Zbornik konference

Jezikovne tehnologije in digitalna humanistika

Proceedings of the Conference on

Language Technologies and Digital Humanities

19.–20. september 2024

Ljubljana, Slovenija

19–20 September 2024

Ljubljana, Slovenia

Uredila / *Edited by*:

Špela Arhar Holdt, Tomaž Erjavec

ZBORNIK KONFERENCE
JEZIKOVNE TEHNOLOGIJE IN DIGITALNA HUMANISTIKA

PROCEEDINGS OF THE CONFERENCE ON
LANGUAGE TECHNOLOGIES AND DIGITAL HUMANITIES

Uredila / *Edited by*: Špela Arhar Holdt, Tomaž Erjavec
Tehnični uredniki / *Technical editors*: Ida Gnidovec,
Jakob Lenardič, Tina Munda, Mihael Ojsteršek

Založil / *Published by*:
Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Izdal / *Issued by*:
Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Za založbo / *For the publisher*:
Andrej Pančur
Direktor / *Director*

Ljubljana, 2024
Prva izdaja / *First edition*

Spletno mesto konference / Conference website:
https://www.sdjt.si/wp/jtdh-2024 / https://www.sdjt.si/wp/jtdh-2024-en/

Publikacija je brezplačno dostopna na / *Publication is available free of charge at*:
https://zenodo.org/records/13912515

CIP - Kataložni zapis o publikaciji

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani

COBISS.SI-ID 211315971

ISBN 978-961-7104-40-0 (PDF)

PREDGOVOR K ZBORNIKU KONFERENCE »JEZIKOVNE TEHNOLOGIJE IN DIGITALNA HUMANISTIKA«

Letošnja konferenca »Jezikovne tehnologije in digitalna humanistika« je štirinajsta v seriji, ki na dve leti poteka že od leta 1998, in peta po vrsti, odkar se je dogodek tematsko razširil, da poleg jezikovnih tehnologij vključuje tudi digitalno humanistiko. Kot v preteklih letih je konferenca pritegnila številne predstavnike in predstavnice raziskovalne, razvojne, študentske in širše skupnosti – tako slovenske kot mednarodne – ki so v dveh dneh predstavljali in spoznavali najnovejše področne raziskave, izsledke, aktivnosti, pa tudi izzive in težave, ki so še pred nami.

Konferenca se tradicionalno odvija s podporo Slovenskega društva za jezikovne tehnologije, Centra za jezikovne vire in tehnologije Univerze v Ljubljani ter raziskovalnih infrastruktur CLARIN.SI in DARIAH-SI. V letu 2024 je organizacijski odbor konference upravljal ZRC SAZU, organizacija pa je potekala v sodelovanju s Fakulteto za elektrotehniko Univerze v Ljubljani, ki je dogodek gostila 19. in 20. septembra 2024.

Poleg slovenskih so na konferenci sodelovali avtorji in avtorice iz držav, kot so Hrvaška, Srbija, Makedonija, Italija, Belgija in Švica. Jezika konference sta bili slovenščina in angleščina. Skupno je bilo predstavljenih 35 prispevkov, vsakega od teh so recenzirali po trije znanstveni recenzenti, člani in članice mednarodnega programskega odbora konference. Izpostaviti gre študentsko sekcijo, ki je mladim na začetku raziskovalne poti zagotovila posebej skrbne recenzije, možnost predstavitve pred naklonjeno strokovno javnostjo in študentsko nagrado za najboljši dosežek.

Konferenca je gostovala vabljeni predavanji, kjer sta uveljavljena raziskovalca iz tujine pregledno osvetlila dva ključna aktualna trenda področja. Simon Dobnik z Univerze v Göteborgu je v svojem predavanju Beyond Pixels and Words (»Onkraj pikslov in besed«) raziskoval preplet med slikovnimi in jezikovnimi podatki ter njuno uporabo v jezikovnih tehnologijah. Barbara McGillivray z univerze King's College London pa je v predavanju Exploring language change computationally: lessons from interdisciplinary collaborations (»Računalniško raziskovanje jezikovnih sprememb: Spoznanja na osnovi interdisciplinarnih sodelovanj«) predstavila interdisciplinarni pristop k računalniškemu proučevanju jezikovnih sprememb.

Program konference je dopolnil zanimiv mednarodni strokovni panel, na katerem so Simon Dobnik (Univerza v Göteborgu), Tanja Samardžić (Univerza v Zürichu), Nikola Ljubešić (Inštitut "Jožef Stefan"), Andrej Žgank (Univerza v Mariboru), Danila Zuljan Kumar (ZRC SAZU) in Kaja Dobrovoljc (Univerza v Ljubljani) z moderiranjem Darinke Verdonik (Univerza v Mariboru) razpravljali o napredku in perspektivah v raziskavah govorne komunikacije.

Pričujoči zbornik vsebuje recenzirane prispevke z raznoliko vsebino. Bogato zastopane so zlasti predstavitve novih virov in orodij za raziskave govorjenega jezika, parlamentarnega govora, uporabniško generiranih vsebin, zgodovinskih virov in pisanja učečih se. Kot v preteklih letih so izpostavljene teme, povezane z jezikoslovnim označevanjem in procesiranjem jezikovnih podatkov, nekoliko novejše pa teme, vezane na gradnjo jezikovnih modelov in preverbo uporabnosti generativnih tehnologij za različne strokovne naloge.

Urednika se zahvaljujeva vsem, ki ste prispevali k uspehu konference in izidu konferenčnega zbornika – avtorjem in avtoricam prispevkov, vabljenima predavateljema, sodelujočim na panelu, članom in članicam programskega in organizacijskega odbora, vodjem sekcij, tehničnim urednikom in urednicam in vsem drugim sodelujočim.

Upava, da bo bralcem in bralkam zbornik prinesel nove vpoglede in raziskovalne spodbude na področju jezikovnih tehnologij in digitalne humanistike, in se veseliva prihodnje konference JT-DH!

Ljubljana, september 2024

Urednika zbornika, Špela Arhar Holdt in Tomaž Erjavec

PREFACE TO THE PROCEEDINGS OF THE CONFERENCE "LANGUAGE TECHNOLOGIES AND DIGITAL HUMANITIES"

This year's "Language Technologies and Digital Humanities" conference marks the fourteenth in a series that has taken place every two years since 1998, and the fifth since the event expanded to include digital humanities alongside language technologies. As in previous years, the conference attracted numerous representatives from research, development, student, and broader communities—both Slovenian and international—who gathered over two days to present and explore the latest research, findings, activities, as well as upcoming challenges in the field.

The conference was held with the support of the Slovenian Language Technologies Society, the Centre for Language Resources and Technologies at the University of Ljubljana, and the research infrastructures CLARIN.SI and DARIAH-SI. In 2024, the organizational committee was led by ZRC SAZU, in cooperation with the Faculty of Electrical Engineering at the University of Ljubljana, which hosted the event on 19 and 20 September 2024.

In addition to Slovenian participants, authors from Croatia, Serbia, Macedonia, Italy, Belgium, and Switzerland also took part. The conference languages were Slovenian and English. A total of 35 papers were presented, each reviewed by three scientific reviewers, members of the international program committee. Of particular note was the student section, which provided young researchers with especially thorough reviews, the opportunity to present to a constructive professional audience, and an award for the best student achievement.

The conference featured keynote lectures, where two established researchers highlighted current trends in the field. In his lecture Beyond Pixels and Words, Simon Dobnik from the University of Gothenburg explored the interplay between visual and linguistic data and their application in language technologies. Barbara McGillivray from King's College London, in her lecture Exploring Language Change Computationally: Lessons from Interdisciplinary Collaborations, presented an interdisciplinary approach to the computational study of language change.

The program was enriched by an international expert panel, where Simon Dobnik (University of Gothenburg), Tanja Samardžić (University of Zurich), Nikola Ljubešić (Jožef Stefan Institute), Andrej Žgank (University of Maribor), Danila Zuljan Kumar (ZRC SAZU), and Kaja Dobrovoljc (University of Ljubljana), moderated by Darinka Verdonik (University of Maribor), discussed progress and perspectives in speech communication research.

These proceedings contain peer-reviewed papers on diverse topics. Notable are presentations of new resources and tools for research on spoken language, parliamentary speech, user-generated content, historical sources, and learner writing. As in previous years, topics related to linguistic annotation and language data processing are prominent, with a growing focus on building language models and evaluating the potential of generative technologies to enrich research methodologies.

We want to thank all who contributed to the success of the conference and the publication of these proceedings—authors, keynote speakers, panel participants, program and organizational committee members, session chairs, technical editors, and all other contributors.

We hope that the proceedings provide readers with new insights and inspire further research in the fields of language technologies and digital humanities. We look forward to the next JT-DH conference!

Ljubljana, September 2024

Editors of the Proceedings, Špela Arhar Holdt and Tomaž Erjavec

PROGRAMSKI ODBOR / *PROGRAMME COMMITEE*

PREDSEDSTVO PROGRAMSKEGA ODBORA / *STEERING COMMITTEE*

Špela Arhar Holdt, predsednica / *Chair*, CJVT
Filozofska fakulteta, Univerza v Ljubljani in Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Arts, University of Ljubljana and Faculty of Computer and Information Science, University of Ljubljana*

Slavko Žitnik, SDJT
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer and Information Science, University of Ljubljana*

Tomaž Erjavec, CLARIN.SI
Odsek za tehnologije znanja, Institut "Jožef Stefan" / *Department of Knowledge Technologies, Jožef Stefan Institute*

Jakob Lenardič, DARIAH-SI
Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Matej Klemen, študentska sekcija / *student section*
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer and Information Science, University of Ljubljana*

Tina Munda, študentska sekcija / *student section*
Filozofska fakulteta, Univerza v Ljubljani Ljubljani / *Faculty of Arts, University of Ljubljana*

David Bordon, študentska sekcija / *student section*
Filozofska fakulteta, Univerza v Ljubljani Ljubljani / *Faculty of Arts, University of Ljubljana*

ČLANI PROGRAMSKEGA ODBORA IN RECENZENTI / PROGRAMME COMMITTEE MEMBERS AND REVIEWERS

Saša Babič
Inštitut za slovensko narodopisje, ZRC SAZU / *Institute of Slovenian Ethnology, ZRC SAZU*

Petra Bago
Filozofska fakulteta, Univerza v Zagrebu / *Faculty of Humanities and Social Sciences, University of Zagreb*

Vuk Batanović
Inovacijski center Elektrotehniške fakultete, Univerza v Beogradu / *Innovation Center of the School of Electrical Engineering in Belgrade*

Narvika Bovcon
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer and Information Science, University of Ljubljana*

Václav Cvrček
Inštitut češkega narodnega korpusa, Karlova univerza v Pragi / *Institute of the Czech National Corpus, Charles University in Prague*

Jaka Čibej
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer and Information Science, University of Ljubljana*

Simon Dobrišek
Fakulteta za elektrotehniko, Univerza v Ljubljani / *Faculty of Electrical Engineering, University of Ljubljana*

