laibacher Wochenblatt* 297, 17 April 1886, n. pag. "Familienabend des laibacher deutschen Turnvereines," *Laibacher Zeitung* CVI/99, 3 May 1887. "Laibacher deutscher Turnverein," *Laibacher Zeitung* CX/157, 14 July 1891, 1309. "Ehrungs-Kneipe," *Laibacher Wochenblatt*, 604, 5 March 1892, n. pag. "Der Laibacher deutsche Turnverein," *Deutsche Stimmen aus Krain Triest und Küstenland, Beilage des Grazer Tagblattes* XI/8, 23 January 1901, 11.

Figure 3: Heinrich Wettach, Landscape, oil on canvas, 55 x 66.5 cm

Source: Private collection

Undoubtedly, this aligned with their general political orientation and engagement, but their interest and commitment perhaps came from the fact that at the beginning of the 20th century, they were parents of four school-age children and found themselves in a position where it was becoming increasingly difficult to secure comprehensive, quality schooling for children and young adults in Carniola in German.[17]

The Wettachs worked with the male and female sections of the *Deutscher Schulverein*, an organisation that we count as one of the so-called German protective societies (*Schützvereine*).[18] It was established in 1880 in Vienna to encourage and support activities of schools and kindergartens in the German language on linguistic margins and in linguistically mixed areas.[19] The jubilee yearbook of the Carniolan *Schulverein's* women's section tells us about Marie Wettach's involvement.[20]

17 Matić, *Nemci v Ljubljani*, 274–86, 373 et passim. Angela Ilić, "Podajmo si roke in srca!," *Stati inu obstati*, No. 23/24 (2016): 199.

18 Wettach became the society's official from 1900 onwards. "Ortsgruppe Laibach des deutschen Schulvereines," *Laibacher Zeitung* CXXIII/82, 30 April 1904, 657. Marie Wettach served on the board of its women's division until 1902. – "Hauptversammlung der Frauenortsgrupe Laibach des Deutschen Schulvereines," *Laibacher Zeitung* CXXI/54, 6 March 1902, 433.

19 More about this in Pieter M. Judson, *Guardians of the Nation: Activists on the Language Frontiers of Imperial Austria* (Cambridge MA: Harvard University Press, 2006), 19–65. Matić, Nemci v Ljubljani, 227–31, 279–82.

20 Vodstvo šole, ed., *Denkschrift zum 25 jährigen Bestande der Frauen-Ortsgruppe »Laibach« des deutschen Schulvereins: 1885–1910* (Laibach: Verlag der Deutsch. Schulvereinsschule in Laibach, 1910), 26, 34.

The text also specifically thanks her husband, Heinrich Wettach, who often participated in decorating the venues for many a *Schulverein's* event.[21]

Figure 4: Marie Wettach in a historical costume, c. 1900, photo.

Source: Private collection

In 1903, the women's section of the *Schulverein* applauded the establishment of the women's division of the *Südmark*, of which Marie Wettach was also a member.[22] *Südmark* was another one of the German protective societies with a more pronounced German nationalist accent. It was established in 1889 in Graz, and its objective was planned migration of poor German-speaking farmers to linguistically mixed rural areas where German was being replaced by Slovenian, for example, in Southern Styria,

21 Ibidem, 4.
22 Ibid., 29. "Frauenortsgruppe Laibach des Vereines 'Südmark'," *Laibacher Zeitung* CXXV/27, 3 February 1905, 230.

Carinthia, Carniola.[23] While the Wettachs were active each in their own section of *Schulverein*, it is not entirely clear if Heinrich Wettach was a member of the *Südmark*'s male division.

After the family converted to Protestantism in 1901, Marie Wettach became involved in the Evangelical Women's Society (Evangelischer Frauenverein Laibach) in Ljubljana. The association was visibly engaged in culture, child-rearing and education. As early as the mid-19th century, a school was established next to the Evangelical Church in Ljubljana, which society members assisted financially and in other ways, and in 1901, the association also founded a kindergarten. Among other things, the association strongly supported the demand for women to have access to study in Austrian universities, faculties of medicine and arts.[24] Members participated in charity work, mostly when it came to working with children, helping the poor and in the event of natural disasters. For this purpose, the association cultivated a strong social visibility, because it often raised funds by organising charity concerts, raffles and other events.[25] Marie Wettach served as the president of the society between 1904 and 1906, and again between 1910 and 1920.[26] As many newspaper articles reveal, the society flourished during her terms, including the time of World War I. In that period, members promoted, among other things, the work of the Red Cross.[27] Marie Wettach was often singled out in the press as a donor of monetary and other contributions for charity purposes.[28]

The Final Years in Carinthia

The dramatic events that took place in the second half of 1918 brought a rapid dissolution of Austria-Hungary. Within only a few months, the majority of Carniola went from Austrian to the new Yugoslav authority. As is clear from the memoirs of Wettach's good friend, the musician Hans Gerstner, individuals who remained partial to the links with the German cultural space and the – at the time already former – imperial Austria, were subjected to considerable pressure in these new circumstances, with many emigrating to the newly created Republic of German-Austria. The new authorities in Carniola banned, among other things, the interests of many institutions and societies that were considered German, such as the Philharmonic Society.[29]

23 Pieter M. Judson, "Versuche um 1900, die Sprachgrenze sichtbar zu machen," in Moritz Csáky and Peter Stachel, eds., *Die Verortung von Gedächtnis*, (Wien: Passagen Verlag, 2001), 171.

24 Aleksandra Serše, "Evangeljsko žensko društvo v Ljubljani 1856–1945," *Etnolog* 11 (2001): 61–64.

25 Ibid., 59–63.

26 Ibid., 63. As the president, she was probably active only until 1919 when the family moved to Carinthia, but remained listed as the president until the new one was elected.

27 Ibid., 64.

28 "5½ avstrijsko vojno posojilo," *Slovenec*, 277, 3 December 1914, 4. "Darila za Rdeči križ," *Slovenski narod*, 39, 18 February 1915, 3. "Spenden für unsere Soldaten im Felde," *Laibacher Zeitung* CXXXV/52, 4 March 1916, 378.

29 Kuret, *Ljubljanska filharmonična*, 443–53. On the problems of the German-leaning population and the "nationalisation" of their institutions after World War I, see Ervin Dolenc, "Deavstrizacija v politiki, upravi in kulturi v Sloveniji," in Dušan Nećak et al., eds., *Slovensko-avstrijski odnosi v 20. stoletju = Slowenischösterreichische Beziehungen*

The property of many such citizens, societies and enterprises in Ljubljana was entrusted to sequestrators after the war.[30]

Heinrich Wettach's family was one of the thirty-five Protestant families who left Ljubljana soon after the war.[31] They moved to their holiday house on Lake Ossiach in Carinthia. *Gottscheer Bote* reported they left in the first third of 1919, when they allegedly sold their two villas situated at a truly elite location in Ljubljana to merchant Grobelnik for 320.000 kroner, which, considering the massive inflation at the time, was a complete bargain.[32] According to their grandson, they were only allowed to take one wagon of possessions with them and even a part of this was lost on the way.[33] Their financial status changed dramatically. Following the meagre profit from the sale of the real estate in Ljubljana and the ban on transferring funds from erstwhile Carniola in the then-new Yugoslav state to the Republic of German-Austria – which were ultimately decimated by inflation – the Wettachs were recipients of social support in Carinthia.[34]

Probably due to financial distress, the Wettachs opened a private educational facility in their house on Lake Ossiach, a home for girls *Heimgard* (*Mädchenheim Heimgard*). The advert in the *Cillier Zeitung* claimed that it was initially intended for girls aged 15 and over who were taught the basic and most important household chores, such as cooking, laundry, sewing, clothes pattern-making, darning and ironing male suits. Heinrich Wettach took over instruction of music, drawing and art history.[35]

im 20. Jahrhundert (Ljubljana: Oddelek za zgodovino Filozofske fakultete, 2004), 81–94, http://hdl.handle.net/11686/26817. Irena Selišnik, "Usode uradnic in uradnikov po prvi svetovni vojni," Retrospektive 7, No. 1 (2024): 11–39. Rok Stergar, "Continuity, Pragmatism, and Ethnolinguistic Nationalism: Public Administration in Slovenia during the Early Years of Yugoslavia," in Peter Becker et al., eds., *Hofratsdämmerung? Verwaltung und Ihr Personal in den Nachfolgestaaten der Habsburgermonarchie 1918 bis 1920* (Wien: Böhlau, 2020), 179–92, https://doi.org/10.7767/9783205211525.179. Irena Selišnik, "Status državljanstva ob nastanku nove Države SHS: Strategije izbire," *Zgodovinski časopis* 164, No. 3–4 (2021): 476–91.

30 In 1920, newspapers reported the names of some of the sequestrators that managed the property of several enterprises and private citizens, including the Wettachs: "Žlahta," *Slovenec* XLVIII/49, 29. 2. 1920, 1. "Žlahta," *Slovenec* XLVIII/56, 9. 3. 1920, 1. About this, see also: Weiss, *Hans Gerstner*, 164–66.

31 Breda Mihelič, "Vilska četrt med Prešernovo cesto in Tivolijem v Ljubljani ter prenova Wettachove vile: Problemi varovanja in prenove stanovanjske četrti," *Varstvo spomenikov* 39 (2003): 142. Cf. Weiss, *Hans Gerstner*, 173.

32 "Besitzwechsel," *Gottscheer Bote* XVI/10, 1 April 1919, 76. Today's value of the 320.000 kroner from 1919 would be approximately 71.330 euros, and the same sum in 1914 would equal 1,99 million euros today. This shows that they sold the house far below its market value, probably in a hurry. The calculations were made using Historischer Währungsrechner: https://finanzbildung.oenb.at/docroot/waehrungsrechner/#, 3 October 2024.

33 Harald Wettach, semi-structured interviews by Beti Žerovc, 2012. Leitenberger remembers: "For me, the difficult times were over, but my parents had to flee Yugoslavia after the war. They received a blue envelope at their address and had 14 days to leave the country. From then on, they could no longer withdraw money from the bank (their account was blocked) or receive aid in food to which they were entitled to. And then they moved. They filled up half a wagon with everything they could, furniture and other belongings. Those were things that they were allowed to take. Where did they go? To Villach. They had acquaintances there who stored these things in their attic. Luckily, they owned a small holiday house near Villach, where they could live. It was furnished, like one furnishes a summer house. Simple, just enough to have a roof over their heads." – Brigitta Leitenberger, a semi-structured interview by Ruth Deutschmann, 1998.

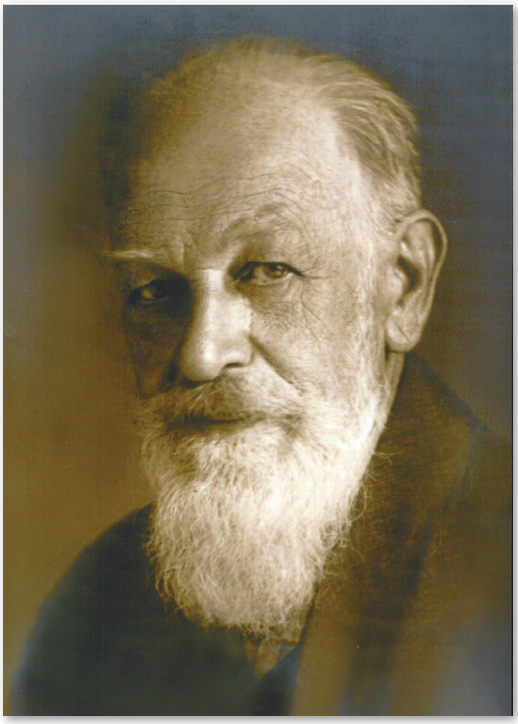
34 Harald Wettach, semi-structured interviews by Beti Žerovc, 2012.

35 "Mädchenheim Heimgard," *Cillier Zeitung* L/92, 15 November 1925, 3. The home for girls probably operated only between 1923 and 1928, when we also find newspaper advertisements for it. – "Announcement 9116-14," Neues *Wiener Journal* XXXVI/12.361, 22 April 1928, 39. For more, see Žerovc and Valant, "The Artistic Formation."

Considering that the programme of the institution changed several times in a few years, we can assume that it was not as successful as they had anticipated.[36]

The name *Heimgard* probably comes from the name of Wettach's holiday house, which appears in earlier correspondence under this name.[37] We can assume that the spouses selected this name following their beliefs, and so "the one that protects home(land)" probably alluded to the protection of the German homeland on the southern border of the German linguistic territory, which was, at the same time, a linguistically mixed territory. In advertisements, the adjective "Aryan" occasionally appeared in the name of the girls' home – *Arisches Mädchenheim Heimgard* – which perhaps testifies to how much and in what direction German nationalism intensified in the spouses as they grew older. The dire financial insecurity could have contributed to this, along with the loss of home and high standard of living that the Wettachs had previously enjoyed. The pain of this loss was, in the words of Marie Wettach's grandson, so great that even though she lived until 1967, she was never able to visit the house and the city in which she lived her most active years and where her four children grew up.[38]

Figure 5: Heinrich Wettach in his later years, photo



Source: Private collection

36 Ibid.

37 See the date on Heinrich Wettach's letter to Elsa Kastl, 7 August 1913, property of Angelika Hribar, Ljubljana.

38 Harald Wettach, semi-structured interviews by Beti Žerovc, 2012. For similar life stories and a broader context of such processes, see the literature listed in Žerovc, "Cultural and Historical, I.," fn 2.

The couple thus continued their life in Carinthia and tried to adjust to the new circumstances, but not only did they bitterly miss their home in the erstwhile Carniola – the painter was deeply affected by the lack of social and cultural life. Disappointed with historical events, including the post-war political order in the new Republic of Austria and with a palpable nostalgia for the time and the city in whose culture and arts he had played such an important role for almost four decades, he wrote to his former student Elsa Kastl:

> Those were beautiful, unforgettable years, full of lively cultural activity and flourishing society, particularly in music, which Ljubljana, once the Slavic hatred and arrogance thwarted everything that was German, will never live to see again. Thank God they are already being rewarded for that. The satisfaction over this is a small compensation for the massive losses that some of us had to suffer.[39]

Heinrich Wettach lived in St. Andrä on Lake Ossiach until his death on 1 October 1929. He is buried in the Protestant cemetery in Sankt Ruprecht (am Moos) near Villach.[40]

The quick departure from Ljubljana after the end of World War I, as well as events in World War II, probably contributed to the fact that Wettach's descendants no longer possess many works by their grandfather. The painter's grandson, Harald Wettach, for example, remembers that as a child in Carinthia, he saw a lot of material, including sketches for the allegories in the Philharmonic Society building, the painter's self-portrait in a dark suit, a huge folder of watercolours, different mountain motifs etc. Based on his narrative and the narrative of his aunt Brigitta, the house was used as the English officers' headquarters, while, considering the awkward situation, Marie Wettach eventually moved out. At that time, perhaps the majority of the painter's remaining works and documents were lost or destroyed.[41]

* * *

Heinrich Wettach taught and painted portraits of a number of Ljubljana citizens; he often exhibited his work in Ljubljana and, together with other Carniolans, competed for the limited opportunities in public tenders for painting commissions. His simultaneous outstanding contributions to the musical, social and cultural life of the Carniolan capital made him truly exceptional among its cultural workers, which leads us to conclude that for decades, he was just as recognisable in the streets of Ljubljana as his contemporaries Ivan Cankar and Rihard Jakopič, today Slovenian national icons.

39 Heinrich Wettach's letter to Elsa Kastl, 22 September 1920, property of Angelika Hribar, Ljubljana.

40 "Zum Tode Heinrich Wettachs," *Freie Stimmen* XLIX/238, 15 October 1929, 6.

41 Harald Wettach, semi-structured interviews by Beti Žerovc, 2012. After World War II, Austria, like Germany, was divided into four zones; Carinthia was under British military command. Brigitta Leitenberger remembers that her mother still lived in the house on Lake Ossiach: *"The English put tremendous pressure on her. They took everything but a small vestibule in front of her tiny attic studio, where she had her bed. And she lived there alone, without her children. She stayed until we told her that this cannot go on. The English then took the entire house from her."* – Brigitta Leitenberger, a semi-structured interview by Ruth Deutschmann, 1998.

It is clear that he, like many other German-leaning individuals, considered Carniola his home and was wholeheartedly dedicated to it,[42] strove to the best of his abilities to ensure that it flourished, yet was erased from its history because he imagined its political and cultural framework in a way that differed from (some) Slovenian-leaning Carniolans.[43] The actual course of history united Carniola after World War I with South Slavs, whereas Wettach belonged to those Carniolans who saw the past, present and future of the land not only in connection with the German cultural, but also political space.[44]

The present study thus tries to at least partially reconstruct the life course and worldview of a seriously overlooked artist in Slovenian art history, and also highlight a part of history without which the knowledge and understanding of Slovenian visual arts and culture at the turn of the century would simply not be complete. Slovenian history without considering the artistic endeavours within the cultural circle of the German-leaning population cannot exist, as German-leaning Carniolans are no less ancestors of contemporary Slovenians than the Slovenian-leaning Carniolans are. Likewise, they are no less important than the latter when it comes to creating the artistic field we have inherited. Without including this segment, we cannot truly understand everything that happened in the Carniolan sphere of visual arts and why, meaning that even "Slovenian" events remain misunderstood and illogically connected, because the events on both sides often happened in sequence, or even as a response to each other.[45]

42 Even upon his death, the Carinthian press called him a *German Carinthian (Deutschkrainer)*. We can explain that as a clear indicator that Wettach, regardless of his origin, felt a sense of belonging and attachment to Carniola. – "Heinrich Wettach," *Villacher Zeitung* XXVII/82, 12 October 1929, 5.

43 It seems telling, that the long-term Ljubljana mayor only mentioned the Wettach family in a single footnote in several hundred pages of his memoirs. The painter's son Reinhard appears in 1928 as "the son of an eminent German family" that allegedly carried out violence against Slovenian students. – Ivan Hribar, *Moji spomini* I, II (Ljubljana: Slovenska matica, 2022); originally in Ivan Hribar, "Moji spomini," *Trgovski list* XI/124, 18 October 1928, 4.

44 The Carniolan German-leaning population for the time being still floats somewhere in the background of the Slovenian perception of the past like some kind of amorphous mass, although there were many very different people whose worldviews were not even remotely unified. The artistic conservativism and national bigotry that we actually sense or at least suspect at Wettach were also not essential traits of this group; the freethinking, development oriented and anti-clerical Carniolans – particularly before the end of the 19th century – often saw connecting with the German cultural or cultural and political space as the only option that would lead to positive development of Carniola. On how the process of national differentiation of the Carniolan bourgeoisie was influenced by the opposition to the (Yugo)slav orientation of the Slovenian politics, which was to have a negative impact on the rights and will overall mean a civilisational step back for Carniola, see for example, Janez Cvirn, "Kdor te sreča, naj te sune, če ti more, v zobe plune," *Zgodovina za vse* XIV, No. 2 (2007): 52–54. It is worth adding that the national polarisation described in the article was predominantly a phenomenon of the small number of elites, while it affected the majority of the population differently and to a lesser extent.

45 For the process of erasing the Carniolan visual arts history that includes some of the actors from this article and leads towards impoverished and incorrect understanding of the Slovenian visual arts past, see Beti Žerovc, "Ivan Meštrović u Ljubljani 1903./1904.: Prijedlozi za dva spomenika i izlaganje s društvom Hagenbund u ljubljanskoj kazini," *Časopis za suvremenu povijest* 56, No. 2 (2024): 249–77.

Acknowledgement

The article is funded by the Slovenian Research and Innovation Agency (ARIS) as a part of the Research Program P6-0199 *History of Art of Slovenia, Central Europe and the Adriatic*.

Sources and Literature

Archival sources

Heinrich Wettach's letters to Elsa Kastl, Angelika Hribar's private archive, Ljubljana.

Bibliography

Cvirn, Janez. "Kdor te sreča, naj te sune, če ti more, v zobe plune." *Zgodovina za vse* 14, No. 2 (2007): 38–56.

Denkschrift zum 25 jährigen Bestande der Frauen-Ortsgruppe »Laibach« des deutschen Schulvereins: 1885–1910. Laibach: Verlag der Deutschen Schulvereinsschule in Laibach, 1910.

Dolenc, Ervin. "Deavstrizacija v politiki, upravi in kulturi v Sloveniji." In *Slovensko-avstrijski odnosi v 20. stoletju = Slowenischösterreichische Beziehungen im 20. Jahrhundert*, edited by Dušan Nećak et al., 81–94. Ljubljana: Oddelek za zgodovino Filozofske fakultete, 2004. http://hdl.handle.net/11686/26817.

Hribar, Ivan. *Moji spomini I, II*. Ljubljana: Slovenska matica, 2022 (reprint of the 1983 edition).

Ilić, Angela. "Podajmo si roke in srca!." *Stati inu obstati*, No. 23/24 (2016): 193–219.

Jahres-Bericht der philharmonischen Gesellscahft in Laibach: für die Zeit vom 1. Oktober 1885 bis 30. September 1886, edited by Friedrich Keesbacher. Laibach: Verlag der philharmonischen Gesellschaft, 1886.

Judson M., Pieter. "Versuche um 1900, die Sprachgrenze sichtbar zu machen." In *Die Verortung von Gedächtnis*, edited by Moritz Csáky and Peter Stachel, 163–74. Wien: Passagen Verlag, 2001.

Judson M., Pieter. *Guardians of the Nation: Activists on the Language Frontiers of Imperial Austria*. Cambridge MA: Harvard University Press, 2006.

Kuret, Primož. "Jubilejna koncertna sezona 1901–1902 Ljubljanske filharmonične družbe." *Muzikološki zbornik* 19 (1983), 41–50.

Kuret, Primož. *Ljubljanska filharmonična družba 1794–1919: Kronika ljubljanskega glasbenega življenja v stoletju meščanov in revolucij*. Ljubljana: Nova revija, 2005.

Matić, Dragan. *Nemci v Ljubljani: 1861–1918*. Ljubljana: Oddelek za zgodovino Filozofske fakultete, 2003.

Mihelič, Breda. "Vilska četrt med Prešernovo cesto in Tivolijem v Ljubljani ter prenova Wettachove vile: Problemi varovanja in prenove stanovanjske četrti." *Varstvo spomenikov* 39 (2003): 139–49.

Mikša, Peter. "'Da je Triglav ostal v slovenskih rokah, je največ moja zasluga.' Jakob Aljaž in njegovo planinsko delovanje v Triglavskem pogorju." *Zgodovinski časopis* 69, No. 1/2 (2015): 112–23.

Peternel, Marija Mojca. *"Ljubiteljem kranjskih Alp!": Kranjska podružnica Nemškega in avstrijskega planinskega društva*. Ljubljana: Založba Univerze, 2023.

Selišnik, Irena. "Status državljanstva ob nastanku nove Države SHS: Strategije izbire." *Zgodovinski časopis* 164, No. 3–4 (2021): 476–91.

Selišnik, Irena. "Usode uradnic in uradnikov po prvi svetovni vojni." *Retrospektive* 7, No. 1 (2024): 11–39.
Serše, Aleksandra. "Evangeljsko žensko društvo v Ljubljani 1856–1945." *Etnolog* 11 (2001): 57–66.
Stergar, Rok. "Continuity, Pragmatism, and Ethnolinguistic Nationalism: Public Administration in Slovenia during the Early Years of Yugoslavia." In *Hofratsdämmerung? Verwaltung und Ihr Personal in den Nachfolgestaaten der Habsburgermonarchie 1918 bis 1920*, edited by Peter Becker et al., 179–92. Wien: Böhlau, 2020.
Valant, Miha. "Ljubljansko društvo Kazina in združenja za likovno umetnost na Kranjskem med leti 1848 in 1918." Phd diss., Ljubljana: Faculty of Arts, 2023.
Valant, Miha, and Beti Žerovc. "Društva za likovno umetnost na Kranjskem v obdobju od 1848 do 1918." *Likovne besede*, No. 113 (2019): 4–13.
Weiss, Jernej. *Hans Gerstner. (1851–1939): Življenje za glasbo*. Maribor: Litera and Pedagoška fakulteta, 2010.
Zajc, Marko. "Kazina skozi čas." In *Zgodovinopisje v zrcalu zgodovine: 50 let inštituta za novejšo zgodovino*, edited by Aleš Gabrič, 127–38. Ljubljana: Inštitut za novejšo zgodovino, 2009.
Železnik, Sara. *Repertoarne smernice Filharmonične družbe v Ljubljani: Katalogi muzikalij Filharmonične družbe*. Ljubljana: Znanstvena založba Filozofske fakultete, 2014.
Žerovc, Beti. "Ivan Meštrović u Ljubljani 1903./1904.: Prijedlozi za dva spomenika i izlaganje s društvom Hagenbund u ljubljanskoj kazini." *Časopis za suvremenu povijest* 56, No. 2 (2024): 249–77.
Žerovc, Beti. "Cultural and Historical Overview of the Life of the Painter Heinrich Wettach (1858–1929), I. The Painter's Beginnings and Settling in Ljubljana." *Prispevki za novejšo zgodovino* 65, No. 3 (2025): 98–117.
Žerovc, Beti, and Miha Valant. "The Artistic Formation of the Painter Heinrich Wettach (1858–1929) and His Educational Work" [forthcoming].

Newspaper sources

Cillier Zeitung, 1925.
Deutsche Stimmen aus Krain Triest und Küstenland, Beilage des Grazer Tagblattes, 1901.
Deutsche Wacht, 1886.
Freie Stimmen, 1929.
Gottscheer Bote, 1919.
Grazer Tagblatt, 1896.
Jahres-Bericht der philharmonischen Gesellscahft in Laibach, 1886.
Laibacher Wochenblatt, 1886, 1889, 1892.
Laibacher Zeitung, 1887, 1891, 1892, 1896, 1901–1906, 1916, 1917.
Neues Wiener Journal, 1928.
Slovenec, 1914, 1920.
Slovenski narod, 1915.
Trgovski list, 1928.
Villacher Zeitung, 1929.

Oral sources

Harald Wettach, semi-structured interviews by Beti Žerovc, 2012.
Brigitta Leitenberger, semi-structured interview by Ruth Deutschmann, 1998 [The Austrian director Deutschmann conducted the interview as part of the *Chronisten* project: https://chronisten.at/].

Beti Žerovc

KULTURNOZGODOVINSKI ORIS ŽIVLJENJA SLIKARJA HEINRICHA WETTACHA (1858–1929), II. SLIKARJEVA DRUŽBENA IN DRUŠTVENA VPETOST V LJUBLJANI TER NJEGOVA ZADNJA LETA NA KOROŠKEM

POVZETEK

Članek obravnava družbeno vpetost slikarja Heinricha Wettacha (1858–1929), ki je deloval v Ljubljani od leta 1885 do konca prve svetovne vojne, ter ga s tem konkretneje usidra tudi v kulturni milje kranjske prestolnice na prelomu stoletja. Pretresa zlasti njegovo povezanost z društvi in organizacijami, s katerimi je bil najmočneje povezan, in opredeljuje naravo njegovega sodelovanja z njimi.

Iz različnih virov je razvidno, da je bil Wettach navdušen glasbenik, ki je že od leta 1885, torej vse od prihoda v Ljubljano, prepeval v moškem zboru Filharmonične družbe. Že leta 1887 ga zasledimo kot violinista v njenem orkestru, prav tako pa je kmalu postal redni član t. i. Gerstnerjevega godalnega kvarteta, v katerem je igral violino in violo. Od leta 1896 je naveden tudi v popisih članov ljubljanskega družabnega društva Kazina, kjer je sodeloval v najrazličnejših okvirih. Da mu za društvene dejavnosti ni bilo škoda časa, priča pogosto prevzemanje zamudnega izdelovanja raznih umetniških okrasitev prostorov, scenografij in kostumografij ter celo živih slik za številne družabne in dobrodelne dogodke. Kazina je v začetku 20. stoletja razvila tudi lasten program za likovno umetnost in k razstavljanju v svojih prostorih povabila nekatera vidnejša srednjeevropska umetniška združenja. Ob tujih in drugih avstrijskih avtorjih so na teh razstavah sodelovali domači kranjski umetniki, tudi Wettach.

Če je v okviru Kazine ali Filharmoničnega društva ter njunih dejavnosti Wettach sodeloval z zelo različno politično usmerjenimi posamezniki predvsem znotraj spektra avstrijsko centralistično in nemško čutečih Kranjcev, pa se je v Ljubljani pridružil še nekaterim izraziteje nemško orientiranim društvom. Vse od svojega prihoda v Ljubljano je, sprva kot zborovodja, nato pa tudi kot funkcionar, sodeloval s telovadnim društvom *Turnverein*. Oba s soprogo sta sodelovala vsak v svojem oddelku *Schulvereina*, ni pa popolnoma jasno, ali je bil Heinrich Wettach član moškega oddelka *Südmark*. Vsekakor je v ženskem oddelku tega društva sodelovala Marie Wettach. Po prestopu družine v evangeličansko vero leta 1901 je bila še posebej aktivna v ljubljanskem Evangelijskem ženskem društvu, tudi kot predsednica v letih 1904–1906 in ponovno 1910–1920.

Zadnji del članka obravnava neprostovoljno odselitev slikarja in njegove družine iz Ljubljane leta 1919 ter njegova zadnja leta na Koroškem. Družina Wettach je bila po prvi svetovni vojni svoji dve ljubljanski vili na zares elitni lokaciji prisiljena prodati za bagatelo. Po skromnem izkupičku od prodanih nepremičnin in prepovedi

prenosa denarnih sredstev iz nekdanje Kranjske v tedaj novi jugoslovanski državi v Republiko Nemško Avstrijo, sta zakonca Wettach na Koroškem prejemala podporo za revne. Naselila sta se v svoji počitniški hiši na Osojskem jezeru in tam, verjetno predvsem zaradi finančne stiske, odprla zasebno izobraževalno ustanovo, dekliški dom *Heimgard* (*Mädchenheim Heimgard*). Zaradi večkratnih sprememb programa ustanove v le nekaj letih lahko domnevamo, da ta ni delovala po njunih pričakovanjih. Slikar je v Sv. Andražu na Osojskem jezeru živel vse do smrti 1. oktobra 1929.

Pričujoča razprava tako po eni strani poskuša rekonstruirati vsaj del življenjske poti in nazorov spregledanega umetnika v slovenski umetnostni zgodovini, po drugi pa tudi košček zgodovine, brez katerega poznavanje in razumevanje slovenske likovne umetnosti in kulture na prelomu stoletja preprosto ne more biti celovito. Slovenske zgodovine brez upoštevanja umetnostnega dogajanja znotraj kulturnega kroga nemško čutečih namreč ne more biti, saj nemško čuteči Kranjci niso nič manj predniki sodobnih Slovencev, kot so to slovensko čuteči Kranjci. Prav tako tudi niso nič manj kot slednji oblikovali umetnostnega polja, ki smo ga podedovali. Brez vključevanja tega segmenta tako ne moremo zares razumeti, kaj vse se je v kranjskem likovnem polju dogajalo in zakaj, pri čemer tudi »slovenski« dogodki ostajajo nepravilno dojeti in nelogično povezani, saj so dogodki na obeh straneh pogosto potekali v sosledju ali celo kot odziv enih na druge.