Helena Dobrovoljc
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU / *Fran Ramovš Institute of the Slovenian Language, ZRC SAZU*

Kaja Dobrovoljc
Filozofska fakulteta, Univerza v Ljubljani Ljubljani / *Faculty of Arts, University of Ljubljana*

Jerneja Fridl
Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti / *Research Centre of the Slovenian Academy of Sciences and Arts*

Polona Gantar
Filozofska fakulteta, Univerza v Ljubljani Ljubljani / *Faculty of Arts, University of Ljubljana*

Vojko Gorjanc
Filozofska fakulteta, Univerza v Ljubljani Ljubljani / *Faculty of Arts, University of Ljubljana*

Jurij Hadalin
Inštitut za novejšo zgodovino / *Institute of Contemporary History* Inštitut

Ivo Ipšić
Univerza na Reki / *University of Rijeka*

Mateja Jemec Tomazin
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU / *Fran Ramovš Institute of the Slovenian Language, ZRC SAZU*

Alenka Kavčič
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer and Information Science, University of Ljubljana*

Iztok Kosem
Filozofska fakulteta, Univerza v Ljubljani Ljubljani / *Faculty of Arts, University of Ljubljana*

Simon Krek
Filozofska fakulteta, Univerza v Ljubljani in Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Arts, University of Ljubljana and Faculty of Computer and Information Science, University of Ljubljana*

Drago Kunej
Glasbenonarodopisni inštitut, ZRC SAZU / *Institut of Ethnomusicology, ZRC SAZU*

Nikola Ljubešić
Odsek za tehnologije znanja, Institut “Jožef Stefan” / *Department of Knowledge Technologies, Jožef Stefan Institute*

Nataša Logar
Fakulteta za družbene vede, Univerza v Ljubljani / *Faculty of Social Sciences, University of Ljubljana*

Matija Marolt

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer and Information Science, University of Ljubljana*

Sanda Martinčić Ipšić

Univerza na Reki / *University of Rijeka*

Mirjam Sepesy Maučec

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru / *Faculty of Electrical Engineering and Computer Science, University of Maribor*

Maja Miličević Petrović

Univerza v Bolonji / *University of Bologna*

Dunja Mladenić

Laboratorij za umetno inteligenco, Institut "Jožef Stefan" / *Artificial Intelligence Laboratory, Jožef Stefan Institute*

Andrej Pančur

Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Matevž Pesek

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer and Information Science, University of Ljubljana*

Karmen Pižorn

Pedagoška fakulteta, Univerza v Ljubljani / *Faculty of Education, University of Ljubljana*

Senja Pollak

Odsek za tehnologije znanja, Institut "Jožef Stefan" / *Department of Knowledge Technologies Jožef Stefan Institute*

Ajda Pretnar

Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Marko Robnik-Šikonja

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer and Information Science, University of Ljubljana*

Tanja Samardžić

Univerza v Zürichu / *University of Zurich*

Miha Seručnik
Zgodovinski inštitut Milka Kosa, ZRC SAZU / *Milko Kos Historical Institute, ZRC SAZU*

Marko Stabej
Filozofska fakulteta, Univerza v Ljubljani Ljubljani / *Faculty of Arts, University of Ljubljana*

Janez Štebe
Fakulteta za družbene vede, Univerza v Ljubljani / *Faculty of Social Sciences, University of Ljubljana*

Mojca Šorn
Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Daniel Vasić
Univerza v Mostarju / *University of Mostar*

Darinka Verdonik
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru / *Faculty of Electrical Engineering and Computer Science, University of Maribor*

Jerneja Žganec Gros
Alpineon d.o.o. / *Alpineon d.o.o., Slovenia*

Andrej Žgank
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru / *Faculty of Electrical Engineering and Computer Science, University of Maribor*

Aleš Žagar
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer and Information Science, University of Ljubljana*

Branko Žitko
Fakulteta za znanost, Univeza v Splitu / *Faculty of Science, University of Split*

ORGANIZACIJSKI ODBOR / *ORGANISING COMMITEE*

Jerneja Fridl, predsednica / Chair, DARIAH-SI
Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti / *Research Centre of the Slovenian Academy of Sciences and Arts*

Mojca Šorn, DARIAH-SI
Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Ana Cvek, DARIAH-SI
Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Simon Dobrišek, CJVT
Fakulteta za elektrotehniko, Univerza v Ljubljani / *Faculty of Electrical Engineering, University of Ljubljana*

Katja Meden, CLARIN.SI
Institut "Jožef Stefan" / *Jožef Stefan Institute*

Kaja Dobrovoljc, SDJT
Filozofska fakulteta, Univerza v Ljubljani Ljubljani / *Faculty of Arts, University of Ljubljana*

Miha Peče, DARIAH-SI
Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti / *Research centre of the Slovenian Academy of Sciences and Arts*

Miha Seručnik, DARIAH-SI
Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti / *Research centre of the Slovenian Academy of Sciences and Arts*

ORGANIZATORJI / *ORGANISERS*

SPONZORJI / *SPONSORS*

URNIK / *TIMETABLE*

ČETRTEK / *THURSDAY*, 19 SEPTEMBER 2024

FAKULTETA ZA ELEKTROTEHNIKO / *FACULTY OF ELECTRICAL ENGINEERING*

08:00–09:00	**Registracija / *Registration***
Lokacija / *Venue*	**Predavalnica P2 / *Lecture Hall P2***
09:00–09:30	**Otvoritev / *Opening***
09:30–10:30	**Vabljeno predavanje / *Keynote* [ENG]** Vodja sekcije / *Session Chair*: Slavko Žitnik **Simon Dobnik (University of Gothenburg):** *Beyond Pixels and Words* [Abstract]
10:30–10:50	**Odmor za kavo / *Coffee break***
10:50–12:30	**Sekcija 1 / *Session 1*: Speech and UGC Resources [ENG]** Vodja sekcije / *Session Chair*: Darinka Verdonik

10:50–11:10	Filip Dobranić & Nikola Ljubešić: *Twitter Before X: Scenes from the Balkans*
11:10–11:30	Kristina Pahor de Maiti Tekavčič, Nikola Ljubešić & Darja Fišer: *Expanding the Frenk Corpus of Socially Unacceptable Discourse to French*
11:30–11:50	Nikola Ljubešić, Peter Rupnik & Tea Perinčić: *Mići Princ - A Little Boy Teaching Speech Technologies the Chakavian Dialect*
11:50–12:10	Kaja Dobrovoljc: *Extension and Evaluation of the Spoken Slovenian Treebank*
12:10–12:30	Tanja Samardžić, Peter Rupnik, Mirjana Starović & Nikola Ljubešić: *Mak na konac: A Multi-Reference Speech-to-text Benchmark for Croatian and Serbian*
12:30–14:00	**12:30 – Skupinsko slikanje (pred glavnim vhodom) / *Group photo (in front of the main entrance)*** **Odmor za kosilo / *Lunch break***
14:00–15:00	**Predstavitev plakatov, odmor za kavo / *Poster session with coffee break* [SLO, ENG]** Koordinacija posterske sekcije / *Poster session coordinator*: Ana Cvek
	Lenka Bajčetić, Vuk Batanović & Tanja Samardžić: *Lemmatizing Serbian and Croatian via String Edit Prediction*

	Janez Štebe: *Strojna preverba internetnih naslovov novičarskih prispevkov v naslov na Wayback Archive*
	Klara Žnideršič, Vid Klopčič, Matevž Pesek & Matija Marolt: *The GOVORI.SI Speech Transcription Platform*
15:00–16:20	**Sekcija 2 / *Session 2*: Govorni, parlamentarni viri in etika [SLO]** Vodja sekcije / *Session Chair*: Špela Arhar Holdt
15:00–15:20	Darinka Verdonik, Nikola Ljubešić, Peter Rupnik, Kaja Dobrovoljc & Jaka Čibej: *Izbor in urejanje gradiv za učni korpus govorjene slovenščine - ROG*
15:20–15:40	Katja Meden, Tomaž Erjavec & Andrej Pančur: *"Parlament je po teoriji polje kontroliranega konflikta": Slovenski parlamentarni korpus siParl 4.0*
15:40–16:00	Aleš Vaupotič & Narvika Bovcon: *Osebni podatki v umetnosti: Njihova zakonita obdelava in vloga etike v novomedijski kulturi*
16:00–16:20	Matej Martinc, Veronika Bajt, Špela Rot & Senja Pollak: *Sistem za zaznavanje sprememb v rabi besed in njegova uporaba za sociolingvistično analizo*
16:20–16:30	**Odmor / *Break***
16:30–18:00	**Panel Napredki in perspektive v raziskavah govorne komunikacije / *Panel Frontiers in Speech Communication Research* [SLO, HR, SR, ENG]**

PETEK / *FRIDAY*, 20 SEPTEMBER 2024

FAKULTETA ZA ELEKTROTEHNIKO / *FACULTY OF ELECTRICAL ENGINEERING*

08:00–09:00	**Registracija / *Registration***	
Lokacija / *Venue*	**Predavalnica P2 / *Lecture Hall P2***	**Predavalnica P6 / *Lecture Hall P6***
09:00–10:00	**Vabljeno predavanje / *Keynote* [ENG]** Vodja sekcije / *Session Chair*: Darja Fišer **Barbara McGillivray (King's College London):** *Exploring Language Change Computationally: Lessons from Interdisciplinary Collaborations* [Abstract]	
10:00-10:20	**Odmor za kavo / Coffee break**	
10:20–11:20	**Sekcija 3 / *Session 3*: Linguistic Annotation, Historic Language Data [ENG]** Vodja sekcije / *Session Chair*: Tomaž Erjavec	
10:20–10:40	Nikola Ljubešić, Luka Terčon & Kaja Dobrovoljc: *CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages*	

10:40–11:00	Katja Meden, Ana Cvek, Vid Klopčič, Matevž Pesek, Mihael Ojsteršek, Mojca Šorn & Andrej Pančur: *Unlocking History: A Redesign of the SIstory 5.0 Portal*
11:00–11:20	Alice Fedotova, Adriano Ferraresi, Maja Miličević Petrović & Alberto Barrón-Cedeño: *Expanding the European Parliament Translation and Interpreting Corpus: A Modular Pipeline for the Construction of Complex Corpora*
11:20–11:30	**Odmor / *Break***
11:30–12:30	**Sekcija 4 / *Session 4*: Development and Application of LLMs [ENG]** Vodja sekcije / Session Chair: Nikola Ljubešić
11:30–11:50	Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič & Marko Robnik-Šikonja: *Generative Model for Less-Resourced Language with 1 Billion Parameters*
11:50–12:10	Jaka Čibej: *First Steps Toward the Compilation of a Safety Dataset for Slovene Large Language Models*