1.01 DOI: https://doi.org/10.51663/pnz.65.3.15

Tjaša Konovšek*

Solidarity, Development, and Socialist Globalisation: The Centre for the Study and Cooperation of Yugoslavia with Developing Countries (1966–1973)

IZVLEČEK

CENTER ZA PROUČEVANJE IN SODELOVANJE JUGOSLAVIJE Z DRŽAVAMI V RAZVOJU (1966–1973): NASTANEK, DELOVANJE IN TRANSFORMACIJA

Center za proučevanje in sodelovanje Jugoslavije z državami v razvoju je bil ustanovljen v Ljubljani, glavnem mestu Socialistične republike Slovenije, v času, ko je bilo Gibanje neuvrščenih – in jugoslovanska vloga v njem – že dobro uveljavljeno, a hkrati na pragu stagnacije. Ustanovitev je bila del odziva na obstoječe stanje: njegov namen je bil okrepiti jugoslovansko neuvrščeno delovanje z institucionalizacijo produkcije znanja o državah v razvoju in njihovem sodelovanju z Jugoslavijo. Njegovo poslanstvo je bilo zasnovano interdisciplinarno, združevalo je raziskovanje, izobraževanje in podporo pri oblikovanju politik. Članek obravnava formativno obdobje centra med letoma 1966 in 1973, ga umešča v socialistično politično okolje in akademski sistem ter sledi njegovi preobrazbi v ustanovo zveznega pomena. Kljub pomanjkanju arhivskih virov članek pokaže, da je center deloval kot stičišče med znanstveno produkcijo, zunanjo politiko in mednarodnim sodelovanjem, ter s tem ponuja nov vpogled v institucionalne

* **PhD, Assistant, Institute of Contemporary History, Privoz 11, SI-1000 Ljubljana, tjasa.konovsek@inz.si; ORCID: 0000-0001-8872-692X**

dimenzije »neuvrščenosti od spodaj«. S tem, ko v ospredje postavlja institucijo in ne zgolj državne diplomacije, članek prispeva k sodobni historiografiji, ki znova preučuje infrastrukture socialistične globalizacije in aktivnosti, povezane z neuvrščenostjo.

Ključne besede: Center za proučevanje in sodelovanje Jugoslavije z državami v razvoju, socializem, znanje, solidarnost, razvoj

ABSTRACT

The Centre for the Study and Cooperation of Yugoslavia with Developing Countries was founded in Ljubljana, the capital of the Socialist Republic of Slovenia, during a time when the Non-Aligned Movement and Yugoslavia's role within it were well established but faced a period of stagnation. A part of the response to this impasse was the creation of this Centre, which aimed to strengthen Yugoslavia's engagement with the Non-Aligned Movement by institutionalising the generation of knowledge about developing countries and their cooperation with Yugoslavia. Its mission was interdisciplinary, combining research, education, and policy support. This article explores the Centre's formative years from 1966 to 1973, situating it within the socialist political environment and academic system, and tracing its evolution into an institution of federal significance. Despite the scarcity of archival sources, the article demonstrates how the Centre operated as a link between academic output, foreign policy, and international cooperation, providing new insights into the institutional aspects of "non-alignment from below." By emphasising an institution rather than state-level diplomacy, the article adds to recent historiography that reexamines the infrastructures of socialist globalisation and activities related to non-alignment.

Keywords: Centre for the Study and Cooperation of Yugoslavia with Developing Countries, socialism, knowledge, solidarity, development

An overview of current literature demonstrates that academic interest in alternative globalisations, socialist modernity, and the Non-Aligned Movement is growing. Many studies are decentralising the dominant Cold War narrative and instead proposing a multipolar perspective on global development.[1] This paper aims to enhance understanding of socialist modernity and globalising processes by examining a lesser-known research centre established in Ljubljana in 1966. The Centre for the Study and Cooperation of Yugoslavia with Developing Countries served as a hub, linking many

1 James Mark, Artemy M. Kalinovsky, and Steffi Marung, "Introduction," in James Mark, Artemy M. Kalinovsky, and Steffi Marung, eds., *Alternative Globalizations: Eastern Europe and the Postcolonial World* (Indiana: Indiana University Press, 2020), 1–32, accessed on 3 June 2025, https://doi.org/10.2307/j.ctvx8b7ph.4. Bojana Videkanić, "Nonaligned Modernism: Yugoslav Culture, Nonaligned Cultural Diplomacy, and Transnational Solidarity," *Nationalities Papers* 49, No. 3 (2021): 506, https://doi.org/10.1017/nps.2020.105.

of these issues within a single institution. While scholars familiar with the Centre's existence agree on its intriguing position and activities, systematic research remains limited, largely due to the lack of primary sources.[2] By acknowledging both the scarcity of sources and the importance placed on the Centre by current scholarship, this paper offers at least a partial overview of its existence, activities, and outputs. It thus supports the argument that the establishment of the Non-Aligned Movement (NAM) and Yugoslavia's related initiatives, rooted in a specific vision of globalisation, were acts of significant political agency, resulting from both the complex reality of the Cold War era and the intellectual and political traditions of the Yugoslav state.[3]

While the early period of the Centre's existence examined here may not correspond to its most productive or influential years, this contribution nonetheless explores its initial achievements and the new knowledge it generated. In subsequent years, this expertise helped forge stronger links between Yugoslavia and developing countries. In the mid-1960s, the founding of an institution dedicated solely to Yugoslavia's cooperation with NAM countries reflected the academic, political, and economic interests and needs of that era, including the export of technology, knowledge, and ideology. This paper demonstrates that the Centre was an institution rooted in its local environment, contributing in its specific way to the broader development of global socialism.

***

> New economic trends have emerged, particularly those leading to an increasing degree of unity and interdependence within the global economic market. Although this process is uneven and heterogeneous, it nevertheless encompasses the entire world, breaking down old structures and creating new forms and methods. The world has become a whole; the global market has established strong connections between all its parts, while simultaneously revealing the differences and antagonisms that exist within it. The main manifestations of these fundamental social contradictions are the conflict between the great powers and the issue of underdeveloped countries.[4]

2 While the Centre was active in different forms since 1966, there are no publicly available archival records. Its early production is scarcely catalogued and often unavailable. For further possible research options and state of the sources, see: Jure Ramšak, "An Attempt at an Alternative Globalization: The Slovenian Economy and the Developing Countries, 1970–1990 [Poskus drugačne globalizacije: slovensko gospodarstvo in dežele v razvoju 1970-1990]," *Acta Histriae* 23, No. 4 (2015): 767.

3 Mark, Kalinovsky, and Marung, "Introduction," 14. Videkanić, "Nonaligned Modernism," 507.

4 Vlado Benko, "The place and role of developing countries in the modern international community [Mesto in vloga dežel v razvoju v sodobni mednarodni skupnosti]," in *Yugoslavia and the Economic Development of Developing Countries: Symposium Papers*, Ljubljana, 23–24 June 1966 (Ljubljana: School of Political Science; Chamber of Commerce of Slovenia, 1966), 1–33.

With these words, Professor Vlado Benko[5] opened the introductory study resulting from a conference titled "Yugoslavia and Economic Development of the Developing Countries" that took place in Ljubljana in 1966. Organised by the College of Political Sciences (*Visoka šola za politične vede*) and the Chamber of Commerce and Industry of Slovenia (*Gospodarska zbornica Slovenije*), the conference brought together a diverse group of participants from Yugoslavia, representing various scientific traditions such as sociology, political science, and economics, to review, present, and plan Yugoslavia's involvement in cooperation with the so-called developing countries (*države v razvoju*). Benko's words encapsulated a key dilemma at the heart of Yugoslavia's engagement with the developing world at the time: how to navigate a global system marked by both integration and inequality. By framing the international order as both interdependent and antagonistic, he set the intellectual stage for establishing the Centre for the Study and Cooperation of Yugoslavia with Developing Countries later that same year. His opening was not only illustrative of contemporary socialist diagnoses of global contradictions but also pointed to the rationale for institutionalising research and cooperation with the developing countries. It anticipated the Centre's dual mission: to generate knowledge about developing countries and to translate the principles of solidarity and development into concrete policies and practices of non-alignment.

As Benko further noted, aiming to establish firmer and more equal economic relations within the developing global market, the conference covered a range of topics: trade in goods, investment cooperation,[6] organisational issues, acceleration of economic development, institutional elements of cooperation, and technical assistance between Yugoslavia and developing countries.[7] The conference thus provided immediate impetus for establishing the Centre for the Study and Cooperation of Yugoslavia with Developing Countries. By 1966, when the conference took place, the NAM was already a significant force in global politics, with Yugoslavia playing a crucial role in its formation and development. Besides Yugoslavia's political and economic position within the Movement, many indicators – such as the number of students from Non-Aligned countries studying at Yugoslav universities – also illustrate the dynamic

5 In 1961, Vladimir (Vlado) Benko (1917–2011) began working as a senior lecturer in the subject of International Relations and Foreign Policy of the SFRY at the then College of Political Sciences in Ljubljana. He pursued further academic training in the United States, France, and Sweden. In 1967, he was appointed Associate Professor of International Relations at the College of Political Sciences. From 1967 to 1970, he served as the Director of the College of Political Sciences (in 1968 renamed as the College of Sociology, Political Sciences, and Journalism). In 1966, he initiated the establishment of a research unit called the Centre for the Study of the SFRY's Cooperation with Developing Countries (now known as the Centre for International Cooperation and Development), which is the focus of this paper. – Boštjan Udovič, Bojko Bučar, and Milan Brglez, "Benko, Vladimir (1917–2011)," *Slovenska biografija* (Ljubljana: ZRC SAZU, 2013), accessed on 3 June 2025, http://www.slovenska-biografija.si/oseba/sbi1017730/#novi-slovenski-biografski-leksikon.

6 This mainly excluded loans, since Yugoslavia faced "a chronic shortage of financial capital". – Jure Ramšak, "Yugoslavia and the unlikely success of the New International Financial Order," *Godišnjak za društvenu istoriju* 31, No. 1 (2024): 39–53, accessed on 3 June 2025, https://udi.rs/godisnjak/godisnjak-za-drustvenu-istoriju-god-xxxi-sveska-1-2024/.

7 *Yugoslavia and the Economic Development of Developing Countries: Symposium Papers*, Ljubljana, 23–24 June 1966 (Ljubljana: School of Political Science; Chamber of Commerce of Slovenia, 1966).

exchange between countries.[8] This indicates that cooperation among individual states remained strong, while the top-down view of NAM activities suggested a relative stagnation of the Movement. In the second half of the 1960s, leadership changes occurred in several Non-Aligned states, including Algeria, Indonesia, Ghana, India, and others. No summits or other multilateral meetings were held. Confronted with this complex situation, the Yugoslav leaders, fully aware that Yugoslavia faced isolation and political constraints in Europe without the global Non-Aligned framework, sought ways to reinvigorate the NAM and draw attention to it once more.[9]

The establishment of the Centre for the Study and Cooperation of Yugoslavia with Developing Countries aimed to provide knowledge and information that would bridge the gap between existing ties and the apparent stagnation. By examining the Centre's emergence, early work, and transformation into a federal institution – although it was located in Ljubljana rather than in the federal capital, Belgrade – this paper builds on the concept of "*non-alignment from below*",[10] exploring how the NAM was constructed beyond, yet always in relation to, global political and economic developments. The paper will focus on two key concepts: solidarity as a cornerstone of socialist internationalism;[11] and development understood as a vital condition that enabled the full potential of a socialist society.[12] In doing so, the contribution will shed light on the Centre's existence and role, while also deepening understanding of specific visions of modernity that the Centre, its collaborators, and, by extension, the Yugoslav federation as a socialist state, projected upon the developing countries during the period of socialist globalisation.[13]

8 Aleš Gabrič, "Cultural and scientific cooperation of non-aligned countries in the shadow of political dilemmas [Kulturno in znanstveno sodelovanje neuvrščenih držav v senci političnih dilem]," in Barbara Predan, ed., *Robovi, stičišča in utopije prijateljstva: spregledane kulturne izmenjave v senci politike* (Ljubljana: Inštitut za novejšo zgodovino: Akademija za likovno umetnost in oblikovanje, 2022), 22. Dugonjic-Rodwin and Mladenović, "Transnational Educational Strategies," 336–42.

9 Jovan Čavoški, "Searching for a new meaning: Yugoslavia and global non-alignment in crisis 1965–1970 [U potrazi za novim smislom: Jugoslavija i kriza globalne nesvrstanosti 1965–1970]," Istorija 20. veka 39, No. 2 (2021): 353–74, accessed on 2 June 2025, https://doi.org/10.29362/ist20veka.2021.2.cav.353-374. Dragan Bogetić, "Yugoslavia and the Non-Aligned Movement," in Duško Dimitrijević and Jovan Čavoški, eds., *The 60th Anniversary of the Non-Aligned Movement* (Belgrade: Institute of International Politics and Economics, 2021), 239–53.

10 Paul Stubbs, "Introduction. Socialist Yugoslavia and the Non-Aligned Movement: Contradictions and Contestations," in Paul Stubbs, ed., *Socialist Yugoslavia and the Non-Aligned Movement: Social, Cultural, Political and Economic Imaginaries* (Montreal: McGill-Queen's University Press, 2023), 4.

11 Mark, Kalinovsky, and Marung, "Introduction," 14.

12 Alessandro Iandolo, "Socialist approaches to development," in Corinna R. Unger, Iris Borowy, and Corinne A. Pernet, eds., *The Routledge Handbook on the History of Development* (London: Routledge, 2022), 34–51, https://doi.org/10.4324/9780429356940-5.

13 Mark, Kalinovsky, and Marung, "Introduction," 13. James Mark and Paul Betts, "Introduction," in James Mark and Paul Betts, eds., *Socialism Goes Global: The Soviet Union and Eastern Europe in the Age of Decolonization* (Oxford: New York: Oxford University Press, 2022), 5.

The Centre's Establishment and Fields of Activity

The initiative to establish an institution dedicated to Yugoslavia's cooperation with developing countries in the early 1960s arose within a broader discussion of anti-imperialism, decoloniality, socialist solidarity, and peaceful coexistence supported by global economic cooperation. At the 1966 conference on economic cooperation, participants emphasised the need for systematic, institutionalised, and long-term monitoring of developing countries as a prerequisite for establishing closer and more stable relations. They acknowledged that the varied conditions of these countries posed specific challenges for economic engagement and that a lack of knowledge about them represented one of the main barriers to cooperation.[14] In response to these concerns, the Centre for the Study and Cooperation of Yugoslavia with Developing Countries was established in Ljubljana in December 1966. Building directly on the conference's conclusions, it aimed to address these knowledge gaps and encourage more effective cooperation.[15] It started operating in 1967 as a research unit of the College of Political Sciences at the University of Ljubljana. In 1973, it restructured itself as an autonomous research institution.

The interdisciplinary work of the Centre focused on two main goals: promoting collaboration between Yugoslavia and developing countries, and acquiring knowledge about the key political situations and existing legislation of the developing countries.[16] At the December 1966 meeting, Vlado Benko, the leading advocate for the Centre's establishment and its first director, stressed that beyond the general social interest and the conference's conclusions, the executive council of the Socialist Republic of Slovenia also backed the initiative.[17] Along with the available infrastructure, academic interest, and the general trend towards decentralising the Yugoslav federation – from the Constitution of 1963 to the constitutional amendments of 1971 – these factors helped ensure that the Centre remained in Ljubljana, despite its federal significance. While it was supported by the republic and the College, its founders also consulted several key organisations at both republican and federal levels – including the Chamber of Commerce and Industry of Slovenia, the Bureau for Market Research,[18] *Jugobanka*

14 Jože Korošec, "Institutional elements of the system of our economic cooperation to date with developing countries [Institucionalni elementi sistema našega dosedanjega ekonomskega sodelovanja deželami v razvoju]," in *Yugoslavia and the Economic Development of Developing Countries: Symposium Papers*, Ljubljana, 23–24 June 1966 (Ljubljana: School of Political Science; Chamber of Commerce of Slovenia, 1966), 1–47 [222–68].

15 *Establishment of the Centre for the Study and Cooperation of Yugoslavia with Developing Countries*. The document is kept at the Centre for International Cooperation and Development, Kardeljeva ploščad 1, 1000 Ljubljana. https://www.cmsr.si/.

16 *Research Centre for Cooperation with Developing Countries Ljubljana-Yugoslavia* (Ljubljana: Center za proučevanje sodelovanja z deželami v razvoju, 1981), 3.

17 *Invitation and record of the 7th sitting of the pedagogical-scientific council at the College of political science, December 10 1966*. The document is kept at the Centre for International Cooperation and Development, Kardeljeva ploščad 1, 1000 Ljubljana. https://www.cmsr.si/.

18 Active at the Chamber of commerce of Slovenia, established in 1963 by Franc Tretjak. – "Tretjak, Franc (1914–2009)," *Slovenska biografija* (Ljubljana: ZRC SAZU, 2013), accessed on 6 June 2025, http://www.slovenska-biografija.si/oseba/sbi722091/#slovenski-biografski-leksikon.

bank,[19] the Institute for International Technical Cooperation,[20] and Intertrade[21] – all of which expressed their willingness to collaborate with the Centre.[22]

To generate new knowledge and provide expertise, the Centre focused on two main activities. It delivered lectures on collaborating with developing countries as part of the College curriculum, connecting students with professionals in economics and international relations. Students also took part in the Centre's research projects, while new findings were incorporated into regular teaching. Additionally, the Centre provided specialised training for commercial officers, economic staff, technical assistance personnel, and other specialists. It organised conferences, meetings, and seminars, and prepared and published studies, manuals, and bibliographies. Furthermore, the Centre planned to introduce a new subject at the College: Socio-Economic Processes in Developing Countries (*družbeno-ekonomski procesi v državah v razvoju*). This course was designed to combine developmental and historical perspectives. Students would investigate colonialism and its impacts, analyse national liberation struggles, and explore the pursuit of state sovereignty and economic independence. Aimed at fostering an ideological foundation, the subject sought to enhance understanding of cooperation and relations between states.[23]

The Key Concepts in Early Scientific Production

The research team employed at the Centre since its inception was interdisciplinary, which, among other factors, directly influenced its knowledge production. It is unclear how many collaborators worked there between 1966 and 1973. However, until 1980, the Centre employed around 20 researchers from fields such as economics, political science, law, sociology, and communications.[24] In its early efforts, research was often framed through the concepts of solidarity and development. Initial interpretations offered little critique of these concepts and were mainly used to criticise the politics and economic activities of Western countries. Nonetheless, the understanding

19 Jugobanka – Yugoslav Bank for Foreign Trade (*Jugoslovanska banka za zunanjo trgovino*) was a part of the National bank of the Federal People's Republic of Yugoslavia. Dušan Mramor, "Overview of the institutional structure of the banking and credit system and certain other parts of Yugoslavia's economic system, 1945–1983 [Prikaz institucionalne ureditve bančno-kreditnega in nekaterih drugih delov ekonomskega sistema Jugoslavije 1945–1983]," *Bančni vestnik: revija za denarništvo in bančništvo* 34, No. 1 (1985): 17–20.

20 Established in 1963. – Gabrič, "Cultural and scientific cooperation," 11.

21 Intertrade was engaged in trade with developing countries, promoting the exchange of goods, the transfer of technology, and the development of industrial enterprises. – Janez Demšar et al., *25 let. Intertrade (Ljubljana)* (Ljubljana: Delo, 1977).

22 *Establishment of the Centre for the Study and Cooperation of Yugoslavia with Developing Countries.*

23 Ibid.

24 *Research Centre for Cooperation with Developing Countries*, 17–18. Cf. Maja Korolija, "Yugoslav science during the Cold War (1945–1960): socio-economic and ideological impacts of a geopolitical shift," *Humanities and Social Sciences Communications* 10, 913 (2023), https://doi.org/10.1057/s41599-023-02414-2. Davor Boban and Ivan Stanojević, "The Institutionalisation of Political Science in Post-Yugoslav States: Continuities and New Beginnings," in Gabriella Ilonszki and Christophe Roux, eds., *Opportunities and Challenges for New and Peripheral Political Science Communities* (Cham: Palgrave Macmillan, 2022), 87–118, https://doi.org/10.1007/978-3-030-79054-7_4.

of development differed from the Soviet model by allowing more room for bottom-up initiatives. Instead of insisting on a strictly state-led approach, the Centre's research acknowledged the importance of the personal experiences of economic experts and promoted building on the limited but vital knowledge of individuals directly involved with developing countries.

Assessing the state of global affairs and Yugoslavia's role within them, the two concepts were also connected to time and speed, which, as the authors noted, was a key issue leading to increasing gaps between the global West and developing postcolonial countries. By training foreign experts and offering material support to developing countries, the Centre argued that Yugoslavia could help accelerate modernisation and thus reduce the disparity between states on the world stage. At the same time, however, the authors acknowledged "subjective factors" related to the ideas of solidarity and development. They identified areas where Yugoslavia and other socialist states could directly implement further changes. These included the political and economic conditions in the developing countries and the politics of NAM, which was viewed as the most effective safeguard for successful economic cooperation, territorial integrity, independence, and the spread of socialist ideas.[25] The need for another NAM conference was clearly expressed in the hope of achieving new, more decisive actions in economic cooperation and issues in the developing countries.[26]

Acknowledgement of "subjective factors", such as domestic economic reform in Yugoslavia, created space for Yugoslavia's agency in relation to developing countries by establishing clearer legislation and enhancing the competencies of state bodies.[27] Because Yugoslavia had limited capacity to finance the economic development of developing countries, technical and scientific cooperation gained greater significance. In the second half of the 1960s, the Centre issued guidelines for such support: indeed, the Yugoslav state was supposed to provide more financial backing for technical and scientific cooperation, but it was Yugoslav economic organisations that were expected to take the lead, mainly by offering expert support alongside the export of materials, machines, and equipment. This was similar to the documents discussed after the NAM ministers' conference in Georgetown, Guyana, in August 1972, and again at the 1974 UNCTAD (United Nations Conference on Trade and Development) meeting, which circulated and were discussed in Yugoslavia at the time. The topics included student

25 Avguštin Lah, "Foreword [Predgovor]," in Rodoljub Jemuović and Avguštin Lah, *Scientific, Technical, and Cultural Cooperation of Yugoslavia with Developing Countries [Naučna, tehnička i kulturna saradnja Jugoslavije sa zemljama u razvoju]* (Ljubljana: Univerza v Ljubljani, Fakulteta za sociologijo, politične vede in novinarstvo: Center za proučevanje sodelovanja z deželami v razvoju, 1972), i.

26 Ivo Pelicon, "Summary Assessment and Certain Problems of Yugoslavia's Economic Cooperation with Developing Countries in the Period from 1966 to 1967 [Sumarna ocena in nekateri problemi gospodarskega sodelovanja Jugoslavije z deželami v razvoju v času od 1966 do 1967]," in *Yugoslavia's Economic Relations with Developing Countries:* Conference Materials, Ljubljana, June 1968 (Ljubljana: School of Political Science, 1968), 1, 2.

27 Ivo Fabinc, *Economic Relations of Yugoslavia with Developing Countries: Final Study [Ekonomski odnosi Jugoslavije sa zemljama u razvoju: finalna studija]* (Ljubljana: Univerza v Ljubljani, Fakulteta za sociologijo, politične vede in novinarstvo: Center za proučevanje sodelovanja z deželami v razvoju, 1972), 11.

exchange, training of planning experts, the exchange of information and ideas, and the establishment of joint schemes and institutions for technology transfer and training.[28]

The Centre found that some of the proposed measures for cooperation with developing countries were financially less profitable or even unprofitable for Yugoslavia. Hence, the researchers employed at the Centre suggested that the state should consider providing compensation to offset the negative effects of such cooperation. When discussing the Centre's findings, the researchers explained that, although cooperation might only yield benefits in the long term or prove unprofitable due to global uncertainties such as war, political upheaval, and shifting economic trends, it was nonetheless understood through the principle of solidarity. It offered incentives for the global integration of the Yugoslav economy.[29] Solidarity also facilitated the activation of capacities such as exchanging and disclosing technical documentation and information to developing countries; establishing expert training centres in those countries; organising workshops; providing scholarships; strengthening ties with Yugoslav experts experienced in working with developing countries; and establishing clear legislation to define relationships between the federation and its units as well as responsibilities of existing political and expert bodies.[30] While solidarity served as a powerful argument for cooperation, it also faced practical issues such as the limited existence of institutions, regulations, and legislation that could effectively channel desired actions – particularly during internal or international conflicts.[31]

In line with the Centre's earliest project, titled "Economic, Political, Scientific-Technical, and Cultural Cooperation of Yugoslavia with Developing Countries" [*Gospodarsko, politično, znanstveno-tehnično in kulturno sodelovanje Jugoslavije z deželami v razvoju*],[32] the Centre established a comprehensive information service: the INDOK. It included literature, data, and information collection, based on the Centre's own studies but greatly supplemented by collaborations with other research institutions, research centres in banks in Yugoslavia and abroad, and its own network

28 SI AS 1140, container 10, unit of description 297. Federal ZAMTES – Centre for the Transfer of Science and Technology among Non-Aligned Countries – UNCTAD [Zvezni ZAMTES – Center za transfer znanosti in tehnologije med neuvrščenimi deželami – UNCTAD.]

29 Jure Ramšak, "Yugoslavia and the ambivalence of the south-south economic cooperation in the 1970s and 1980s [Jugoslavija i ambivalentnost ekonomske saradnje Jug-Jug u sedamdesetim i osamdesetim godinama]," *Tokovi istorije* 1 (2024), 204–24.

30 Ibid., 22, 23. Miloš Vuksanović, "Technical cooperation as an important element of our relations with developing countries [Tehnično sodelovanje kot pomemben element naših odnosov z deželami v razvoju]," in *Yugoslavia's Economic Relations with Developing Countries:* Conference Materials, Ljubljana, June 1968 (Ljubljana: School of Political Science, 1968), 112.

31 Mirko Žarić, "Basic Assumptions and Dilemmas of Action-Oriented Measures and a More Organized Role of Non-Alignment on the International Level [Osnovne pretpostavke i dileme akcionog dejstva i organizirovanije uloge nesvrstanosti na medjunarodnom planu]," in Mirko Žarić and Dragoslav Pejić, eds., *Non-alignment and Yugoslavia [Nesvrstanost i Jugoslavija]* (Ljubljana: Univerza v Ljubljani, Fakulteta za sociologijo, politične vede in novinarstvo: Center za proučevanje sodelovanja z deželami v razvoju, 1972), 56.

32 The project was supported by major Yugoslav institutions working with developing countries: Jugobanka, Federal Secretariat for Foreign Trade, Investment Bank, Fund for the Financing and Insurance of Export Transactions, Federal Chamber of Commerce, Federal Secretariat for Information, and Federal Fund for the Financing of Scientific Activities. – Janez Terček, "Editorial," in *Bulletin [Bilten]*. Ljubljana: Center za proučevanje sodelovanja z deželami v razvoju 1, No. 1 (1970): 3–7.

of correspondents from developing countries and international organisations. Besides managing the Centre's production, the INDOK also provided information to support economic cooperation with developing countries, such as legislative overviews, and organised symposiums, seminars, workshops, and consultations for both Yugoslav and foreign personnel and students.[33]

Transformation: An Independent Research Institution

By the late 1970s, the recognition of the power of the scientific-technical revolution became the central argument when pursuing further collaborations and strengthening trade with developing countries to support their economic development by enabling them to rely on their own potential, as well as regarding the efforts to establish the New International Economic Order and further cooperation within the NAM.[34] Science became a factor that helped establish the level of both economic and social development. By acknowledging international developments in technical cooperation,[35] the Centre extended criticism towards the concentration of wealth, knowledge, and power in developing countries.[36]

Throughout the socialist period, the Centre remained a vital institution, providing expertise, information, and support to the state and Yugoslav companies, which relied on existing studies when planning long-term strategies for cooperation with developing countries.[37] It has, however, faced many transformations. The first one occurred in the early 1970s. By that time, the Centre's primary focus had shifted more notably from its initial plan to research the economic and social development of developing countries equally. It focused predominantly on the economic aspect, while still including the interdisciplinary focus established in its early years. In March 1973, the Centre therefore separated itself from the former College of Political Sciences (since 1970, the Faculty of Sociology, Political Science and Journalism) and renamed itself the Centre

33 Ibid.

34 Justyna Pierzyńska, "Collective self-reliance: A portrait of a Yugoslav development strategy," *Miscellanea Geographica Sciendo*, 16, No. 2 (2012): 30–35, https://doi.org/10.2478/v10288-012-0024-3.

35 Mainly the 1978 Conference in Buenos Aires and the subsequent Plan of Action for Promoting and Implementing Technical Cooperation among Developing Countries (TCDC). – *Buenos Aires Plan of Action* (1978), accessed on 25 July 2025, https://unsouthsouth.org/bapa40/documents/buenos-aires-plan-of-action/. About the economic and diplomatic cooperation of Slovenia (Yugoslavia) with developing countries, see: Jure Ramšak, "'Socialist' economic diplomacy: activities of the Socialist Republic of Slovenia in the field of international economic relations 1974–1980 ['Socialistična' gospodarska diplomacija: dejavnost Socialistične republike Slovenije na področju mednarodnih ekonomskih odnosov 1974-1980]," *Annales* 24, No. 4 (2014): 733–48, http://www.zdjp.si/sl/docs/annales/sociologia/n24-4/ramsak.pdf.

36 Janez Rogelj, *Scientific and Technical Cooperation of the Socialist Republic of Slovenia with Developing Countries in the Period 1976–1979/1980 [Znanstveno-tehnično sodelovanje SR Slovenije z deželami v razvoju v obdobju 1976 – 1979/1980]* (Ljubljana: Center za sodelovanje proučevanja z deželami v razvoju Ljubljana, 1981), 1–4.

37 SI AS 1140, container 71, unit of description 1058. Elements of the Strategy for Economic Cooperation of the Socialist Republic of Slovenia with Developing Countries (Concept), Republican Committee for International Cooperation. [Elementi strategije ekonomskega sodelovanja Socialistične Republike Slovenije z Državami v razvoju (koncept), Republiški komite za mednarodno sodelovanje.]

for the Study of Cooperation with Developing Countries [*Center za proučevanje delovanja z deželami v razvoju*].[38]

This reorganisation affected both the internal operations of the Centre and its links with other institutions. Firstly, the group of founding members expanded, and their membership became more stable while their roles became more defined. The new members included the Executive Council of the Socialist Republic of Slovenia, the Yugoslav bank for foreign economic relations and *Ljubljanska banka*, the Chamber of Commerce of Yugoslavia, the Research Community of Slovenia, the Self-Managed Interest Community for Foreign Economic Relations Vojvodina – Novi Sad, the Self-Managed Interest Community for Foreign Economic Relations Slovenia – Ljubljana, and the Federal Secretariat for Foreign Affairs.[39] While before the transformation, research projects were financed by federal institutions (for example, the Federal Fund for the Financing of Scientific Activities [*Savezni fond za finansiranje naučnih aktivnosti*]) and "*other interested parties*",[40] the clearly defined and registered financial responsibilities of the founding members became one of the primary financial sources for the Centre.[41] The funding members, along with other permanent members, contributed about one-third of the Centre's budget.[42]

While the transformation allowed the Centre to select research topics and broaden its activities more independently, the framework of its efforts remained aligned with its original purpose. It continued to be a research institution that advised, educated, and informed its members, clients, and state institutions about the economic, social, and political situations of developing countries, as well as their roles within the international community. It also contributed to fostering more democratic and solidarity-based international relations. By combining practical experience gained from previous collaborations with developing countries and insights from other Western nations, the Centre also provided a platform for discussion, knowledge creation, and information exchange concerning activities within the NAM and among developing countries.[43]

The Centre primarily concentrated on economic cooperation between Yugoslavia and developing countries within the framework of the New International Economic Order. This resulted in a series of studies on individual countries and research into

38 *Statute of the Centre for the Study of Cooperation with Developing Countries*. The document is kept at the Centre for International Cooperation and Development, Kardeljeva ploščad 1, 1000 Ljubljana, https://www.cmsr.si/.