12:10–12:30	Slobodan Beliga & Ivana Filipović Petrović: *Large Language Models Supporting Lexicography:* *Conceptual Organization of Croatian Idioms*
12.30–14:00	**Odmor za kosilo / *Lunch break***
14:00–15:00	**Študentska sekcija, odmor za kavo /** ***Student session, Coffee break* [SLO, ENG]** Vodja sekcije / *Session Chair*: David Bordon
14:00–14:20	Camile Lendering, Manfred González & Joaquín Figueira: *Efficient Fine-Tuning Techniques for Slovenian* *Language Models*
14:20–14:40	Luka Terčon: *Uporaba šestih mer skladenjske kompleksnosti za* *primerjavo jezika v govornem in pisnem korpusu*
14:40–15:00	Matej Klemen: *Test poznavanja splošnih besed v slovenščini med* *udeleženci Mladinske poletne šole*
15:00–15:10	**Odmor / *Break***

15:10–16:05	**Sekcija 5 / *Session 5*: Various Specialised Corpora and Studies [ENG]** Vodja sekcije / *Session Chair*: Kaja Dobrovoljc	**Sekcija 6 / *Session 6*: Različni specializirani korpusi in študije [SLO]** Vodja sekcije / *Session Chair*: Polona Gantar
15:10–15:25	Anna Kryvenko: *Degrees of Belonging to Europe in Parliamentary Discourse: A Comparative Corpus-Assisted Study*	Mojca Stritar Kučuk: *KOST 2.0: Predstavitev korpusa in potek označevanja jezikovnih napak*
15:25–15:45	Ajda Pretnar Žagar: *A Corpus Linguistic Characterisation of sPeriodika*	Jaka Čibej & Tina Munda: *Metoda polavtomatskega popravljanja lem in oblikoskladenjskih oznak na primeru učnega korpusa govorjene slovenščine ROG*
15:45–16:05	Jakob Lenardič: *Slovenian particle: Not a Syntactic Category*	Diana Košir & Tomaž Erjavec: Korpus CVET 1.0: *Izdelava, opis in analiza zbirke starejših besedil v verski periodiki*
16:05–16:25	**Zaključek / *Closing***	
16:30–18:00	**Občni zbor SDJT / *SDJT Annual Meeting***	

KAZALO / *TABLE OF CONTENTS*

VABLJENI PRISPEVKI / *INVITED TALKS*

PRISPEVKI / *PAPERS*

RAZŠIRJENI POVZETKI / *EXTENDED ABSTRACTS*

ŠTUDENTSKI PRISPEVKI / *STUDENT PAPERS*

"BEYOND PIXELS AND WORDS"

Simon DOBNIK
University of Gothenburg
Box 200, 40530 Göteborg
simon.dobnik@gu.se

ABSTRACT

Words are not used in isolation. When we communicate we relate them to our background knowledge, the intent of interaction – what is the purpose of what we want to say, who is our partner, what has been said before – our common ground, our senses and perception of the physical world and situations around us. Speech is also not the only way to convey information with: we interact in writing, symbols, with different kinds of texts, with eye-gaze, gestures and other properties of our bodies. Language models in language technology extract meaning primarily from text and sometimes a few other modalities such as images and acoustic signal. This poses two questions: (i) to what extent can these modalities be a proxy for representing semantic knowledge for different natural language processing tasks and applications; and (ii) how can we port semantic knowledge captured in these modalities to different modalities – how can we bring large language models to the real world and take them for a walk? In this talk I will describe our research towards answering these questions and outline our challenges awaiting ahead.

BIO

Simon Dobnik is a professor of computational linguistics at the Department of Philosophy, Linguistics and Theory of Science (FLoV) at University of Gothenburg, Sweden. He is a member of the Centre for Language Technology (CLT) and the Centre for Linguistic Theory and Studies in Probability (CLASP) where he leads the Cognitive systems research group. He has worked on (i) data models and machine learning of meaning representations for language, action and perception, (ii) semantic models for language, action and perception (computational semantics), (iii) representation learning in language, inference and interpretability, (iv) interpretation and generation of spatial descriptions and reference, (v) interactive learning with small data, (vi) data bias and privacy, and (vii) multimodal dialogue, robotics and related topics.

IZVLEČEK PREDAVANJA

Besed ne uporabljamo v izolaciji. Ko govorimo, jih povezujemo s predhodnim znanjem, z namenom interakcije – s kakšnim namenom govorimo, kdo je naš sogovornik, kaj je bilo povedano prej – našim skupnim razumevanjem, s čuti in zaznavanjem fizičnega sveta in situacij okrog nas. Govor pa ni edini način, s katerim prenašamo informacije: interakcije potekajo v pisanju, simbolih, z različnimi vrstami besedil, s pogledom, gestami in drugimi lastnostmi naših teles. Jezikovni modeli v jezikovnih tehnologijah pomen primarno luščijo iz besedil, občasno pa tudi iz drugih modalnosti, kot so slike in akustični signal. To nas postavlja pred dve vprašanji: (i) v kolikšni meri te modalnosti predstavljajo in posredujejo semantično znanje za različne naloge in uporabe obdelave naravnega jezika in (ii) kako semantično znanje iz teh modalnosti prenesti v druge modalnosti – kako bi velike jezikovne modele pripeljali v resnični svet in jih peljali na sprehod? Na predavanju bom predstavil raziskave, s katerimi naslavljamo ta vprašanja, in orisal izzive, ki nas čakajo v prihodnosti.

O PREDAVATELJU

Simon Dobnik je profesor računalniškega jezikoslovja na oddelku za filozofijo, jezikoslovje in teorijo znanosti na univerzi Göteborg na Švedskem. Deluje kot član centra Centre for Language Technology (CLT) in član centra Centre for Linguistic Theory and Studies in Probability (CLASP), kjer vodi skupino za kognitivne sisteme. Njegove raziskave vključujejo: (i) podatkovne modele in strojno učenje pomenskih reprezentacij za jezik, delovanje in zaznavanje, (ii) semantične modele jezika, delovanja in zaznavanja (računalniška semantika), (iii) učenje reprezentacij v jeziku, sklepanju in razložljivosti, (iv) interpretacijo in generiranje prostorskih opisov in referenc, (v) interaktivno učenje z majhnimi količinami podatkov, (vi) pristranskost in varnost podatkov, ter (vii) multimodalni dialog, robotiko in s tem povezane teme.

EXPLORING LANGUAGE CHANGE COMPUTATIONALLY: LESSONS FROM INTERDISCIPLINARY COLLABORATIONS

Barbara McGILLIVRAY
King's College London, Department of Digital Humanities
Strand Campus, Strand, London, WC2R 2LS
barbara.mcgillivray@kcl.ac.uk

ABSTRACT

Advanced computational methods allow us to analyse vast datasets and uncover previously inaccessible patterns. However, few natural language processing algorithms properly account for the dynamic nature of language, particularly its semantics, which is crucial to humanistic inquiry. Efforts are underway to improve AI systems' understanding of historical context and language dynamics, such as in the automatic detection of semantic change, but human annotation and interpretation is still needed to capture the nuances of language and its cultural context. In this talk I will report on a collaborative project involving digital humanists, computational linguists, software engineers and library curators to analyse the effects of mechanisation on the English language of the nineteenth century. I will discuss the challenges and insights gained from combining voluntary crowdsourcing for historical language annotation with algorithms and design experiments. Integrating these approaches allows us to reach a nuanced understanding of language evolution in response to mechanization and, more broadly, contribute to interdisciplinary research at the intersection of AI and the humanities.

BIO

Barbara McGillivray is a lecturer in digital humanities and cultural computation in the Department of Digital Humanities at King's College London and a Turing fellow at the Alan Turing Institute. She is the editor-in-chief of the *Journal of Open Humanities Data* and the convenor of the MA programme Digital Humanities at King's. She also serves as the president of the ACL Special Interest Group on Language Technologies for the Socio-Economic Sciences and Humanities, as well as the convenor of the Turing special interest group Humanities and Data Science. Her research focuses on computational methods for studying language change in both historical languages and contemporary data. She has been a co-investigator of the Living with Machines project, a large collaboration between the Alan Turing Institute and the British Library, aimed at investigating the effects of mechanization through the analysis of British historical newspaper collections. Her most recent book is *Applying Language Technology in Humanities Research: Design, Application, and the Underlying Logic* (co-authored with Gábor Mihály Tóth, Palgrave Macmillan, 2020).

IZVLEČEK PREDAVANJA

Napredne računalniške metode nam omogočajo analizo ogromnih količin podatkov in odkrivanje prej nezaznavnih vzorcev. Le malo algoritmov za obdelavo jezika pa ustrezno upošteva dinamično naravo jezika, predvsem semantiko, ki je ključna za humanistične raziskave. Potekajo prizadevanja, da bi sistemi umetne inteligence bolje razumeli zgodovinski kontekst in dinamiko jezika, na primer pri samodejnem zaznavanju semantičnih premikov, vendar za zaznavanje finih odtenkov jezika in kulturnega konteksta še vedno potrebujemo človeško označevanje in interpretacijo. Na predavanju bom poročala o interdisciplinarnem projektu, v katerem smo za analizo učinkov mehanizacije na razvoj angleškega jezika v 19. stoletju sodelovali digitalni humanisti, računalniški jezikoslovci, programerji in knjižničarji. Razpravljala bom o izzivih in spoznanjih, ki jih je prineslo združevanje množičenja za označevanje zgodovinskega jezika, algoritmov in oblikovalskih eksperimentov. Integracija omenjenih pristopov nam namreč omogoča večplastno razumevanje evolucije jezika ob prihodu mehanizacije ter v širšem prispeva k interdisciplinarnim raziskavam na stičišču umetne inteligence in humanističnih ved.

O PREDAVATELJICI

Barbara McGillivray is predavateljica digitalne humanistike in računalniške kulturologije na oddelku za digitalno humanistiko King's College London in raziskovalka na inštitutu Alan Turing. Je glavna urednica revije *Journal of Open Humanities Data*, koordinatorka magistrskega programa Digitalna humanistika na King's College London, predsednica posebne interesne skupine za jezikovne tehnologije za socio-ekonomske vede in humanistiko pri združenju ACL ter koordinatorka posebne interesne skupine Humanistika in podatkovna znanost na inštitutu Alan Turing. Njene raziskave se osredotočajo na računalniške metode za raziskovanje jezikovnih sprememb tako v zgodovinskem jeziku kot v sodobnih podatkih. Bila je med glavnimi raziskovalci projekta Living with Machines (Živeti s stroji), obsežnega sodelovanja med inštitutom Alan Turing in narodno knjižnico British Library, pri katerem so skozi analizo britanskih zbirk zgodovinskega časopisja raziskovali učinke mehanizacije. Njena najnovejša knjiga je *Applying Language Technology in Humanities Research. Design, Application, and the Underlying Logic* (soavtor Gábor Mihály Tóth, Palgrave Macmillan 2020).