39 The rest was contributed by research communities or gained in the framework of the projects, funded by the federation, republic, or international institutions or organisations. – *Research Centre for Cooperation with Developing Countries*, 33.

40 Centre for the Study of Cooperation with Developing Countries, "Foreword," in Ivo Fabinc, *Economic Relations of Yugoslavia with Developing Countries: Final Study [Ekonomski odnosi Jugoslavije sa zemljama u razvoju: finalna studija]* (Ljubljana: Univerza v Ljubljani, Fakulteta za sociologijo, politične vede in novinarstvo: Center za proučevanje sodelovanja z deželami v razvoju, 1972). Žarko Lazarević, "Yugoslavia: economic historiography between national and international context," in Antonie Dolezalova and Catherine Albrecht, eds., *Behind the Iron Curtain: Economic Historians during the Cold War, 1945–1989* (Basingstoke: Palgrave Macmillan, 2023), 200.

41 *Centre for International Cooperation and Development, 6 April 1973, at the Court in Ljubljana RGZ UV 380/1*. The document is kept at the Centre for International Cooperation and Development, Kardeljeva ploščad 1, 1000 Ljubljana, https://www.cmsr.si/.

42 *Research Centre for Cooperation with Developing Countries*, 4.

43 *Statute of the Centre for the Study of Cooperation with Developing Countries.*

specific issues related to cooperation (transport, legislation, investments, etc.). The Centre also enhanced its cooperation with the United Nations and organisations within the federation dedicated to similar activities. Such collaborations included work with the International Centre for Public Enterprises in developing countries, where the legal and economic aspects of technology transfer from Yugoslavia to these countries, along with the role of communication and information systems, gained further significance.[44]

In Summary: Prospects

While both the topics and the Centre as an institution warrant further research, initial findings, combined with existing literature on the matter, indicate that the Centre played a significant role within the federal political system and scientific production. Through its publications, expertise, and educational activities, it provided vital support for collaboration between Yugoslavia and developing countries. Moreover, by analysing the Centre's output – particularly its key arguments, framework, and rationale for the studies it conducted and published – it is possible to identify two core concepts that were both established and adaptable: development and solidarity. Each carried its own political implications, which were often reflected in the Centre's research focus. However, they gained additional and distinct meanings when utilised in the context of the Centre's scientific work: offering ideological justification while also masking any potential adverse effects of their findings and proposals. These concepts thus not only linked the Centre to existing political and economic power structures but also granted considerable agency to its activities. While many questions regarding the Centre's research practices, daily operations, and closer ties with economic organisations and political structures remain unanswered, this early research demonstrates potential for connecting global developments to hands-on political, scientific, and economic practices. Further investigation into the Institute for Developing Countries – another research centre established in 1963 in Zagreb – would also shed more light on the questions raised (and partially addressed) in this contribution.

Acknowledgement

This article has been prepared as part of the ERC Perspective Research Project *Socialist Management in a Global Context: Technocratic Developments in the Soviet Union and Yugoslavia, 1955–1991*, funded by the Slovenian Research and Innovation Agency (ARIS), project code N6-0399 (B); and research program *Political history*, code P6-0281.

44 *Research Centre for Cooperation with Developing Countries*, 3, 4.

Sources and Literature

Archival sources

SI AS – Arhiv Republike Slovenije:
SI AS 1140, ZAMTES – Zavod za mednarodno znanstveno in tehnično sodelovanje SR Slovenije.

Literature

Boban, Davor, and Ivan Stanojević. "The Institutionalisation of Political Science in Post-Yugoslav States: Continuities and New Beginnings." In *Opportunities and Challenges for New and Peripheral Political Science Communities*, edited by Gabriella Ilonszki and Christophe Roux, 87–118. Cham: Palgrave Macmillan, 2022. https://doi.org/10.1007/978-3-030-79054-7_4.

Bogetić, Dragan. "Yugoslavia and the non-aligned movement." *The 60th Anniversary of the Non-Aligned Movement*, edited by Duško Dimitrijević and Jovan Čavoški, 239–53. Belgrade: Institute of International Politics and Economics, 2021.

Čavoški, Jovan. "Searching for a new meaning: Yugoslavia and global non-alignment in crisis 1965–1970 [U potrazi za novim smislom: Jugoslavija i kriza globalne nesvrstanosti 1965–1970]." *Istorija 20. veka* 39, No. 2 (2021): 353–74. Accessed 2 June 2025. https://doi.org/10.29362/ist20veka.2021.2.cav.353-374.

Dugonjic-Rodwin, Leonora, and Ivica Mladenović. "Transnational Educational Strategies during the Cold War: Students from the Global South in Socialist Yugoslavia, 1961–91." In *Socialist Yugoslavia and the non-aligned movement: social, cultural, political and economic imaginaries*, edited by Paul Stubbs, 331–57. Montreal: McGill-Queen's University Press, 2023.

Gabrič, Aleš. "Cultural and scientific cooperation of non-aligned countries in the shadow of political dilemmas [Kulturno in znanstveno sodelovanje neuvrščenih držav v senci političnih dilem]." In *Robovi, stičišča in utopije prijateljstva: spregledane kulturne izmenjave v senci politike*, edited by Barbara Predan, 9–27. Ljubljana: Inštitut za novejšo zgodovino : Akademija za likovno umetnost in oblikovanje, 2022.

Iandolo, Alessandro. "Socialist approaches to development." In *The Routledge Handbook on the History of Development*, edited by Corinna R. Unger, Iris Borowy, and Corinne A. Pernet, 34–51. London: Routledge, 2022. https://doi.org/10.4324/9780429356940-5.

Jakovina, Tvrtko. *The Third Side of the Cold War [Treća strana hladnog rata]*. Zaprešić: Fraktura, 2011.

Korolija, Maja. "Yugoslav science during the Cold War (1945–1960): socio-economic and ideological impacts of a geopolitical shift." *Humanities and Social Sciences Communications* 10, 913 (2023). https://doi.org/10.1057/s41599-023-02414-2.

Lazarević, Žarko. "Yugoslavia: economic historiography between national and international context." In *Behind the Iron Curtain: Economic Historians during the Cold War, 1945–1989*, edited by Antonie Dolezalova and Catherine Albrecht, 195–221. Basingstoke: Palgrave Macmillan, 2023.

Lüthi, M. Lorenz. "The Non-Aligned Movement and the Cold War, 1961–1973." *Journal of Cold War Studies* 18, No. 4 (2016): 98–147. Accessed 20 May 2025, https://www.jstor.org/stable/26925642.

Mark, James, and Paul Betts. "Introduction." In *Socialism Goes Global: The Soviet Union and Eastern Europe in the Age of Decolonization*, edited by James Mark and Paul Betts, 1–24. Oxford: New York: Oxford University Press, 2022.

Mark, James, Artemy M. Kalinovsky, and Steffi Marung. "Introduction." In *Alternative Globalizations: Eastern Europe and the Postcolonial World*, edited by James Mark, Artemy M. Kalinovsky, and

Steffi Marung. Indiana: Indiana University Press, 2020, 1–32. Accessed 3 June 2025. https://doi.org/10.2307/j.ctvx8b7ph.4.

Mišković, Nataša. "Introduction." In *The Non-Aligned Movement and the Cold War. Delhi – Bandung – Belgrade*, edited by Nataša Mišković, Harald Fischer-Tiné and Nada Boškovska, 1–18. London: New York: Routledge, 2014.

Mramor, Dušan. "Overview of the institutional structure of the banking and credit system and certain other parts of Yugoslavia's economic system, 1945–1983 [Prikaz institucionalne ureditve bančno-kreditnega in nekaterih drugih delov ekonomskega sistema Jugoslavije 1945–1983]." *Bančni vestnik: revija za denarništvo in bančništvo* 34, No. 1 (1985): 14–25.

Ramšak, Jure. "An Attempt at an Alternative Globalization: The Slovenian Economy and the Developing Countries, 1970–1990 [Poskus drugačne globalizacije: slovensko gospodarstvo in dežele v razvoju 1970-1990]." *Acta Histriae* 23, No. 4 (2015): 765–82.

Ramšak, Jure. "'Socialist' economic diplomacy: activities of the Socialist Republic of Slovenia in the field of international economic relations 1974–1980 ['Socialistična' gospodarska diplomacija: dejavnost Socialistične republike Slovenije na področju mednarodnih ekonomskih odnosov 1974-1980]." *Annales* 24, No. 4 (2014): 733–48. http://www.zdjp.si/sl/docs/annales/sociologia/n24-4/ramsak.pdf.

Ramšak, Jure. "Yugoslavia and the ambivalence of the south-south economic cooperation in the 1970s and 1980s [Jugoslavija i ambivalentnost ekonomske saradnje Jug-Jug u sedamdesetim i osamdesetim godinama]." *Tokovi istorije* 1 (2024): 204–24.

Ramšak, Jure. "Yugoslavia and the unlikely success of the New International Financial Order." *Godišnjak za društvenu istoriju* 31, No. 1 (2024): 39–53. Accessed 3 June 2025. https://udi.rs/godisnjak/godisnjak-za-drustvenu-istoriju-god-xxxi-sveska-1-2024/.

Rothermund, Dietmar. "The era of non- alignment." In *The Non-Aligned Movement and the Cold War. Delhi – Bandung – Belgrade*, edited by Nataša Mišković, Harald Fischer-Tiné, and Nada Boškovska, 19–34. London: New York: Routledge, 2014.

Stubbs, Paul. "Introduction. Socialist Yugoslavia and the non-aligned movement: contradictions and contestations." In *Socialist Yugoslavia and the Non-aligned Movement: Social, Cultural, Political and Economic Imaginaries*, edited by Paul Stubbs, 3–33. Montreal: McGill-Queen's University Press, 2023.

Udovič, Boštjan, Bojko Bučar, and Milan Brglez. "Benko, Vladimir (1917–2011)." *Slovenska biografija*. Ljubljana: ZRC SAZU, 2013. Accessed 3 June 2025. http://www.slovenska-biografija.si/oseba/sbi1017730/#novi-slovenski-biografski-leksikon.

Uredništvo. "Tretjak, Franc (1914–2009)." *Slovenska biografija* Ljubljana: ZRC SAZU, 2013. Accessed 6 June 2025. https://www.slovenska-biografija.si/oseba/sbi722091/#slovenski-biografski-leksikon.

Videkanić, Bojana. "Nonaligned Modernism: Yugoslav Culture, Nonaligned Cultural Diplomacy, and Transnational Solidarity." *Nationalities Papers* 49, No. 3 (2021): 504–22. https://doi.org/10.1017/nps.2020.105.

Životić, Aleksandar, and Jovan Čavoški. "On the Road to Belgrade: Yugoslavia, Third World Neutrals, and the Evolution of Global Non-Alignment, 1954–1961." *Journal of Cold War Studies* 18 (2016): 79–97. Accessed 20 May 2025. https://doi.org/10.1162/JCWS_a_00681.

Other sources

Center for International Cooperation and Development, April 6, 1973, at the Court in Ljubljana RGZ UV 380/1. The document is kept at the Centre for international cooperation and development, Kardeljeva ploščad 1, 1000 Ljubljana. https://www.cmsr.si/.

Establishment of the Center for the Study and Cooperation of Yugoslavia with Developing Countries. The document is kept at the Centre for international cooperation and development, Kardeljeva ploščad 1, 1000 Ljubljana. https://www.cmsr.si/.

Invitation and record of the 7th sitting of the pedagogical-scientific council at the College of political science, December 10 1966. The document is kept at the Centre for international cooperation and development, Kardeljeva ploščad 1, 1000 Ljubljana. https://www.cmsr.si/.

Statute of the Center for the study of cooperation with developing countries. The document is kept at the Centre for international cooperation and development, Kardeljeva ploščad 1, 1000 Ljubljana. https://www.cmsr.si/.

Printed sources

Benko, Vlado. "The place and role of developing countries in the modern international community [Mesto in vloga dežel v razvoju v sodobni mednarodni skupnosti]." In *Yugoslavia and the Economic Development of Developing Countries: Symposium Papers*, Ljubljana, 23–24 June 1966, 1–33. Ljubljana: School of Political Science; Chamber of Commerce of Slovenia, 1966.

Center for the study of cooperation with the developing countries. "Foreword." In Ivo Fabinc, *Economic Relations of Yugoslavia with Developing Countries: Final Study [Ekonomski odnosi Jugoslavije sa zemljama u razvoju: finalna studija]*, i–ii. Ljubljana: Univerza v Ljubljani, Fakulteta za sociologijo, politične vede in novinarstvo: Center za proučevanje sodelovanja z deželami v razvoju, 1972.

Demšar, Janez, et al. *25 let. Intertrade (Ljubljana)*. Ljubljana: Delo, 1977.

Fabinc, Ivo. *Economic Relations of Yugoslavia with Developing Countries: Final Study [Ekonomski odnosi Jugoslavije sa zemljama u razvoju: finalna studija]*, 10, 11. Ljubljana: Univerza v Ljubljani, Fakulteta za sociologijo, politične vede in novinarstvo: Center za proučevanje sodelovanja z deželami v razvoju, 1972.

Jemuović, Rodoljub, and Avguštin Lah. "Activities and Experiences from the Cooperation of Some International Organizations and Developed Countries with Developing Countries [Aktivnosti i iskustva iz saradnje nekih nedjunarodnih organizacija i razvijenih zemaljasa zemljama u razvoju]." In Rodoljub Jemuović and Avguštin Lah, *Scientific, Technical, and Cultural Cooperation of Yugoslavia with Developing Countries [Naučna, tehnička i kulturna saradnja Jugoslavije sa zemljama u razvoju]*, 19–32. Ljubljana: Univerza v Ljubljani, Fakulteta za sociologijo, politične vede in novinarstvo: Center za proučevanje sodelovanja z deželami v razvoju, 1972.

Korošec, Jože. "Institutional elements of the system of our economic cooperation to date with developing countries [Institucionalni elementi sistema našega dosedanjega ekonomskega sodelovanja deželami v razvoju]." In *Yugoslavia and the Economic Development of Developing Countries: Symposium Papers*, Ljubljana, 23–24 June 1966. 1–47 [222–68]. Ljubljana: School of Political Science; Chamber of Commerce of Slovenia, 1966.

Lah, Avguštin. "Foreword [Predgovor]." In Rodoljub Jemuović and Avguštin Lah, *Scientific, Technical, and Cultural Cooperation of Yugoslavia with Developing Countries [Naučna, tehnička i kulturna saradnja Jugoslavije sa zemljama u razvoju]*, i–iii. Ljubljana: Univerza v Ljubljani, Fakulteta za sociologijo, politične vede in novinarstvo: Center za proučevanje sodelovanja z deželami v razvoju, 1972.

Pelicon, Ivo. "Summary Assessment and Certain Problems of Yugoslavia's Economic Cooperation with Developing Countries in the Period from 1966 to 1967 [Sumarna ocena in nekateri problemi gospodarskega sodelovanja Jugoslavije z deželami v razvoju v času od 1966 do 1967]." In *Yugoslavia's Economic Relations with Developing Countries:* Conferende Materials. Ljubljana, June 1968, 1–24. Ljubljana: School of Political Science, 1968.

Pierzyńska, Justyna. "Collective self-reliance: A portrait of a Yugoslav development strategy." *Miscellanea Geographica Sciendo* 16, No. 2 (2012): 30–35. https://doi.org/10.2478/v10288-012-0024-3.

Research Centre for Cooperation with Developing Countries Ljubljana-Yugoslavia. Ljubljana: Center za proučevanje sodelovanja z deželami v razvoju, 1981.

Rogelj, Janez. *Scientific and Technical Cooperation of the Socialist Republic of Slovenia with Developing Countries in the Period 1976–1979/1980 [Znanstveno-tehnično sodelovanje SR Slovenije z deželami v razvoju v obdobju 1976 – 1979/1980]*, 1–4. Ljubljana: Center za sodelovanje proučevanja z deželami v razvoju Ljubljana, 1981.

Terček, Janez. "Editorial." In *Bulletin [Bilten]*. Ljubljana: Center za proučevanje sodelovanja z deželami v razvoju 1, No. 1 (1970): 3–7.

Vuksanović, Miloš. "Technical cooperation as an important element of our relations with developing countries [Tehnično sodelovanje kot pomemben element naših odnosov z deželami v razvoju]." In *Yugoslavia's Economic Relations with Developing Countries:* Conference Materials, Ljubljana, June 1968, 89–113. Ljubljana: School of Political Science, 1968.

Yugoslavia and the Economic Development of Developing Countries: Symposium Papers, Ljubljana, 23–24 June 1966. Ljubljana: School of Political Science; Chamber of Commerce of Slovenia, 1966.

Žarić, Mirko. "Basic Assumptions and Dilemmas of Action-Oriented Measures and a More Organized Role of Non-Alignment on the International Level [Osnovne pretpostavke i dileme akcionog dejstva i organizirovanije uloge nesvrstanosti na medjunarodnom planu]." In *Non-alignment and Yugoslavia [Nesvrstanost i Jugoslavija]*, edited by Mirko Žarić and Dragoslav Pejić, 51–79. Ljubljana: Univerza v Ljubljani, Fakulteta za sociologijo, politične vede in novinarstvo: Center za proučevanje sodelovanja z deželami v razvoju, 1972.

Tjaša Konovšek

CENTER ZA PROUČEVANJE IN SODELOVANJE JUGOSLAVIJE Z DRŽAVAMI V RAZVOJU (1966–1973): NASTANEK, DELOVANJE IN TRANSFORMACIJA

POVZETEK

Center za proučevanje in sodelovanje Jugoslavije z državami v razvoju je nastal leta 1966 v Ljubljani. Njegova ustanovitev spada v kontekst hladne vojne, dekolonizacije in vzpona Gibanja neuvrščenih (NAM), ki ga je Jugoslavija soustvarjala kot eden ključnih akterjev. Prispevek raziskuje vlogo centra kot veznega člena med znanstveno produkcijo, zunanjepolitičnim delovanjem in razvojnim sodelovanjem z državami globalnega juga, hkrati pa se umešča v pristop raziskovanja »neuvrščenosti od spodaj« – razumevanje gibanja, ki ne izhaja zgolj iz politične elite, temveč tudi iz institucionaliziranih praks, znanstvene produkcije in sodelovanja na terenu. Center je nastal neposredno po konferenci Jugoslavija in gospodarski razvoj dežel v razvoju, ki se je leta 1966 odvila v Ljubljani. Združila je strokovnjake s področij sociologije, ekonomije in politologije, poudarek pa je bil na pravičnejših in stabilnejših gospodarskih odnosih z državami v razvoju. Udeleženci so pozvali k ustanovitvi raziskovalne institucije za trajnostno in interdisciplinarno razumevanje razmer v teh državah. Center je tako sprva nastal kot raziskovalna enota Visoke šole za politične vede, nato pa je leta 1973 postal samostojen zavod.

Deloval je na dveh področjih: izobraževalnem in raziskovalnem. Na prvem je organiziral predavanja, seminarje in usposabljanja, na drugem pa spremljal gospodarsko sodelovanje med Jugoslavijo in državami v razvoju, analiziral procese modernizacije ter pripravljal študije, priročnike in bibliografije. V povezavi z znanstveno produkcijo centra sta se kot ključna izkazala koncepta solidarnosti in razvoja. Solidarnost je pogosto pomenila ideološko-politično načelo, ki je presegalo ekonomske interese in temeljilo na skupnih zgodovinskih izkušnjah kolonializma in boja za suverenost, medtem ko je razvoj razumljen kot večdimenzionalen proces tehnološkega napredka, gospodarske neodvisnosti in politične avtonomije. Center je pri tem poudarjal odgovornost Zahoda za globalne neenakosti ter pomen tehnološkega prenosa kot poti k odpravi tehnološkega kolonializma in vzpostavitvi miru. Pri tem so posebno vlogo imeli zbiranje, posredovanje in uporaba informacij. V ta namen je center v okviru svojega delovanja vzpostavil informacijski sistem INDOK, v sklopu katerega so sodelavci centra zbirali, obdelovali in posredovali podatke o političnih in ekonomskih razmerah v državah v razvoju. Po letu 1973 je center doživel preobrazbo. Postal je samostojna raziskovalna ustanova z zveznim pomenom. Slednje je omogočalo razširitev delovanja, hkrati pa tudi tesnejše povezave z zunanjepolitičnimi cilji Jugoslavije. Center ni bil zgolj raziskovalna ustanova, temveč tudi politični akter v mednarodnem prostoru. Z analitičnimi prispevki in poročili je vplival na oblikovanje jugoslovanske politike do držav v razvoju ter prispeval k uresničevanju koncepta socialistične modernosti kot alternative zahodnim modelom razvoja.

Solidarnost in razvoj sta v kontekstu centra delovala kot osrednja konceptualna okvira za povezovanje v okviru NAM ter artikulacijo globalne prihodnosti. Center je bil pomemben akter v produkciji znanja. Predstavljal je prostor, kjer so se prepletale raziskovalne, politične in diplomatske prakse socialistične Jugoslavije. Za globlje razumevanje njegovih institucionalnih povezav ter sodelovanja z drugimi akterji s področja znanosti, gospodarstva in politike je potrebno nadaljnje raziskovanje, ki bo pripomoglo k boljšemu razumevanju jugoslovanske globalne politike in vloge znanstvenih ustanov v njej.

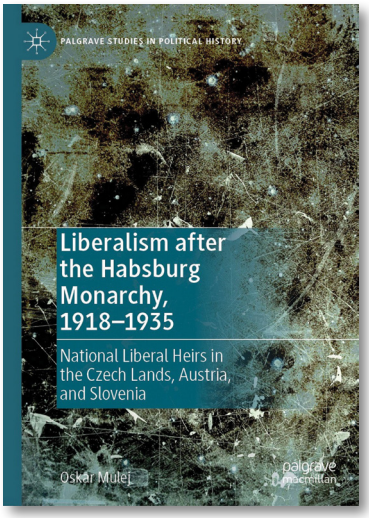
Jelka Piškurić

NA ROBU LJUBLJANSKEGA BARJA

Ocene in poročila – Reviews and Reports

Oskar Mulej, Liberalism after the Habsburg Monarchy, 1918–1935: National Liberal Heirs in the Czech Lands, Austria, and Slovenia.

Cham: Palgrave Macmillan, 2024, 368 str.



Knjiga Oskarja Muleja *Liberalism after the Habsburg Monarchy, 1918–1935* prinaša poglobljeno in primerjalno zasnovano analizo liberalnih političnih tradicij v treh nekdanjih habsburških prostorih – v čeških deželah, Avstriji in na Slovenskem. Gre za obsežno delo, ki združuje politično zgodovino, zgodovino idej, diskurzivno analizo in primerjalno metodologijo, ob tem pa odpira eno ključnih vprašanj evropske politike med obema vojnama – kaj je v teh razmerah sploh pomenilo biti »liberalec«.

Avtor knjige, dr. Oskar Mulej, je zgodovinar in raziskovalec, zaposlen na Univerzi na Dunaju, kjer se ukvarja s politično zgodovino srednjeevropskega prostora. Doktoriral je na Srednjeevropski univerzi v Budimpešti. Pri svojem delu redno prepleta slovenski, avstrijski in širši srednjeevropski kontekst. Njegovo raziskovalno zanimanje obsega nacionalizme, liberalizem ter spreminjanje in analiziranje političnih ideologij v 19. in 20. stoletju. Pričujoča monografija je rezultat večletnega sistematičnega dela in predstavlja pomemben prispevek tako k domačemu kot mednarodnemu zgodovinopisju.

Že v uvodu avtor opozori na temeljno protislovje, ki ga delo raziskuje: kako to, da so številne politične stranke, ki jih danes v zgodovinopisju označujemo kot liberalne, same sebe redko imele za takšne ali pa sploh nikoli. Včasih so se od te oznake celo odmikale. Avtor s tem zastavi zapleteno vprašanje o politični identiteti in spominu. Kot uvodno izhodišče uporabi izjavo češkega liberalnega publicista Ferdinanda Peroutka (1895–1978) iz leta 1923, da je liberalizem po prvi svetovni vojni »brez moči«.

To izhodišče primerja s sodobnimi vprašanji o krizi liberalizma in naraščajočem vplivu neliberalnih političnih usmeritev v srednji in vzhodni Evropi.

Mulej uvede ključen analitični pojem »nacionalno-liberalni dediči« (*national liberal heirs*), s katerim označuje politične akterje, ki so izšli iz liberalnih strank 19. stoletja, a so se v medvojnem obdobju od teh ideoloških izhodišč pogosto oddaljili. Ti dediči so po njegovem pogosto ohranili organizacijsko strukturo in družbeno bazo liberalnih predhodnikov, vendar so vse bolj prevzemali elemente nacionalizma, konservativizma in včasih celo avtoritarizma. Avtor tako opozarja, da je njihova liberalnost neredko obstajala bolj v očeh sodobnikov ali zgodovinarjev kot v njihovem dejanskem političnem delovanju.

V knjigi avtor analizira tri osrednje primere: češko *Československo národní demokracijo* (ČsND), avstrijsko *Grossdeutsche Volkspartei* (GdVP) in slovenske liberalce, bolj ali manj združene v *Jugoslovanski demokratski stranki* (JDS) ter njenih naslednicah, zlasti *Samostojni demokratski stranki* (SDS) in *Jugoslovanski nacionalni stranki* (JNS). Vse tri primere odlikuje raznolika mešanica stare liberalne tradicije in novih političnih okoliščin, v katerih so bile prisiljene delovati med množično demokracijo, naraščajočim populizmom in gospodarsko krizo.

V poglavju o semantiki liberalizma avtor prek analize političnega jezika pokaže, kako je pomen pojma liberalizem v medvojnem obdobju razpadal in se radikalno preoblikoval. V češkem, avstrijskem in slovenskem prostoru je bil pojem pogosto nekoliko ironično uporabljen ali pa so ga politični nasprotniki izkoriščali kot negativno oznako. V Avstriji je bilo še posebej značilno, da so se nekateri nacionalni liberalci, zlasti iz stranke GdVP, javno opredeljevali do protiklerikalizma in nemškega nacionalizma, pa tudi do antisemitizma, kar kaže na to, da oznaka »liberalen« ni več pomenila tistega, kar danes običajno razumemo pod tem pojmom.

V nadaljevanju avtor podrobno obravnava razvoj nacionalno-liberalnih tradicij v drugi polovici 19. stoletja in ugotavlja, da je do pomembnega ideološkega preobrata prišlo že pred letom 1918. Nacionalni liberalizem je v srednjeevropskem prostoru v veliki meri prepletel načela narodne emancipacije s cilji politične modernizacije. Vendar pa se je že proti koncu stoletja začel oddaljevati od klasičnega liberalnega programa svobode posameznega človeka. V mnogih primerih je liberalizem postal predvsem izraz nacionalne buržoazije in njenega boja proti starim družbenim ureditvam ter cerkvi, ne pa nujno proti avtoritarnosti sami.

Avtor uporabi bolj uravnotežen pristop. Namesto da bi ocenjeval, ali so omenjene stranke »prave« liberalne stranke, jih proučuje kot produkt določenih zgodovinskih okoliščin in ideoloških premikov. To velja tudi za njihov gospodarski program, ki ga obravnava v poglavju o vlogi države v obdobju gospodarske krize. V njem ugotavlja, da so politične stranke pogosto zagovarjale razširjeno vlogo države v gospodarstvu, kar se zdi v nasprotju s klasičnim liberalizmom, a je to mogoče razumeti kot odziv na realne potrebe časa, še posebej po gospodarski krizi leta 1929. Še pomembnejše pa je, da se nekateri predstavniki omenjenih političnih strank niso zavzemali za odprto avtoritarno oblast, ampak so pogosto poskušali ohraniti osnovna načela parlamentarizma in pravne države.

Ena od pomembnih prednosti knjige je tudi širši pogled, ki presega splošne okvire strankarske politike. V zadnjem poglavju, z naslovom *A Glimpse Beyond Party Politics*, Mulej analizira tudi liberalne intelektualce in civilnodružbene ideje in pobude, ki so v medvojnem obdobju razvijali druge oblike liberalizma, bolj v obliki socialnega liberalizma ter skozi kritično distanco do že obstoječe politike. S tem pokaže, da liberalizem ni bil le strankarski pojav, temveč tudi del širšega kulturnega sveta.

Knjiga je pregledno razdeljena na sedem poglavij, uvod in zaključek. Prvi dve poglavji uvajata osnovne pojme in zgodovinski okvir, tretje obravnava politične pomene pojma »liberalizem« v različnih kontekstih, četrto poglavje podrobno analizira izbrane politične stranke, peto se osredotoča na nacionalistične pristope analiziranih strank, šesto na gospodarske in ustavne poglede, sedmo pa predstavi intelektualno ozadje in neodvisne liberalne pobude. Takšna struktura omogoča preglednost in bralcu daje možnost, da se poglobi v posamezne vidike v skladu z lastnim zanimanjem.

Avtor v svojem delu večkrat opozori na povezave med medvojnim obdobjem in sodobnostjo. Po njegovem mnenju je vzpon neliberalnih oblik demokracije v sodobni vzhodni Evropi, kot primer izpostavi Viktorja Orbana iz Madžarske, mogoče razumeti tudi kot nadaljevanje zgodovinskih procesov, ki jih obravnava njegovo delo. Ta perspektiva mu daje dodatno aktualnost in odpira možnosti za nadaljnja raziskovanja.

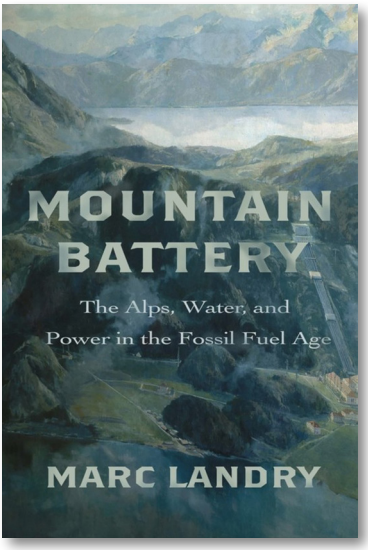
Kljub visoki znanstveni ravni je knjiga napisana dostopno in berljivo. Mulej pogosto razlaga uporabljene strokovne izraze, uvaja jasne konceptualne razmejitve (denimo med liberalizmom kot filozofijo, kot ideologijo in kot politično prakso) in s tem omogoča razumevanje tudi bralcem, ki se z zgodovino političnih idej šele seznanjajo. Knjiga sicer ne ponudi enotne ocene o usodi liberalizma v tem obdobju, kar pa je najverjetneje tudi njen glavni namen. Delo odpira odprt razmislek o obravnavani temi in ne skuša podati dokončne sodbe.