LEMMATIZING SERBIAN AND CROATIAN VIA STRING EDIT PREDICTION

Lenka BAJČETIĆ[1,2], Vuk BATANOVIĆ[2], Tanja SAMARDŽIĆ[3]
[1]School of Electrical Engineering, University of Belgrade
[2]Innovation Center of the School of Electrical Engineering, University of Belgrade
[3]University of Zurich

In this paper, we examine the effectiveness of lemmatizing texts in Serbian and Croatian using a pre-trained large language model fine-tuned on the task of string edit prediction. We define lemmatization as a tagging task, where each word-lemma transformation is represented as a string edit tag which encodes the necessary prefix and suffix alterations. Our approach is verified using the BERTić large language model and leads to improved results on the standard Serbian SETimes.SR and the standard hr500k Croatian dataset, as well as on ReLDI-NormTagNER-sr and ReLDI-NormTagNER-hr datasets. Its additional advantage is that it does not rely on any lexical databases, making it easily applicable to different text domains and language variants.

Keywords: Lemmatization, BERTić, SETimes.SR, hr500k, ReLDI-NormTagNER-sr, ReLDI-NormTagNER-hr

1 INTRODUCTION

Lemmatization entails assigning to each word form its base form (e.g., ’write’ → ’write’, ’writes’ → ’write’, ’wrote’ → ’write’, ’written’ → ’write’). It used to be a crucial task in linguistic analysis and text processing, especially for highly inflected languages like Serbian and Croatian, but its role is changing in modern approaches.

The importance of lemmatization for NLP tasks when dealing with morphologically rich languages is tested by (Kutuzov & Kuzmenko, 2019), who “critically evaluate the widespread assumption that deep learning NLP models do not require lemmatized input”. They conclude that the decisions about text pre-processing before training language models should consider the linguistic nature of the language in question. As a matter of fact, lemmatization may not be necessary for English, but using lemmatized training and testing data for Russian yielded small but consistent improvements for word sense disambiguation. When it comes to Serbian, a positive impact of text lemmatization on model performances has been shown in models for sentiment analysis (Batanović & Nikolić, 2017) and semantic similarity (Batanović et al., 2018).

In morphologically rich languages, diverse sets of grammatical information are encoded within each word using inflections. Both Serbian and Croatian have seven grammatical cases, three genders, and two grammatical numbers and this information is represented through a wide variety of inflections, usually in the suffix form.

In highly inflected languages such as these, it is typical to tackle the problem of lemmatization using inflectional lexicons. These lexicons can easily become extremely large since the high number of different inflectional variants of many words dramatically increases the vocabulary size. Even if semi-automatic techniques are employed in their construction, the process of creating inflectional dictionaries is costly and time-consuming. Furthermore, the resulting lexicons are inherently limited in size and scope, especially when it comes to particular or niche domains. This leads to the issue of out-of-vocabulary words which cannot be properly processed in this approach.

An alternative to using inflectional dictionaries is to redefine lemmatization as a task of predicting sets of string edits. The generic transformation from a word to a lemma is done in four steps: 1) remove a suffix of length N_s; 2) add a new lemma suffix, L_s; 3) remove a prefix of length N_p ; 4) add a new lemma prefix, L_p. In the end, the tuple $[N_s; L_s; N_p; L_p]$ defines the word-to-lemma transformation. For example, the tuple of necessary string edits for the word 'učio' to get the lemma 'učiti' is [1, 'ti', 0, 0]. In this way, lemmatization can be understood as a sequence labeling task where each token's tag is actually the tuple which represents the set of necessary string edits. This approach is inherently more easily applicable to different text domains and language variants, as it does not rely on any lexical databases.

This technique was proven to work well both on Serbian (Gesmundo & Samardžić, 2012b) and a set of eight different languages (Gesmundo & Samardžić, 2012a), but all previous experiments were performed without deep and transfer learning, and relied on hand-crafted features instead of embeddings. In this paper, we examine the proposed lemmatization method when used in conjunction with a modern large language model. Specifically, we fine-tune a pre-trained language model for string edit sequence prediction, with each token's string edit tag being formulated based on the set of edits necessary to transform it into its lemma. For this purpose we rely on BERTić (Ljubešić & Lauc, 2021), a large language model based on the ELECTRA architecture, which was trained on more than 8 billion tokens of text in Bosnian (800 million), Croatian (5.5 billion), Montenegrin (80 million) and Serbian (2 billion).

2 RELATED WORK

The history of the task of lemmatization includes many approaches, from applying hand-written set of rules (Koskenniemi, 1984; Plisson et al., 2004) to general character-level transducers, which learn the lemmatization rules from example pairs (word form, lemma) (Dreyer et al., 2008; Nicolai & Kondrak, 2016; Eger et al., 2016). With the introduction of deep learning, character-level transducers were redefined as a case of sequence-to-sequence models and the task was solved with bi-LSTM encoder-decoder networks (Bergmanis & Goldwater, 2018; Kondratyuk et al., 2018).

Character-level transducers can be too expressive leading to over-generalization and other avoidable mistakes. A solution to this is to learn a set of edits (edit scripts or edit trees) as a single label (Chrupala, 2006; Chrupala et al., 2008; Gesmundo & Samardžić, 2012b), which makes the task of lemmatization similar to part-of-speech tagging (POS) or named entity recognition (NER). In this way, one can incorporate the information about the context in predicting lemmas, which can help avoid mistakes caused by ambiguity. Lemmatization is often solved jointly with POS or morphosyntactic tagging due to their inter-dependence (Garabík & Mitana, 2022).

Although these general solutions are elegant and reusable, their performance can be limited by irregularities, which is why it is common to devise additional filters with dictionary look-up (Jursic et al., 2010). For example, the HuSpaCy model for Hungarian contains a hybrid lemmatizer utilizing both a neural model, dictionaries and hand-crafted rules(Berkecz et al., 2023).

The CLASSLA-Stanza package is a pipeline for automatic linguistic annotation of the South Slavic languages, including Croatian and Serbian (Terčon & Ljubešić, 2023). The lemmatizer model is trained after morphosyntactic tagging is already performed, so it is utilizing both the tokens and the morphosyntactic tags, and it relies on an inflectional lexicon which serves as an additional controlling element during lemmatization. The Croatian model is trained using the hrLex 1.3 inflectional lexicon (Ljubešić, 2019a), while the Serbian model relies on srLex 1.3 inflectional lexicon (Ljubešić, 2019b).

3 METHODOLOGY

The goal of this work is to see whether defining lemmatization as a string edit prediction task will prove to be a suitable framing for a large language model to learn, thereby avoiding the reliance on an inflectional lexicon. For this purpose, we compare two lemmatization approaches:

- The baseline approach - here the large language model is first fine-tuned for the task of morphosyntactic tagging. MSD predictions obtained from the trained model are then used as input for an inflectional lexicon in order to perform word lemmatization.
- The proposed approach - here the large language model is fine-tuned for the task of predicting string edit tags which encode the transformations necessary to turn a given surface token into its lemma. Lemmatization is performed by directly applying the predicted string edits to each token.

In order to train and evaluate our models, we use four datasets:

1. The Serbian linguistic training corpus SETimes.SR 2.0 - contains around 100,000 tokens of newswire texts (Batanović et al., 2023)

2. The Croatian linguistic training corpus hr500k 2.0 - contains about 500,000 tokens of texts from different genres, including newswire, blog posts, messages from online forums, etc. (Ljubešić & Samardžić, 2023)

3. The Serbian Twitter training corpus ReLDI-NormTagNER-sr 3.0 - contains around 100,000 tokens of Twitter texts (Ljubešić et al., 2023b)

4. The Croatian Twitter training corpus ReLDI-NormTagNER-hr 3.0 - contains around 100,000 tokens of Twitter texts (Ljubešić et al., 2023a)

All four datasets have been manually annotated for a variety of NLP tasks, including morphosyntactic tagging and lemmatization, but none of them have previously been used to evaluate the proposed lemmatization approach. Our training and evaluation process was conducted in two settings: one using the predefined train-dev-test data splits in each of the datasets, and another using 10-fold cross-validation. For both approaches we evaluated model performances after multiple fine-tuning lengths, ranging from 1 to 25 epochs.

For the baseline approach we first use the train data gold MSD tags to fine-tune BERTić on the task of morphosyntactic tagging. We then use its output on the test data and the test data surface tokens to query an inflectional lexicon and obtain lemma predictions. Similarly to (Terčon & Ljubešić, 2023), we use the hrLex 1.3 inflectional lexicon for Croatian, and the srLex 1.3 lexicon for the Serbian data. The hrLex 1.3 lexicon contains 164,206 entries, while srLex 1.3 contains 169,328 entries. The lookup function is implemented to be robust, and it functions as a sieve. It first checks whether the lexicon has an entry which fits the lookup constraints completely, meaning that it has an exact match for both the token and the provided morphosyntactic tag. If this exact lookup does not yield any results, the lookup function checks whether the lexicon contains an entry with different token capitalization variants (lowercase, uppercase and all caps), and the exact match for the MSD tag. If this lookup also does not prove successful, the next step is to search for the entry which has the exact same token, but where only the part-of-speech is matched, rather than the whole morphosyntactic tag. Again, a failed lookup for the exact token is followed by trying different token capitalization variants. Finally, if none of the attempts prove fruitful, the last lookup is conducted only based on the token, disregarding the morphosyntactic tag altogether. If the token does not exist in the lexicon in any shape or form, the lemmatizer will simply assume that the lemma is the same as token (uppercased if the morphosyntactic tagger classifies it as proper noun). Additionally, there are several rules implemented to handle punctuation and abbreviations.

For the proposed new approach, the transformation tuples i.e. string edits are created for each word using the method proposed by (Gesmundo & Samardžić, 2012a). The first step is to extract the longest common substring between the token and the lemma. If the token and lemma have no common substring, the set of necessary string edits can be arbitrarily defined. For example in the case of token ’was’ and the lemma ’be’, the tuple could be [3, ’be’, 0, 0] or [0, 0, 3, ’be’]. In this case we have opted for the first possibility, so all the changes are modelled as suffix changes. If the token and its respective lemma have a common substring of at least one character, the procedure is as follows: the part of the token which comes after the longest common substring is considered as the suffix which needs to be removed for lemmatization, while the part of the lemma which comes after the longest common substring is considered as the suffix which needs to be added. The same logic is applied to define the prefix transformations. The pre-trained BERTić model is then fine-tuned on the task of per-token tag prediction, where the tags are defined by the transformation tuples (string edits).

4 DATASET ANALYSIS

It is worth noting that the number of distinct tags varies greatly between the datasets. In SETimes.SR 2.0 there are 310 tags, or different ways in which a token is transformed into its respective lemma. The number of distinct string edits is almost twice as large in hr500k 2.0, with 597 tags, which is a consequence of the Croatian dataset being five times larger than the Serbian one. However, the highest number of different tags is found in ReLDI-NormTagNER-hr 3.0 which has 2056 distinct string edits, with 1314 of them being singleton (appearing for only one token-lemma pair). Similarly, ReLDI-NormTagNER-sr 3.0 contains 1825 distinct tags, with over 1100 singletons. For comparison, SETimes.SR 2.0 has only 72 singleton tags, and hr500k 2.0 has 151. It is evident that the number of distinct string edit tags grows rapidly when working with non-standard textual data. In our experiments we treat all string edit tags equally and leave for future consideration the issue of data sparsity resulting from a large number of singleton tags in non-standard language.