Knjiga Oskarja Muleja je dragocen prispevek k razumevanju politične zgodovine srednje Evrope. Vključitev slovenskega primera v primerjalno analizo je še posebej dobrodošla in prispeva k večji vidnosti slovenske zgodovine v mednarodni znanstveni skupnosti. Delo je gotovo zanimivo za poznavalce tematike kot tudi za mlajše raziskovalce, ki iščejo uvod v kompleksne politične preobrazbe po razpadu habsburške monarhije.

Tamara Logar

Marc Landry, Mountain Battery: The Alps, Water, and Power in the Fossil Fuel Age.

Stanford, California: Stanford University Press, 2025, 314 str.



Leta 1889 so si obiskovalci svetovne razstave v Parizu (*Exposition Universelle*) med drugim lahko ogledali razstavni predmet, naslovljen BELI PREMOG (HOUILLE BLANCHE). Šlo je za reliefni zemljevid predela francoskih Alp, nad katerim je bila postavljena turbina, ki ga je za svetovno razstavo pripravil francoski inženir Aristide Bergès. S tem primerom Marc Landry začenja svojo najnovejšo monografijo *Mountain Battery: The Alps, Water, and Power in the Fossil Fuel Age.*

Prek uvoda, sedmih poglavij in zaključka avtor bralcem pripoveduje zgodbo o gorah, vodi in energiji v Evropi od sredine 19. stoletja do danes. Gre za študijo o zgodovini alpske hidroenergije, v kateri se sprašuje po razlogih, ki so privedli do tega, da se je hidroenergija na začetku 20. stoletja izkazala za glavno alternativo premogu v proizvodnji električne energije. Sočasno pa se sprašuje tudi, zakaj in kako so ravno Alpe postale eno od najpomembnejših središč za proizvodnjo električne energije v Evropi, s čimer bralcem razkrije mnogokrat neizpostavljene delčke zgodovine držav alpskega loka (z izjemo Slovenije, ki ni vključena v monografijo).

Rezultat izkoriščanja Alp za proizvodnjo električne energije vidi v nastanku gigantskega sistema za shranjevanje vodne energije, ki ga poimenuje *evropska baterija* (*Europe's battery*). Kot argumentira, so Alpe s tem postale hibridna pokrajina, v kateri je meja med naravo in tehnologijo zabrisana. Ne gre le za zgodovinsko študijo energetskega vira, ampak za zgodovino alpske energetske pokrajine in njenega preoblikovanja pod vplivom človeka. Landry se ob tem zlasti posveča izumiteljem in tehnološkim inovacijam, dejanjem, mislim ter domišljiji politikov in inženirjev, pa tudi resnici, ki se je skrivala za herojsko opevanimi prizadevanji in načrti o preoblikovanju alpskega vodnega gospodarstva. Za prikaz te zapletenosti posega tudi po primerih na lokalni ravni. Ne nazadnje opozori še na nezaželene posledice človeškega preoblikovanja alpske pokrajine.

V prvem poglavju se avtor posveti začetkom dojemanja alpskih vodotokov kot belega premoga. Ob tem poudari dejstvo, da pri belem premogu ne govorimo o novem viru energije, saj so se izbrani vodotoki že stoletja pred tem izkoriščali s pomočjo tradicionalnih tehnologij. Govorimo namreč le o spremembi načina izkoriščanja vodne energije, do česar je pripeljalo iskanje novih virov energije, ki bi lahko konkurirali premogu. Kot pokaže, je prehod na fosilna goriva namreč sprožil ponovno ovrednotenje

potenciala vodne energije v Alpah. Ker so bile tedaj, kljub vpeljanemu tradicionalnemu izkoriščanju vodne energije s pomočjo vodnih koles, velike količine vodne energije še neizkoriščene, se je začelo zavzemanje za izrabo teh skritih moči alpskih vodotokov. To je imelo za posledico izum novega pogonskega stroja – turbine –, ki je omogočil dojemanje Alp kot nove energetske pokrajine.

V nadaljevanju poglavja Landry pogled usmeri v francoske Alpe in najzaslužnejšega inženirja za širjenje ideje o belem premogu – Aristida Bergèsa. Za zaključek preide k iztočnici za poglavje, ki sledi – kako vodno energijo prenesti v mesta in industrijska središča.

Drugo poglavje obravnava tesno povezavo med alpsko hidroenergijo in elektrifikacijo. To povezavo Landry ponazori s pomočjo življenja inženirjev, ki so svoje kariere posvetili razvoju nove tehnologije, znane kot izmenični tok. Mednje sodi tudi bavarski inženir Oskar von Miller, ki je leta 1882 na mednarodni električni razstavi v Münchnu opozoril na pomen elektrike za prenos vodne energije, leta 1891 pa na električni razstavi v Frankfurtu s sodelavci prvič demonstriral prenos izbrane energije. Dogodku, ki je danes razumljen kot ključni preobrat v zgodovini oskrbe z električno energijo, je v nadaljnih desetletjih sledila sprememba odnosa do alpske vodne energije, z dramatičnim povečanjem števila posegov v alpsko hidrologijo.

V nadaljevanju Landry pozornost posveti še vizijam in prepričanjem, ki so vzniknila v navezavi na beli premog, ter vprašanju pravice do razpolaganja z njim. Kot ugotavlja, so se v osemdesetih letih 19. stoletja v Alpah, zlasti v Švici, Avstriji, Franciji in nemški deželi Baden, pojavila gibanja za nacionalizacijo vodnih virov.

Legenda o jezeru Walchensee na Bavarskem predstavlja uvod tretjega poglavja, namenjenega Alpam kot pokrajini s potencialom za shranjevanje energije. Na prelomu iz 19. v 20. stoletje so strokovnjaki začeli pozornost posvečati alpskim jezerom in predelom v gorah, ki bi vodo kot vir električne energije naredili bolj podobno premogu. Govora je torej o shranjevanju vodne energije v jezerih, kar Landry predstavi na primeru jezera Walchensee. Sočasno v zgodbo vpelje še načrte za elektrifikacijo železnice, ki so med inženirji pravzaprav spodbudili zavzemanje za shranjevanje energije v jezerih. Del poglavja nameni tudi odporu proti načrtovanim projektom, ki bi privedli do sprememb okolja in posledičnih družbenih pretresov.

Landry ugotavlja, da so Alpe med letoma 1914 in 1918 predstavljale gospodarsko fronto, v kateri je ključno vlogo igral ravno beli premog. Države zahodne in srednje Evrope so namreč energetske potrebe v vojnem času poskušale zadovoljiti z alternativnimi viri energije. Za primer tokrat vzame francoske Alpe in razvojno pot hidroenergije, ki jo je sprožila vojna. Hkrati opozori na ključni pomen proizvedene energije za gospodarstvo v vojnih letih in po njej. Spregovori tudi o okoljskih razmerah, ki so gore naredile težavne za delavce, in splošno o problematiki delovne sile zaradi mobilizacije. Kot pokaže v nadaljevanju, so vojna in njene posledice omogočile tudi izvedbo projektov, katerih usoda je bila pred letom 1914 še negotova.

S koncem prve svetovne vojne in v obdobju, ki je sledilo, so se začeli sprejemati še bolj drastični ukrepi za odpravo domnevnih pomanjkljivosti belega premoga.

Landry ugotavlja, da je izkoriščanje gora zaradi njihove zmožnosti shranjevanja vodne energije v medvojnem obdobju postalo pravi trend, in to obdobje označi kot ero gradnje jezov v Alpah. Vse skupaj je spremljalo vprašanje prenosa belega premoga daleč prek meja samih gora, ob sočasni razpravi o izkoriščanju alpske energije v duhu mednarodnega sodelovanja. Temu se je že ob koncu dvajsetih let 20. stoletja pridružila konkurenčna želja po izkoriščanju alpskih vodotokov za doseganje gospodarske samozadostnosti, ki je bila zlasti prisotna v Italiji in Nemčiji. Med drugim avtor pozornost nameni Južni Tirolski, katere vodotoki so do izbruha druge svetovne vojne proizvedli stokrat več električne energije kot leta 1918, ter izgradnji omrežja z visoko napetostjo, ki je povezovalo industrijska središča Zahodne Nemčije z Alpami na zahodu Avstrije.

V šestem poglavju se avtor posveti predvsem daljši zgodovini prizadevanj nacionalsocialistov za vključitev avstrijske vodne energije v nemško oskrbo z električno energijo, s čimer podaja alternativno zgodovino priključitve Avstrije Nemčiji. Pozornost je ob tem namenjena projektu Kaprun v predelu Visoke Ture. Prek takšnih zavzemanj, pogajanj in izvedb projektov pred in med drugo svetovno vojno opozarja na vidike konflikta, ki so do sedaj ostali brez širše pozornosti.

Že v letih pred začetkom druge svetovne vojne se je v Evropi oblikovalo mnenje, da je treba Alpe izkoristiti zlasti za shranjevanje vodne energije, potrebne za razvijajoče se električno omrežje. Dokončanje te evropske gorske baterije pa je, kot ugotavlja Landry, s koncem vojne postalo del hladne vojne. Na primeru avstrijske gorske pokrajine opiše razmere, ki so spodbudile povojno gradnjo jezov v Alpah. Med drugim sta poudarjena tudi Marshallov plan in njegov vpliv na realizacijo načrtov za preoblikovanje alpske pokrajine in nadaljno elektrifikacijo. Obravnava še nevarnosti, ki so spremljale pridobivanje energije iz visokogorskih predelov, ter nekatere najbolj dramatične posledice izkoriščanja alpske energije. V zvezi s tem se posveti zlasti Vajontu, ki velja za eno najhujših katastrof, do sedaj povezanih z jezovi.

Zaključno poglavje je postavljeno v sodobnost in vključuje razmisleke o položaju nove alpske energetske pokrajine v sodobnem svetu ter o posledicah preusmerjanja vodne energije gora. Temu doda trditev, da Alpe še vedno igrajo ključno vlogo v evropski oskrbi z električno energijo, zlasti kot pokrajina, polna jezov, namenjena shranjevanju energije. Z osvetlitvijo vplivov na spreminjanje Alp in organizacije, kot je CIPRA (Commission Internationale pour la Protection des Régions Alpines), pozornost nameni še okoljskim posledicam izgradnje te nove pokrajine. Monografijo zaključi s podpoglavjem, naslovljenim *Prihodnost gorske baterije* (*The Future of the Mountain Battery*), v katerem problematizira pomen alpskih vodotokov v luči globalnega segrevanja ter političnih in znanstvenih razprav.

Sara Šifrar Krajnik

Satoshi Murayama, Žarko Lazarević in Aleksander Panjek (ur.), Changing Living Spaces: Subsistence and Sustenance in Eurasian Economies from Early Modern Times to the Present.

Slovenska znanstvena zbirka za humanistiko 12. Koper: University of Primorska Press, 2024, 283 str., ilustr.

Pričujoča znanstvena monografija je izšla leta 2024 kot plod dela v okviru notranjega raziskovalnega programa »ŽIVS – Življenjski prostori Slovenije: preteklost – sedanjost – prihodnost« na Univerzi na Primorskem. Uredniške dolžnosti za monografijo si delijo zgodovinarji Satoshi Murayama, Žarko Lazarević in Aleksander Panjek, sestoji pa se iz teoretično-metodološke uvodne razprave in enajstih študij primera. Slednje so uredniki razdelili v tri tematske sklope.

Zgodovinar Satoshi Murayama nas v poglobljenem, tako rekoč programatskem uvodnem sestavku seznanja tako z občimi raziskovalnimi cilji monografije kot tudi s teoretičnim ozadjem in osnovnimi izhodišči koncepta življenjskih prostorov (*living spaces*). Kot osnovni namen monografije opredeljuje kolektivno diskusijo o razvojnih stopnjah, ki vodijo od poljedelske družbe v industrializirano in končno tudi poindustrijsko družbo. Kot je razvidno iz naslova dela, je osnovni geografski okvir za monografijo sicer Evrazija, po Murayamovih besedah pa naj bi raziskovalni kolektiv zanimali predvsem tisti lokalni konteksti, kjer prevladujejo drobne posestniške strukture (7). Koncept življenjskih prostorov opredeljuje kot »holističen pristop, ki je antitetičen pristopu analitičnega členjenja« in naj bi zajemal deset glavnih tematskih sklopov. Med te spadajo živali, rastline, mikroorganizmi, vode, zrak, kopno, nesreče, hrana, odpadki in človeštvo (15). Kot v svoji študiji primera zapišeta Aleksander Panjek in Gregor Kovačič, velja koncept življenjskih prostorov razumeti predvsem kot »interdisciplinarni pristop, ki v dolgotrajnem in holističnem smislu obravnava odnose med človekom in okoljem in povezuje preteklost s sedanjostjo, potencialno pa tudi s prihodnostjo« (255).

Če vzamemo v ozir raziskovalni *credo* in geografsko koncepcijo monografije, nas ne bo presenečalo dejstvo, da njen analitični del obsega izrazito raznolik nabor študij, ki izhodiščni analitični koncept preizkušajo v zelo različnih geografskih in časovnih okvirih. Prvi tematski sklop v monografiji se ukvarja z družbenimi in gospodarskimi konteksti (*Social and Economic Contexts*) ter obsega štiri razprave. Pri teh velja kot rdečo nit poudariti predvsem vprašanje posestniških struktur in njihovega vpliva na gospodarski

razvoj v danem lokalnem kontekstu. Prva razprava (avtorjev Luce Mocarellija in Paola Tedeschija) nas tako seznanja s privatizacijo srenjske zemlje v alpskih dolinah Lombardije v 19. stoletju. Avtorja kot svoj glavni prispevek izpostavljata uvid v negativne ekološke posledice, ki jih je povzročilo neodgovorno upravljanje z naravnimi viri že na samem začetku industrializacije, ter v spremenljive, potencialno tudi negativne vplive procesa privatizacije na lokalne podeželske skupnosti. Razmeroma preglednejše narave je prispevek zgodovinarja in urednika monografije Žarka Lazarevića, ki se osredotoča na strukture kmečkega gospodarstva v Sloveniji med obema vojnama. Bralke in bralce seznanja z razdrobljeno lastniško strukturo kmečke posesti, njeno nizko donosnostjo in družbenimi posledicami, ki so jih omenjeni dejavniki pustili na podeželju v obravnavanem obdobju. Ti so se kazali predvsem v visoki (pogosto dejansko »skriti«) nezaposlenosti, prenaseljenosti in končno tudi v intenzivnih (sezonskih in stalnih) migracijah. Kmečko gospodarstvo je prav tako predmet razprave avtorice Haruhise Asada, ki pa omenjeno problematiko obravnava v sodobnosti – znotraj indijske zvezne države Assam. Glavni predmet prispevka so značilnosti samooskrbnega kmetijstva pri različnih etničnih skupnostih v dolini Brahmaputra, v ospredju avtoričinega interesa pa se nahajajo predvsem ekološki konteksti in tehnološki vidiki kmetijske proizvodnje. Zadnji prispevek v pričujočem tematskem sklopu (avtorja Josefa Grulicha) se naposled podrobneje posveča fenomenu migracij na primeru selitev s podeželja v Češke Budjejovice v drugi polovici 18. stoletja. Kot ključni sklep svoje študije poudarja dejstvo, da so migracije v danem kontekstu sicer zajele vse sloje podeželskega prebivalstva, še najpogosteje pa so se selili posamezniki, ki jim zakonodaja ni zagotavljala deleža pri dedovanju družinske kmečke posesti.