Table 1 presents an overview of the most frequent string edit tags in the four datasets we considered. While there are significant differences in the number of distinct tags between the datasets, particularly singleton ones, there are only slight differences in the ten most frequent tags in each corpus.

Table 1: Most frequent string edit tags

	SETimes.SR 2.0	ReLDI-NormTagNER-sr 3.0	hr500k 2.0	ReLDI-NormTagNER-hr 3.0
1	[0, ”, 0, ”]	[0, ”, 0, ”]	[0, ”, 0, ”]	[0, ”, 0, ”]
2	[1, ”, 0, ”]	[1, ”, 0, ”]	[1, ”, 0, ”]	[1, ”, 0, ”]
3	[1, ’a’, 0, ”]	[1, ’a’, 0, ”]	[1, ’a’, 0, ”]	[1, ’a’, 0, ”]
4	[2, ’biti’, 0, ”]	[1, ’ti’, 0, ”]	[2, ’biti’, 0, ”]	[2, ’biti’, 0, ”]
5	[2, ’ti’, 0, ”]	[2, ’biti’, 0, ”]	[2, ’ti’, 0, ”]	[1, ’ti’, 0, ”]
6	[1, ’i’, 0, ”]	[0, ’ti’, 0, ”]	[1, ’i’, 0, ”]	[0, ’ti’, 0, ”]
7	[2, ”, 0, ”]	[2, ’ti’, 0, ”]	[2, ”, 0, ”]	[2, ’ti’, 0, ”]
8	[0, ’ti’, 0, ”]	[0, ’be’, 0, ”]	[0, ’ti’, 0, ”]	[1, ’i’, 0, ”]
9	[2, ’an’, 0, ”]	[1, ’i’, 0, ”]	[1, ’ti’, 0, ”]	[0, ’be’, 0, ”]
10	[0, ’be’, 0, ”]	[2, ”, 0, ”]	[0, ’be’, 0, ”]	[2, ”, 0, ”]

By far the most frequent tag is the one which indicates that nothing is to be done to the token in order to lemmatize it, i.e. the token is already in the lemma form. Most of the tags are expected because they add either infinitive suffix (’-ti’) or typical nominal suffixes like ’-a’ for nouns of feminine gender or ’-an’ for adjectives. The tag [2, ’biti’, 0, ”] covers almost all cases of lemmatization for tokens which are conjugations of the verb ’to be’ (’biti’). The only tag which might not be intuitively understood is [0, ’be’, 0, ”] because ’-be’ is not a typical suffix in Serbian or Croatian. However, this tag explains/encodes the lemmatization for token ’se’ (whose lemma is ’sebe’) and is a very frequent reflexive pronoun, making this tag quite prominent in all four datasets.

We can also notice that none of the most frequent string edits have information encoded regarding the prefix. This is not particularly surprising, since both Croatian and Serbian store most of the inflective information in the suffixes. However, a substantial portion of string edit tags do have prefix transformations encoded, with the ratio ranging between 20% in SETimes.SR (60 out of 310 distinct tags) and 40% in both ReLDI-NormTagNER-sr 3.0 and ReLDI-NormTagNER-hr 3.0 datasets. This difference in the frequency of prefix encodings is probably due to the irregularities in the Twitter data and the fact that the string edits produced for these two datasets often (accidentally) include spelling corrections and re-diacritization. A detailed breakdown of the distribution of string edit tags and their characteristics across the four datasets we consider is shown in Table 2.

Table 2: Distribution of string edit tags and their characteristics across the four datasets

	SETimes.SR 2.0	hr500k 2.0	ReLDI-NormTagNER-sr 3.0	ReLDI-NormTagNER-hr 3.0
Suffix only	249	419	1078	1226
Prefix only	3	11	180	255
Both	57	166	566	574

The predicted sets of necessary string edits are evaluated by verifying whether the token edited in the proposed way really does convert into the lemma. This verification generally confirmed that the transformations are properly produced, but in ReLDI-NormTagNER-sr and ReLDI-NormTagNER-hr there are a number of cases where applying the proposed set of string edits did not correctly create the expected lemma. This occurs in 799 tokens in the ReLDI-NormTagNER-sr dataset, and in 931 cases in the ReLDI-NormTagNER-hr data. The reason why this happens only in the non-standard data is because the original tokens here sometimes have misspellings or are written without diacritic marks, so in these cases the lemmatization process should entail token normalization as a first step. Text written incorrectly and/or without diacritics is quite common in both Serbian and Croatian web corpora, so it is important to have a strategy to deal with this issue when working with text from online sources. While fixing spelling errors is not suitable to be defined using string edits, the issue of undiacriticized text could potentially be addressed with a preprocessing step using a dedicated tool (Ljubešić et al., 2016) in cases where this is an evident problem.

In all four datasets combined there are 3391 different string edit tags. Their overlap of can be seen in Figure 1. In the diagram, we can see that the overlap of string edit tags between all four datasets is 215. Since SETimes.SR has the smallest number of distinct tags (310), we can conclude that the overlap is proportionately quite high. The highest overlap count can be found between the ReLDI-NormTagNER-hr and ReLDI-NormTagNER-sr data. This is a consequence of these two datasets both having a much higher number of different tags than the other two. We can conclude that the non-standard tokens

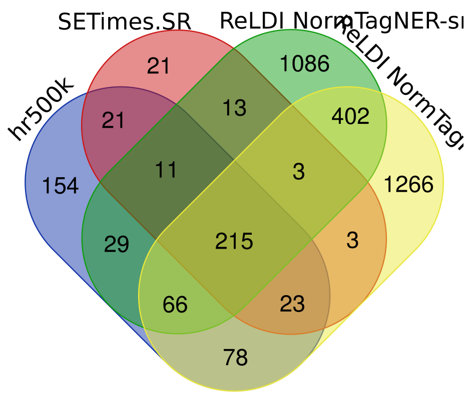


Figure 1: Venn diagram of string edit tags in the four datasets

which are encountered in tweets increase the number and variety of string edit tags far more than the size of the dataset, since hr500k which is five times larger than all the other datasets has only 154 tags that do not appear in other datasets, while ReLDI-NormTagNER-hr and ReLDI-NormTagNER-sr have 1266 and 1066 respectively.

5 EVALUATION RESULTS AND DISCUSSION

Table 3 contains the evaluation results for all four datasets. The best results in each evaluation setting are shown in boldface.

As mentioned in the previous section, the lookup function used in the baseline model is not trivial. This explains why the model based on morphosyntactic tagging and lexicon lookup performs better across all datasets for a smaller number of epochs. Basically, the errors in morphosyntactic tagging are being compensated by the robust lookup function. However, when models are fine-tuned for ten or more epochs, the approach based on string transformations noticeably outperforms the 'standard' baseline model. On no datasets does the baseline model accuracy increase by more than 1-1.5% through additional training, and these improvements can even be as low as 0.15% on hr500k. On the other hand, our proposed approach typically shows clear accuracy improvements as the number of fine-tuning epochs is increased, and only on hr500k does it reach a performance plateau after 10-15 epochs.

Table 3: Results of model evaluation

	Epochs	*10-fold CV*		*train-dev-test*		
		MSD + Lexicon	String edits	MSD + Lexicon	String edits	CLASSLA
SETimes.SR	1	95.1	84.96	94.9	81.23	98.02
	5	96.14	95.77	95.98	95.57	
	10	96.2	97.23	96.06	97.03	
	15	96.24	97.65	96.13	97.36	
	20	96.23	97.81	96.06	97.58	
	25	96.24	**97.86**	96.07	**97.76**	
hr500k	1	96.27	94.64	96.28	94.33	98.02
	5	96.41	98.1	96.48	98.16	
	10	96.4	98.38	96.5	98.54	
	15	96.4	98.43	96.5	98.58	
	20	96.4	**98.44**	96.5	**98.63**	
	25	96.41	98.43	96.5	98.62	
Reldi SR	1	85.78	76.86	86.25	76.04	94.92
	5	87.02	90.07	87.46	89.57	
	10	87.1	92.64	87.6	92.30	
	15	87.13	93.63	87.65	93.36	
	20	87.11	94.12	87.61	93.72	
	25	87.11	**94.39**	87.63	**94.06**	
Reldi HR	1	85.44	79.3	85.48	76.88	93.36
	5	86.33	90.31	86.48	89.9	
	10	86.38	92.3	86.57	91.67	
	15	86.41	92.98	86.56	92.58	
	20	86.40	93.43	86.59	93.04	
	25	86.40	**93.64**	86.64	**93.13**	

Both models perform significantly worse when trained and evaluated on Twitter data, for both Croatian and Serbian. This is, of course, due to the fact that these datasets contains non-standard language, so they can be expected to contain many out-of-vocabulary words, as well as unexpected symbols, non-standard punctuation uses and spelling errors. All of these factors have a significantly greater negative effect on the lexicon-based approach, which attains around 10% lower accuracy scores on the non-standard language datasets than on the standard ones. Conversely, the performance of the model based on string edits is only around 5% lower on the non-standard data, which indicates that this approach is more adaptable to datasets which are further from the linguistic norm. Nevertheless, the same trend is noticeable as in the previous two datasets: the approach based on morphosyntactic tagging and lexicon lookup performs better when the model is trained for a small number of epochs, but it is easily outperformed by the approach based on string edit prediction when the fine-tuning length is increased.

In order to see to what extent the 'traditional' model is affected by the lexicon we have conducted an analysis of the predictions done by the models on different datasets, and this can be seen in Table 4. We have classified the errors in three groups: out-of-vocabulary words; tokens which exist in the lexicon but whose lexicon lemmas are different from the gold standard in the datasets; and cases of ambiguity where the token exists in the lexicon with multiple lemmas, one of which is equivalent to the dataset gold standard, but the model makes a mistake by selecting a different lemma/meaning. We can see that in all the cases, the majority of issues could not be avoided with model improvement because they are lexicon related.

Table 4: Error distribution for lemmatization models relying on inflectional lexicons

	SETimes.SR	hr500k	ReldiSR	ReldiHR
Out-of-vocabulary	26%	33%	72%	59%
Lexicon issues	36%	26.5%	17%	13%
Ambiguity issues	38%	40.5%	11%	28%

Even though the results of lemmatization on all four datasets vary substantially, they all follow the same two main patterns. Firstly, extended fine-tuning of the baseline model never yields more than 1.5% accuracy improvement. This is to a certain extent a consequence of the robust lookup function, but it also indicates that the lemmatization models based on morphosyntactic tagging and inflectional lexicons have an inherent limitation in performance, likely due to the size and scope limitations of the inflectional lexicons themselves. Secondly, after a certain number of epochs the models based on predicting string edit tags outperform the lexicon-based models, despite the advanced lexicon lookup function. This indicates that defining lemmatization as a string edit prediction task in the proposed way may truly be more suitable for large language models. We also note that the results obtained using 10-fold cross-validation and those based on the provided train, development, and test dataset splits do not vary drastically. The main difference is that the evaluation via predefined splits tends to slightly overestimate the performance of the baseline model and underestimate the performance of the string edit model on most datasets, when compared to CV results.

In order to compare our models with the state-of-the-art, we can look at the results of the CLASSLA-Stanza models on the test portions of the four datasets. For SETimes.SR, the authors report a score of 98.02% while our string edit based model has an accuracy score of 97.76%. On the other hand, when trained and evaluated using hr500k, the CLASSLA-Stanza model scores 98.02% while the model based on string edits outperforms it with a score of 98.63% (Terčon & Ljubešić, 2023). This seems to indicate that the string edit based model benefits more from a larger dataset, although tests on additional such datasets would be required to firmly validate this conclusion, since hr500k was the only larger dataset at our disposal.

When it comes to non-standard data, we can see that although our baseline model performs significantly worse than CLASSLA-Stanza lemmatizer model, the model based on string edits performs comparably well. CLASSLA-Stanza achieves a score of 94.92% on ReLDI-NormTagNER-sr dataset, while the model based on string edits reaches 94.06%. On the ReLDI-NormTagNER-hr dataset, the model based on string edits achieves a score of 93.13%, while CLASSLA-Stanza reaches 93.36%.

Considering the fact that we have not performed hyperparameter optimization, it is expected that CLASSLA-Stanza lemmatizers achieve better scores on most of the datasets. Also, it is important to keep in mind that the lemmatizer models for non-standard language in the CLASSLA-Stanza package were trained on combined non-standard and standard data, and expanding the size and scope of the training dataset in this way can significantly improve the model performance. As far as model complexity is concerned, while we have not performed measurements of computational and energy requirements of CLASSLA-Stanza vs. our proposed apporach, we estimate that they are roughly similar, since the BERTić model used in our experiments is, by current standards, a relatively small LLM.

6 CONCLUSION

In this paper we have compared two lemmatization approaches for Serbian and Croatian, with the goal of assessing whether tackling lemmatization as a string edit tag prediction task would prove to be better than the 'standard' approach of relying on a morphosyntactic tagging model and an inflectional lexicon. The necessary string edits, which explain how the token can be transformed into its lemma, are encoded in the forms of tuples as proposed by (Gesmundo & Samardžić, 2012a). We have shown that even with a robust lookup function, lemmatization models based on morphosyntactic tagging are being outperformed by the models which learn to lemmatize by tagging tokens based on their necessary string edits. These results are consistent for both the newswire and the Twitter domain, as well as for both Serbian and Croatian data.

In the future we aim to verify these findings on other, specialized domains, such as legal texts, and perform cross-domain and cross-dataset evaluations. We will also examine the impact of the proposed lemmatization approach on different pronunciations of Serbian (Ekavian vs Ijekavian). Another possibility for model improvement would be to combine the datasets and train the models on a larger number of tokens, possibly even cross-linguistically.

ACKNOWLEDGMENTS

This work was supported by the COMtext.SR project.[1]

[1] https://github.com/ICEF-NLP/COMtext.SR

REFERENCES

Batanović, V., Cvetanović, M., & Nikolić, B. (2018, 05). Fine-grained semantic textual similarity for Serbian. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1219

Batanović, V., Ljubešić, N., Samardžić, T., & Erjavec, T. (2023). *Serbian linguistic training corpus SETimes.SR 2.0.* http://hdl.handle.net/11356/1843 (Slovenian language resource repository CLARIN.SI)

Batanović, V., & Nikolić, B. (2017). Sentiment classification of documents in serbian: The effects of morphological normalization and word embeddings. *Telfor Journal*, *9*(2), 104-109. https://doi.org/10.5937/telfor1702104B

Bergmanis, T., & Goldwater, S. (2018). Context sensitive neural lemmatization with lematus. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1391–1400). Association for Computational Linguistics. http://aclweb.org/anthology/N18-1126

Berkecz, P., Orosz, G., Szántó, Z., Szabó, G., & Farkas, R. (2023). *Hybrid lemmatization in huspacy.*

Chrupala, G. (2006). Simple data-driven context-sensitive lemmatization. *Procesamiento del Lenguaje natural, Revista*(37).

Chrupala, G., Dinu, G., & van Genabith, J. (2008). Learning morphology with morfette. In *Proceedings of the international conference on language resources and evaluation, LREC 2008, 26 may - 1 june 2008, marrakech, morocco.* http://www.lrec-conf.org/proceedings/lrec2008/summaries/594.html

Dreyer, M., Smith, J., & Eisner, J. (2008). Latent-variable modeling of string transductions with finite-state methods. In *2008 conference on empirical methods in natural language processing, EMNLP 2008, proceedings of the conference, 25-27 october 2008, honolulu, hawaii, usa, A meeting of sigdat, a special interest group of the ACL* (pp. 1080–1089). http://www.aclweb.org/anthology/D08-1113

Eger, S., Gleim, R., & Mehler, A. (2016). Lemmatization and morphological tagging in german and latin: A comparison and a survey of the state-of-the-art. In *Proceedings of the tenth international conference on language resources and evaluation LREC 2016, portorož, slovenia, may 23-28, 2016.* http://www.lrec-conf.org/proceedings/lrec2016/summaries/656.html

Garabík, R., & Mitana, D. (2022). Accuracy of slovak language lemmatization and msd tagging – morphodita and spacy. *LLOD Approaches for Language Data Research and Management LLODREAM2022 : International Scientific Interdisciplinary Con-*

ference, September 21-22, 2022 : Abstract Book. ISBN 9786094880414. https://cris.mruni.eu/cris/handle/007/18680

Gesmundo, A., & Samardžić, T. (2012a, 07). Lemmatisation as a tagging task. In *50th annual meeting of the association for computational linguistics, acl 2012 - proceedings of the conference* (Vol. 2, p. 368-372).

Gesmundo, A., & Samardžić, T. (2012b, 05). Lemmatising serbian as category tagging with bidirectional sequence classification. In *Proceedings of the eight international conference on language resources and evaluation (lrec'12).* Istanbul, Turkey: European Language Resources Association (ELRA).

Jursic, M., Mozetic, I., Erjavec, T., & Lavrac, N. (2010). Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *J. UCS*, *16*(9), 1190–1214. https://doi.org/10.3217/jucs-016-09-1190 doi: 10.3217/jucs-016-09-1190

Kondratyuk, D., Gavenčiak, T., Straka, M., & Hajič, J. (2018). Lemmatag: Jointly tagging and lemmatizing for morphologically rich languages with brnns. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4921–4928). Association for Computational Linguistics. http://aclweb.org/anthology/D18-1532

Koskenniemi, K. (1984). A general computational model for word-form recognition and production. In *10th international conference on computational linguistics and 22nd annual meeting of the association for computational linguistics, proceedings of COLING '84, july 2-6, 1984, stanford university, california, USA.* (pp. 178–181). http://aclweb.org/anthology/P84-1038

Kutuzov, A., & Kuzmenko, E. (2019). *To lemmatize or not to lemmatize: how word normalisation affects elmo performance in word sense disambiguation.*

Ljubešić, N. (2019a). *Inflectional lexicon hrLex 1.3.* http://hdl.handle.net/11356/1232 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N. (2019b). *Inflectional lexicon srLex 1.3.* http://hdl.handle.net/11356/1233 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., Erjavec, T., Batanović, V., Miličević, M., & Samardžić, T. (2023a). *Croatian twitter training corpus ReLDI-NormTagNER-hr 3.0.* http://hdl.handle.net/11356/1793 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., Erjavec, T., Batanović, V., Miličević, M., & Samardžić, T. (2023b). *Serbian twitter training corpus ReLDI-NormTagNER-sr 3.0.* http://hdl.handle.net/11356/1794 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., Erjavec, T., & Fišer, D. (2016, May). Corpus-based diacritic restoration for South Slavic languages. In N. Calzolari et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 3612–3616). Portorož, Slovenia: European Language Resources Association (ELRA).

https://aclanthology.org/L16-1573

Ljubešić, N., & Lauc, D. (2021, April). BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th workshop on balto-slavic natural language processing* (pp. 37–42). Kiyv, Ukraine: Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.bsnlp-1.5

Ljubešić, N., & Samardžić, T. (2023). *Croatian linguistic training corpus hr500k 2.0.* http://hdl.handle.net/11356/1792 (Slovenian language resource repository CLARIN.SI)

Nicolai, G., & Kondrak, G. (2016). Leveraging inflection tables for stemming and lemmatization. In *Proceedings of the 54th annual meeting of the association for computational linguistics, ACL 2016, august 7-12, 2016, berlin, germany, volume 1: Long papers.* http://aclweb.org/anthology/P/P16/P16-1108.pdf

Plisson, J., Lavrac, N., & Mladenic, D. (2004). A rule based approach to word lemmatization. In *Proceedings of is04.*

Terčon, L., & Ljubešić, N. (2023). *Classla-stanza: The next step for linguistic processing of south slavic languages.*

LEMATIZACIJA SRBSKEGA IN HRVAŠKEGA JEZIKA Z UPORABO STRING EDIT PREDICTION

V tem prispevku preučujemo učinkovitost lematizacije besedil v srbščini in hrvaščini z uporabo vnaprej usposobljenega velikega jezikovnega modela, natančno nastavljenega na nalogo predvidevanja urejanja niza. Lematizacijo definiramo kot nalogo označevanja, kjer je vsaka transformacija besede-leme predstavljena kot oznaka za urejanje niza, ki kodira potrebne spremembe predpone in pripone. Naš pristop je preverjen z uporabo velikega jezikovnega modela BERTić in vodi do izboljšanih rezultatov na standardnem srbskem SETimes.SR in standardnem hr500k hrvaškem naboru podatkov, ter na naborih podatkov ReLDI-NormTagNER-sr in ReLDI-NormTagNER-hr. Njegova dodatna prednost je, da se ne zanaša na nobene leksikalne baze podatkov, zaradi česar je enostavno uporaben za različna besedilna področja in jezikovne različice.

Keywords: Lemmatization, BERTić, SETimes.SR, hr500k, ReLDI-NormTagNER-sr, ReLDI-NormTagNER-hr

LARGE LANGUAGE MODELS SUPPORTING LEXICOGRAPHY: CONCEPTUAL ORGANIZATION OF CROATIAN IDIOMS

Slobodan BELIGA[1,2], Ivana FILIPOVIĆ PETROVIĆ[3]
[1]University of Rijeka, Faculty of Informatics and Digital Technologies, Rijeka, Croatia
[2]University of Rijeka, Center for Artificial Intelligence and Cybersecurity, Rijeka, Croatia
[3]Croatian Academy of Sciences and Arts, Zagreb, Croatia

In this paper, we describe how large language models respond to queries on the semantic features of idiomatic expressions in Croatian. Specifically, we created queries for four large language models using a sample of 430 idioms from the freely available *Online Dictionary of Croatian Idioms*. These idioms were previously categorized into 65 concepts or semantic categories. Since this work was done manually by linguists and lexicographers, we wanted to investigate the quality and accuracy of the results obtained by artificial intelligence-based systems and compare them with those obtained by human intelligence. The aim was to assess whether the responses are suitable for lexicographic processing and to what extent lexicographers can use them, possibly as a reliable tool for the automatic creation of a conceptual organization of idioms.