Drugi tematski sklop monografije je posvečen uporabi virov (*Utilization of Resources*) in obsega štiri razprave. Prva med njimi (avtorja Tara Takemota) obravnava izkoriščanje trave in lesa na cesarski skupni zemlji ter njeno postopno pretvarjanje v novo kategorijo zaščitenega gozda (*conservation forest*) v japonski prefekturi Yamanashi na začetku 20. stoletja. Razprava tako nudi vpogled v zgodnje poskuse zakonodajne regulacije izkoriščanja gozdov in lokalnih naravnih virov na Japonskem, katere namen je bil predvsem preprečevanje poplavljanja. V japonski prostor je umeščena tudi naslednja študija (avtorja Miyukija Takahashija), ki se dotika vprašanja razprostranjenosti uporabe konj v gospodarstvu zgodnjesrednjeveške Japonske. Avtor na osnovi empiričnih podatkov iz prefektur Asaka in Katsushika navaja raznovrstne namene, za katere je lokalno prebivalstvo vzrejalo in izkoriščalo konje, ter tako nudi intervencijo v širšo razpravo o spreminjajoči se vlogi človeške in živalske delovne sile v japonskem kmečkem gospodarstvu v omenjenem obdobju. Tretja študija v pričujočem sklopu (avtorjev Laitpharlanga Cajeeja in Monice Mawlong) nas znova popelje v sodobni indijski kontekst. Predmet razprave je potencialni prispevek tradicionalnega lončarstva k trajnostnemu razvoju v severovzhodu države, avtorja pa se v analizi posvečata predvsem okoljskim dejavnikom, ki opredeljujejo značaj omenjene obrti. Sklop naposled zaključuje razprava (avtorice Noriko Yuzawa), ki obravnava cirkulacijo gnojil

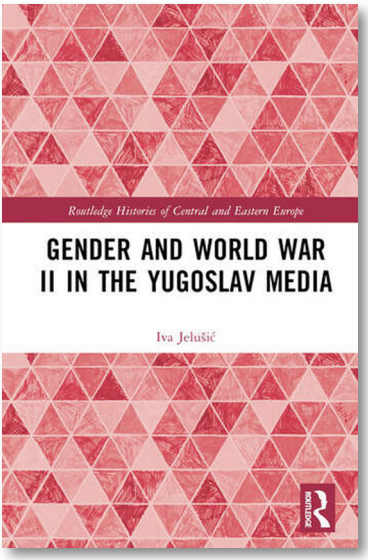
na Japonskem v zadnjih dveh stoletjih. Študija primera se osredotoča na spremembe v uporabi človeškega blata (*night soil*) kot gnojila v prefekturi Aichi v času rastoče urbanizacije in industrializacije.

Zadnji tematski sklop v monografiji obsega tri daljše razprave, ukvarja pa se z »naravnimi spremenljivkami« (*Natural Variables*), ki opredeljujejo življenjske prostore. Prva študija v sklopu (avtorja Masanorija Takashime) obravnava zgodovino pridelave riža na Japonskem v širokem časovnem loku, ki sega od osmega stoletja do začetkov industrializacije v devetnajstem stoletju. Raziskava temelji predvsem na kvantitativnih podatkih in daje vpogled tako v naravne kot tudi v politične dejavnike, ki so vplivali na razvoj omenjene kulture v danem časovnem okviru. Drugi prispevek v sklopu (avtorjev Satoshija Murayame, Hiroka Nakamure, Noboruja Higashija in Toruja Teraoja) je prav tako lokaliziran v Japonsko, ukvarja pa se s poljedelskimi krizami, ki so jih povzročile poplave, suše in pomanjkanje sončne svetlobe na območju t. i. Vzhodnoazijskega monsuna. Sklop, pa tudi celotno monografijo, naposled zaključuje študija (avtorjev Aleksandra Panjeka in Gorazda Kovačiča), ki nas vrača v slovenski geografski kontekst in na osnovi izrazito lokalne historiografske dileme obravnava širše vprašanje človekovega upravljanja z vodami skozi zgodovino. Da bi ugotovila, kako velja razumeti načrtno izsuševanje manjšega jezera v okolici Štanjela glede na siceršnje pregovorno pomanjkanje vode na območju Krasa, sta avtorja proizvedla impresivno in izrazito interdisciplinarno mikroštudijo, ki omenjeno problematiko analizira v štiristoletnem časovnem loku.

Čeprav se posamezne študije primera znotraj monografije lahko na prvi pogled zazdijo izrazito medsebojno oddaljene tako v zemljepisnem, kulturnem kot tudi kronološkem oziru, jih Murayamov koncept življenjskih prostorov ter iz njega izhajajoča široka metodološka zasnova monografije vendarle povezujeta v smiselno celoto. Metodološko univerzalnost pa velja poudariti kot še dodaten razlog, zakaj si pozornost vsekakor zaslužijo tudi študije o bralki oziroma bralcu morebiti manj znanih področjih in epohah – skozi njihovo prebiranje se namreč seznanjamo z metodološkimi pristopi in orodji, ki bi jih s pridom lahko uporabili tudi v raziskavah, ki se dotikajo slovenskega ali tudi poljubnega drugega kulturnozgodovinskega konteksta. Monografijo zato velja priporočiti predvsem bralkam in bralcem s splošnim teoretično-metodološkim interesom za okoljsko, družbeno in gospodarsko zgodovino, prispevka izpod peres slovenskih avtorjev pa nudita tudi podrobnejše vpoglede v okoljsko in gospodarsko zgodovino slovenskega Krasa ter slovenskega podeželja nasploh.

Oliver Pejić

Iva Jelušić, Gender and World War II in the Yugoslav Media.
London: Routledge, cop. 2025, 230 str.



Eden od pogostejših sloganov komunističnih aktivistk v Jugoslaviji je, da so si ženske svoje pravice izborile s participacijo v vojni. Ko je govora o ženskah v vojni, se pred nami izriše množica podob: podoba pogumne vojakinje, matere, ki je v vojni izgubila sina, zdravnice ali negovalke, pa tudi skrbnice. Prav s tem raznolikim nizom podob in vprašanjem, ali se je spomin na ta boj za pravice zares ohranil, se sooča avtorica Iva Jelušić v delu *Gender and World War II in the Yugoslav Media* (*Spol in druga svetovna vojna v jugoslovanskih medijih*). Z analizo tega, kako so se v popularnih revijah med letoma 1945 in 1980 spominjali, si predstavljali in tudi zamolčali podobe ženskih partizanskih bork ter, širše, lik »nove ženske«, Jelušić osvetljuje ambivalentne in pogosto protislovne načine, na katere so se socialistični mediji ukvarjali z brezprimernim sodelovanjem žensk v vojni. Knjiga prepričljivo pokaže, da emancipacija žensk v Jugoslaviji ni bila niti enostaven dosežek niti zgodba o neizogibnem nazadovanju, temveč dinamičen in neprekinjen proces, ki so ga enako oblikovale politične direktive, uredniške odločitve in vsakodnevne kulturne prakse.

Iva Jelušić, doktorica primerjalnega zgodovinopisja s Srednjeevropske univerze (Central European University), je knjigo zasnovala na svojem doktorskem delu. Monografija izstopa po izredno natančni analizi revij in bogati metodološki zasnovi. Avtorica združuje arhivsko raziskovanje, dokumentarne vire in spominsko literaturo s podrobnim branjem revij, pri čemer analizira ne le besedilne vsebine, temveč tudi vizualno podobo, retorične strategije in implicitna pričakovanja bralcev. Tako razkriva večplastne in pogosto protislovne načine ustvarjanja in rabe spomina, povezanega s spolnimi vlogami v socialistični Jugoslaviji. Knjiga je razdeljena na tri smiselno zaokrožene dele, ki se med seboj elegantno povezujejo.

V prvem delu avtorica oriše historiat položaja ženske v socialistični Jugoslaviji in lik »nove ženske«, predstavi razvoj jugoslovanskih medijev ter metodološke pristope, uporabljene pri analizi. Osrednje mesto v obravnavi zavzema partizanka, ki je kot simbol vstopa žensk v moško domeno bojevanja utelešala obljubo emancipacije v revolucionarnem trenutku. Vendar Jelušić prepričljivo pokaže, da ženska vojakinja nikoli ni bila enakovredna moškemu borcu. Čeprav je bila njena vloga v povojnih letih slavljenja izpostavljena, je bila hitro ponovno umeščena v tradicionalno paradigmo matere, medicinske sestre ali vestne aktivistke. Ti arhetipi so se prepletali in krepili,

ustvarjajoč repertoar podob, ki so priznavale prispevek žensk k vojni, a hkrati omejevale njihove povojne možnosti.

Jedro knjige predstavljata drugi in tretji del z analizo štirih revij. Drugi del je posvečen dvema hrvaškima ženskima revijama, ki sta bili kljub sočasnemu izhajanju izrazito različni. Revija *Žena u borbi* (*Ženska v boju*), ki so jo ustvarile komunistične aktivistke in urejale vidne članice državne ženske organizacije, je imela pomembno vlogo pri oblikovanju kanona ženske emancipacije neposredno po vojni. Poudarjala je sodelovanje žensk v narodnoosvobodilnem boju kot domoljubno dolžnost in revolucionarni dosežek, a hkrati skrbno oblikovala podobo vojakinje, tako da je njeno junaštvo uravnotežila s pričakovanji skromnosti in predanosti družini. Rezultat je bila paradoksalna figura emancipirane ženske, ki je bila hkrati moderna in tradicionalna, revolucionarna in domača. Ta protislovja nazorno ponazarjajo dvojni značaj socialistične spolne politike, ki je skušala povzdigniti ženske, ne da bi ji resnično uspelo razgraditi uveljavljene hierarhije in odpraviti patriarhalni sistem.

Povsem drugačno mesto v kulturnem spominu je imela modna in življenjska revija *Svijet* (*Svet*). Na prvi pogled zasnovana po vzoru zahodnih modnih revij, polna glamurja in potrošniških podob, je občasno vendarle objavljala prispevke, značilne za državni ženski tisk: poročila o mednarodnem dnevu žensk, obletnicah Antifašistične fronte žensk in portrete posameznih junakinj odporniškega gibanja. Toda ti prispevki so bili prej izjema kot pravilo. Žensk v vojni so se sicer lotevali v dveh kolumnah: prva je bila o ženskah in špijonaži, kjer so bile ženske pogosto prikazane v seksualiziranih vlogah, druga pa je bila rubrika o ulicah, poimenovanih po ženskah, ki bi skoraj bolj sodila v *Ženo u borbi*. Prav ta dvojnost je ključna: s prepletanjem emancipacijskih pripovedi in potrošniškega okvira je *Svijet* odražal kulturno hibridnost jugoslovanskega socializma, ki je združeval ideološke zaveze s težnjo po modernem življenjskem slogu.

Takšna analiza ponuja svež pogled na to, kako so ženske v socialistični Jugoslaviji usklajevale večplastne in pogosto protislovne identitete. Bralke so bile nagovorjene kot dedinje partizanskih junakinj, kot udeleženke socialistične modernizacije in kot sodobne potrošnice. Ta hibridnost je spodbijala staro uveljavljeno predstavo o linearnem nazadovanju emancipacije, ki naj bi doživela vrhunec v prvih povojnih letih: spomin na odpor žensk je ostal, čeprav v preoblikovani, hibridni obliki, ki je odražala položaj Jugoslavije med Vzhodom in Zahodom, socializmom in potrošništvom.

V tretjem delu knjige se avtorica posveti revijama *Arena* in *Start*, ki sta izhajali pri isti založbi kot *Svijet* (*Vjesnik*). *Arena*, družinska ilustrirana tedenska revija s številčnim ženskim bralstvom, je v vsaki številki vključevala zgodovinska besedila in veliko jih je bilo o drugi svetovni vojni. Le majhen del je obravnaval partizanke in ženske, vključene v vojno, pogost pa je bil lik partizanske matere in žene. Ta narativ je poudarjal materinsko vztrajnost in žrtvovanje, kar je sodelovanje žensk prestavilo iz sfere boja v pripoved o družini. Kot primer kolumne lahko uporabimo *To je moj život*, ki je začela izhajati v sedemdesetih. V kolumni o življenjskih zgodbah ljudi je izšlo osem zgodb žensk, ki so bile dejavne v drugi svetovni vojni in so predstavljale raznolike

vloge, od zdravnic, ilegalk do aktivistk. Le ena je bila vojakinja. Ta premik ni v celoti izbrisal pomena žensk, je pa njihove prispevke interpretiral v skladu s tradicionalnimi spolnimi normami.

Start, prva jugoslovanska moška revija, je bil namenjen predvsem spektaklu potrošnje in seksualiziranim podobam. Toda paradoksalno je odpiral prostor za razpravo o emancipaciji, enakosti spolov in celo neofeminizmu, saj je kar precej novih feministk v poznih sedemdesetih in osemdesetih pisalo prav za *Start*. Čeprav partizank v *Startu* skoraj ni bilo, so se teme emancipacije pojavile v povsem novih in presenetljivih kontekstih. Revija je obenem objektivizirala ženske in ponujala forum za razmislek o spolnih normah.

Analizo obeh revij avtorica prepleta z razmislekom o podobnih prispevkih v obeh ženskih revijah in tako jasno pokaže različnost v obravnavi iste tematike. V nekaterih primerih lahko opazujemo celo intervjuje z istimi ženskami, a v zelo različnih kontekstih. Analiza *Arene* in *Starta* je dragocena dopolnitev, saj pokaže, da se kulturni spomin ni oblikoval le v ženskih revijah, temveč je prežemal tudi žanrsko drugačne in komercialne publikacije. S tem Jelušić jasno pokaže, da se pomen partizanske izkušnje žensk ni izgubil, temveč je bil nenehno reinterpretiran v različnih medijskih kontekstih, pogosto v nasprotnih tonih sentimentalnosti, potrošništva in seksualizacije.

Dosežek knjige ni le v empirični bogatosti, temveč tudi v konceptualni jasnosti. Avtorica revije obravnava hkrati kot besedila in kot kulturne institucije ter pokaže, kako se spomin ustvarja skozi uredniške odločitve, retorične strategije, vizualne podobe in interpretativne prakse bralcev. Razvoj uredniških politik opazuje tudi skozi prizmo političnih sprememb v Jugoslaviji, kar je ključno za razumevanje širše politične slike.

Sklenem lahko, da je knjiga Ive Jelušić pomemben prispevek k raziskovanju kulturnega spomina in spolov v državnem socializmu. Zapleta uveljavljene zgodbe o emancipaciji žensk, poudarja vlogo medijev pri oblikovanju kolektivnega spomina in jugoslovanski primer umešča v regionalno in transnacionalno perspektivo. S tem, ko v ospredje postavlja ženske revije, hkrati pa analizira tudi širši komercialni tisk, avtorica odpira nove poti raziskovanja in jasno pokaže, da je kulturni spomin vedno predmet pogajanj, izpodbijanja in preoblikovanja skozi čas.

Nesa Vrečer


56 ZBIRKA
RAZPOZNAVANJA
RECOGNITIONES
Nataša Henig Miščič
CARNIOLAN SAVINGS BANK
Shaping Financial Landscape in Slovenian Territory

UDC/UDK	94(497.4)„18/19„
ISSN	0353-0329 (tiskana izdaja)
	2463-7807 (spletna izdaja); https://ojs.inz.si/pnz
DOI	https://doi.org/10.51663/pnz.65.3

INŠTITUT
ZA NOVEJŠO ZGODOVINO