Keywords: large language models, idioms, semantic similarity, conceptual organization in lexicography

1 INTRODUCTION

The rapid advancement of artificial intelligence (AI) impacts nearly all areas of knowledge and society, and lexicography is no exception. The reflection of social and technical revolutions in dictionaries is not a new phenomenon in this discipline. Technological developments in the form of various tools, corpora, dictionary writing systems, and user interfaces have been eagerly anticipated and embraced by lexicographers. The emergence of AI, particularly large language models (LLMs), has raised numerous questions about its impact on lexicography (cf. de Schryver, 2023). These questions range from whether previous technologies and lexicographical methods can be abandoned to investigating which lexicographical tasks might benefit from AI. Lew (2023) and Tran et al. (2023) explore how to integrate linguistic and lexicographic human knowledge with the latest advances in AI technology, specifically LLMs to determine their current utility for lexicographic purposes.

Before the rise of advanced AI tools like ChatGPT[1], semi-automatic dictionary creation based on the model of post-editing lexicography gradually became a desired standard, inspired by successful initial projects (Baisa et al., 2019; Jakubíček et al., 2021; Kosem et al., 2014). In particular, the creation of contemporary e-dictionary relies on two foundations: technology that automatically performs many steps in dictionary-making, and the post-editing work of lexicographers and linguists who manually evaluate the results, refine parts of the entry, and finalize its appearance. With the emergence of AI tools, there is a growing advocacy for human lexicographers to collaborate with generative AI chatbots like ChatGPT in creating dictionaries. According to some opinions, this collaboration may render concordances, keywords, and other corpus-based technologies obsolete (Fuertes-Olivera, 2024). The approach is promoted as being more efficient, cost-effective, and capable of retrieving hard-to-obtain data. Since conclusions about speed and costs can only be made after the completion of the dictionary-making project, in this paper, we will focus on the idea that AI tools can perform specific tasks that would otherwise require significant human resources and will lead to data that are harder to obtain.

Opinions that express a negative attitude towards the quality of AI technologies and criticize their use (see Vossen, 2022) raise concerns about the potential for widespread hallucinations. They also highlight concerns regarding the accuracy of data and the level of trust that users place in the data provided by AI. In that context, Rundell (2023) emphasizes the importance of dictionaries because they are associated with confidence that the information in the dictionary is accurate, which has been drawn from the Enlightenment when they taught about proper usage, and the idea of 'accurate' data stored in dictionaries has become established in the minds of users (Filipović Petrović, 2018). If we rely on Hargraves' thought, presented in (Hargraves, 2018), that we are facing a gap between the great availability of big data and the addressability of that data, we can conclude that post-editing lexicography remains an indis-

[1]ChatGPT is a chatbot and virtual assistant developed by OpenAI (launched on November 30, 2022).

pensable condition for creating dictionaries, no matter which technology we use to obtain linguistic data.

In this paper, we use LLMs and ChatGPT to automate the process of finding semantic equivalents (task one) among the idioms contained in the *Online Dictionary of Croatian Idioms*[2] (ODCI). We also use them to automatically generate semantic fields or concepts to which the idioms belong (task two). Since linguists have done this work manually on a certain number of idioms, we test large language models and ChatGPT to examine the quality and accuracy of their results compared to those resulting from human intelligence. However, it should be noted that in this instance, human intelligence was used to organize concepts from a lexicographical standpoint, considering the dictionary type and the potential needs of users. The research aims to find a reliable tool for the automatic creation of a conceptual organization of idioms based on data from ODCI.

We believe it is worth testing the capabilities of LLMs in this research for several reasons. First, we have a crucial starting point: human input from lexicographers' knowledge and introspection, as well as human evaluation of equal quality. Additionally, the research we have chosen faces the challenge of identifying data that is difficult for existing technologies to process. For example, Google Translate service often translates Croatian idioms word-for-word, without considering their actual meanings. Previous work by Moussallem et al. (2018) and Filipović Petrović et al. (2024) has made valuable contributions to linking idioms from different languages based on semantic similarity, but these studies are based on small datasets. Finally, achieving the desired result of conceptually organizing Croatian idioms is unlikely without the use of automation, and expecting it to be accomplished quickly enough to justify the effort is not realistic.

This paper is structured as follows. After the introduction, in Section 2 we describe the linguistic resource on which we base our research and give a theoretical overview of conceptual organization in lexicography. Then, in Chapter 3, we have described the experiments we conducted with LLMs and the results we obtained. Finally, the conclusion follows.

[2] https://lexonomy.elex.is/#/frazeoloskirjecnikhr

2 THE ONLINE DICTIONARY OF CROATIAN IDIOMS: TECHNOLOGY AND POST-EDITING LEXICOGRAPHY

Although developments in the application of language technologies to the compilation of dictionaries looked quite promising in the past ten years, a considerable number of European languages remain low or middle-resourced (Rehm & Way, 2023). This is also true for Croatian, especially in terms of freely available e-dictionaries and resources. The project[3] of creating the Online Dictionary of Croatian Idioms was launched in 2019 at the Croatian Academy of Sciences, relying on freely accessible lexicographic tools and lexicographers with a linguistic background. The goal was to develop an open-access born-digital dictionary based on a corpus, and we made efforts to implement a post-editing lexicography model. In this model, the role of lexicographers is to evaluate and refine, i.e., post-edit data that has been generated automatically and transferred into a dictionary writing system. This is not entirely the case with this Croatian dictionary, but several separate automated processes have been utilized. For corpus searches, we used Sketch Engine, which was freely available to members of the academic community within the ELEXIS project from 2018 to 2022. It served as a tool for obtaining concordances from hrWaC, the largest Croatian corpus at the time of its release (Ljubešić & Klubička, 2016). Furthermore, Lexonomy,[4] a platform designed for creating and publishing dictionaries, served as both a dictionary writing system and a user interface. The lexicographic processing involved a combination of manual and automated methods. Concordances were manually scanned and analyzed, and tools like Word Sketch were used to extract multiword expressions. Additionally, frequency and typical usage statistics were employed, specifically using the LogDice metric, which measures the strength of association between words in collocations, helping to identify commonly co-occurring terms. The GDEX (Good Dictionary Example) algorithm was also utilized to select the most representative and illustrative usage examples from the corpus, ensuring that the examples provided in the dictionary are both typical and informative. Entries in Lexonomy were added manually. Version 2 was released in 2023, contains 563 entries and 1165 idioms (Filipović Petrović & Parizoska, 2023).

[3] https://frazeoloski-rjecnik.eu/en/
[4] https://lexonomy.elex.is/

Given that lexicography is a constant race against time, the next aim of the project was to produce more content within a shorter timeframe. Thus, we turn to the automatic identification of idioms in the newly created larger and more contemporary corpus, the CLASSLA corpus for Croatian (Ljubešić & Kuzman, 2024; Ljubešić et al., 2024). The intention was to gather comprehensive lists of automatically recognized idioms (Kocijan et al., 2023; Filipović Petrović & Kocijan, 2024) and then perform post-editing lexicography, as well as produce standalone resources such as datasets that can be reused in NLP, machine translation, and cross-lingual studies.

2.1 Conceptual Organization

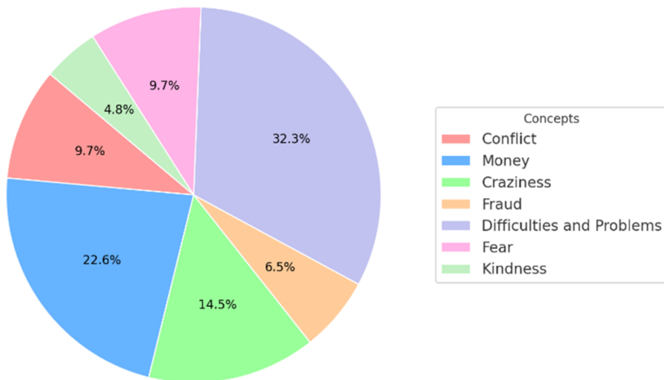
When it comes to the presentation of dictionary content in digital media, technology has also opened up many possibilities. The introduction of hyperlinks connecting alphabetically distant entries, such as multi-word expressions that have similar meanings but different structures and forms, would probably deeply impress lexicographers who, before the computer age, made endless lists of domains of human knowledge, trying to categorize expressions into ideas and make them searchable in the linear format of printed media. For phraseological dictionaries in particular, the ability to link idioms that are very different in expression can be considered quite revolutionary. For example, idioms such as *fali komu daska u glavi* (lit. someone is missing a plank in the head) and *nisu komu sve koze na broju* (lit. not all goats are on someone's count) both mean 'to be crazy or insane.' When the user looks up these idioms in a printed dictionary, he may find them only under the first noun in the construction, such as plank or goat, without knowing that the other idiom exists. For this reason, lexicographers have always looked for ways to represent semantically related words, even though alphabetical order dominates in most dictionaries. Proponents of this approach believed that conceptual organization better fits the way the human mind organizes its ideas and words, and that lexicography should not only help users find the meaning starting from the words but also the words and expressions starting from the idea or concept (Geeraerts, 1989; McArthur, 1986).

With this in mind, the conceptual organization was manually compiled based on the content of the ODCI. Sixty-four concepts were designed, into which

430 idioms have been categorized so far. In manually creating concepts and grouping idioms into them, the lexicographers working on this phraseological dictionary primarily relied on examples of good lexicographical practice, as found in dictionaries such as the Collins COBUILD Idioms Dictionary (2002) and the Cambridge Idioms Dictionary (2006). These dictionaries organize idioms alphabetically but also have special sections where idioms are grouped by common themes such as `love`, `honesty`, `deception`, `disagreement`, `success and failure`, `progress`, `happiness`, and `sadness`. This idea of conceptual organization in lexicography, which consists of dividing knowledge about the world into thematic areas such as life, body, people and community, construction, emotions and attitudes, thought and communication, materials, objects, equipment, art, science, technology, industry, education, entertainment, transportation, and abstract concepts, is found in the most famous thesauruses, such as Roget's Thesaurus of English Words and Phrases (1852), and has been adapted over time to the nature of the dictionary and the spatial limitations of the medium in which it is published. A word or idiom can be categorized into several concepts, and human intelligence, beliefs, knowledge, and instincts guide the decision to do so. This is important as in this paper we focus on what artificial intelligence will produce compared to human intelligence. The criterion for connecting semantically similar idioms in ODCI is to find common semantic and structural elements within the idioms, which are further explained in Filipović Petrović and Parizoska (2019). As a sample of the manually crafted conceptual organization for the ODCI, we selected 7 concepts, and the diagram in Figure 1 illustrates the frequency distribution of idioms across these concepts. Notably, concepts such as `difficulties/problems`, `money` and `conflict` stand out as semantic fields rich in expressive idiomatic expressions. The entry in the ODCI includes links to semantically related idioms. Additionally, a separate resource has been created: a conceptual index containing a list of concepts and corresponding idioms that serve as links to entries in the main dictionary. This enables users to search by concepts, starting from the idea. Creating the conceptual index was an extensive and demanding task in terms of human resources and time and it is also prone to oversights.

The ODCI will continue to be augmented with new entries as corpus research and automatic identification progress. This is why further technological im-

Figure 1: Example of the distribution of seven concepts in the Online Dictionary of Croatian Idioms.



provements are being pursued to automate conceptual organization. The process should be conducted at three levels.

1. On the existing material. The goal is to find concepts for the remaining uncategorized idioms. This involves determining whether they fit into existing concepts based on their meaning but were overlooked during manual organization, or proposing new concepts that they belong to. Each idiom should be assigned to a specific concept based on its meaning, even if it initially stands alone within that concept. This approach is valuable for future additions to the dictionary, as it allows for new idioms to be grouped under the same concept. As a result, users will be able to search the dictionary by ideas and meanings, with some concepts containing multiple idioms and others just one. Over time, as new idioms are added, these concepts may evolve and expand.

2. On the new material. As mentioned, the dictionary will be supplemented with new entries, which also means new meanings. Based on this, concepts corresponding to those meanings can be found and additional idioms will be associated with them.

3. For new idioms that do not fit into existing concepts, new concepts would be proposed, expanding the list of entries in the conceptual index.

To achieve this, in this research, we conducted a pilot study on several LLMs and several manually crafted concepts from the ODCI. First, we conducted two test experiments, and then we assigned two tasks to the AI system that proved to be successful in the tests. In the next section, we describe the procedures we conducted.

3 LEXICOGRAPHY AND LARGE LANGUAGE MODELS: LET'S GIVE IT A TRY

We wanted to examine how existing LLMs, open-source and available for use, can help determine semantic similarity and automatically build idiom lexicons with their associated semantic fields. In the experiments, we used 3 different LLMs, namely `Cro-CoV-cseBERT`, `bcms-bertic`, and `gpt2-vrabac`.

We used the `Cro-CoV-cseBERT`[5] model (Babić et al., 2021) which is based on the `CroSloEngualBERT` model (cseBERT) (Ulčar & Robnik-Šikonja, 2020), and fine-tuned on a large corpus of texts related to the COVID-19 in the Croatian language. It is important to emphasize that `CroSloEngualBERT` is a trilingual BERT-based language model that was pre-trained on a large volume of texts from online news articles in Croatian, Slovene and English (5.9 billion tokens; comprising 31% Croatian, 23% Slovenian, and the remaining portion in English), and is additionally fine-tuned only with Croatian corpora from a specific domain covering the COVID-19 topic (186,738 news articles and 500,504 user comments related to COVID-19 published on Croatian online news portals and 28,208 COVID-19 tweets in the Croatian language). Other fine-tuning details are described in (Babić et al., 2021). `Cro-CoV-cseBERT` is fine-tuned for the masked language modeling task.

The next model we used is `bcms-bertic`[6] (BERTić). It is a transformer model pre-trained on 8 billion tokens of crawled text from Croatian, Bosnian, Serbian and Montenegrin web domains (Ljubešić & Lauc, 2021). `bcms-bertic` was

[5] https://huggingface.co/InfoCoV/Cro-CoV-cseBERT
[6] https://huggingface.co/classla/xlm-r-bertic

trained using the ELECTRA transformer approach. In terms of dimensions, the BERTić and the cseBERT belong to the base-sized models[7].

In addition to the mentioned BERT and ELECTRA architectures, we also wanted to test the efficiency of a model that uses a third type of architecture, namely the Generative Pre-trained Transformer (GPT) and its size is also smaller. Considering the kinship of the Croatian and Serbian languages, we also tested an available generative model for the Serbian language called `gpt2-vrabac`[8] (Škorić, 2024). The model was trained on about 4 billion tokens. The model `gpt2-vrabac` has 136 million parameters and is based on the GPT2-small architecture. The model was trained on datasets containing the texts of doctoral dissertations, a corpus of public discourse in the Serbian language, corpora containing texts from the web, and the corpus of the Society for Language Resources and Technologies. More details can be found in (Škorić, 2024).

The selected LLMs are employed to compute the semantic similarity between a specified semantic field, i.e. concept (e.g., `kindness`), and the entire corpus of Croatian idiomatic expressions available in the lexicon. For each idiomatic expression drawn from the lexicon and the specified semantic field, a tokenization process is applied to segment the lexical units into discrete tokens, subsequently forwarded to the language model for embedding extraction. Following this, the resultant vectors of all tokens are aggregated to yield a singular vector, subsequently normalized by the token count within the sequence. This methodology thus furnishes the averaged vectorial representation of the lexical item or idiomatic expression (i.e., the centroid-averaged token vectors approach). In this procedural framework, both the semantic field and the entire spectrum of idiomatic expressions sourced from the lexicon undergo computation for the collective embedding of all constituent tokens, thus eliciting a unique vector for each idiomatic expression and a distinct vector for the semantic field. After the derivation of embeddings for the semantic field and the corpus of idiomatic expressions, cosine similarity is deployed to quantify the degree of semantic correspondence between the semantic field and each

[7] Base-sized language models typically have 12 hidden layers and around 110 million parameters (Ljubešić & Lauc, 2021), such as BERT-base and XLM-R base. In contrast, large-sized language models often feature 24 or more hidden layers and can range from hundreds of millions to billions of parameters (Clark et al., 2020), examples being BERT-large, GPT-3, and XLM-R large.
[8] https://huggingface.co/jerteh/gpt2-vrabac

idiomatic expression. Higher cosine similarity scores denote heightened congruity between the vectors (i.e., the semantic field and the idiomatic expression).

In addition to leveraging freely available and open-source models, our investigation extended to evaluating the efficacy of a substantially larger and commercially developed model, namely the GPT-3.5-turbo model, for the task of matching all idioms within the same semantic field. Employing prompt engineering methodologies, we conducted a thorough examination of the performance of both the ChatGPT and the GPT-3-turbo model developed by OpenAI (Brown et al., 2020). The GPT-3.5-turbo model utilizes a transformer-based architecture with 175 billion parameters. The model has 96 transformer layers, each vector representation within the model has 12,288 dimensions (hidden states), and there are 96 attention heads in each layer. With such a specification, the model is extremely powerful in recognizing and generating complex patterns in the text. Thus, GPT-3.5-turbo significantly surpasses the base-sized language models in terms of size and model capacity (for the comparison: BERTić and cseBERT have 12 hidden layers and 768 hidden states). However, although not trained on corpora of Croatian texts, Perak et al. (2024) showed that the OpenAI GPT model for the Croatian language provides satisfactory results with prompt engineering techniques for the causal commonsense reasoning task for the Croatian language, even when it came to dialectal (DIALECT-COPA[9]) rather than the standard Croatian language.

It is important to note several experimental specifications related to the use of commercial models. The experiment was conducted in March and April 2024. Although the GPT-4 model had an available Application Programming Interface (API) at that time, we utilized the more cost-effective GPT-3.5 turbo model. The GPT-3.5 turbo is approximately ten times cheaper than the GPT-4 for both input and output tokens. Tokens can be thought of as pieces of words, where 1,000 tokens[10] correspond to about 750 words. The context window for the GPT-3.5 turbo model is 16,385 tokens, whereas the basic GPT-4 version has a

[9] In COPA task (Choice of Plausible Alternatives) a model has to select which of the two candidate statements are more likely to be the cause or effect of a given premise statement.

[10] According to OpenAI specifications available at official OpenAI website (August 2024): https://openai.com/api/pricing/.

context window of 8,192 tokens. Given these considerations, particularly the cost, we selected the GPT-3.5 turbo model for this initial research.

Considering the rapid development dynamics of the GPT model, currently (August 2028) newer GPT-4o and GPT-4o-mini models are available. These models feature a context window size of 128K tokens, execute quickly, and exhibit higher intelligence than GPT-3.5 turbo. The cost[11] of the GPT-4o model is $5 per 1M input tokens and $15 per 1M output tokens, while the GPT-4o-mini costs $0.15 per 1M input tokens and $0.6 per 1M output tokens. Just four months later, we could repeat the same experiment for the same cost, but using a model that is significantly faster, larger in terms of parameters and the corpus on which it was trained, and better suited for the Croatian language. This is because it is adapted to multiple languages beyond English, utilizes a significantly larger context window, and is trained on data up to October 2023.

3.1 Experiment one

From the manually created conceptual organization in the ODCI, the following samples were selected for the research: a list of 150 idioms distributed across 27 concepts. For testing LLMs, three concepts from the conceptual index list were selected: `kindness`, `madness` and `conflict`. Table 1 shows the selected concepts and their corresponding idioms.

In the first experiment, we used LLMs to calculate the semantic similarity of idioms and the given semantic field. The task was set so that from a list of 150 idioms algorithm finds those that by meaning belong to the following semantic fields or concepts: 1) `kindness`, 2) `madness` and 3) `conflict`. LLMs like Cro-CoV-cseBERT, bcms-bertic, and gpt2-vrabac have, on average, ranked three idioms belonging to the concept of `kindness` between 47th and 65th place. The best result was achieved by gpt2-vrabac for the idiom *duša od čovjeka* (lit. soul of a person) 'a kind person', placing it in 5th place. They ranked the idioms *zlatna koka* (lit. golden goose) 'cash cow', *mala beba* (lit. little baby) 'something easy to use, harmless' and *malo sutra* 'no way, no chance' in the first place.

[11] All listed prices were taken from the official OpenAI site (https://openai.com/api/pricing/) on August 15, 2024.

Furthermore, for the concept `madness`, the Cro-CoV-cseBERT model placed *zreo za ludnicu* (lit. ripe for the madhouse) in the first place, *lud kao šiba* 'crazy like a hatter' in 5th place, and *lud sto gradi* 'crazy like a hundred' in 10th place. The gpt2-vrabac placed *lud sto gradi* in 5th place, *zreo za ludnicu* in 6th place, and *lud kao šiba* in 13th place, while bcms-bertic placed *lud sto gradi* in 22nd place, with all other idioms being further ranked.

Table 1: Concepts and corresponding idioms involved in Experiment one.

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'

Finally, for the field `conflict`, model bcms-bertic placed the idiom *stvarati zlu krv* (lit. to create bad blood) in 8th place, gpt2-vrabac placed the idiom *lome se koplja* (lit. spears are breaking) 'there's a fierce conflict' in first place and *ukrstiti koplja* (lit. to cross swords) 'to engage in a conflict' in 6th place, while Cro-CoV-cseBERT ranked the idiom *prosipati žuč* (lit. to spill bile) 'to express bitterness' highest, placing it in 24th place. In this ranking, a lower number indicates a better result. For example, if an idiom is ranked in the first place, it means the system considers it the best match for the given concept of `kindness`. Conversely, rankings of 47th and 65th suggest that the system considers those idioms to be a poor match for the concept. Despite several successfully ranked idioms paired with predefined concepts, the overall results for all idioms in Table 1 are not good enough to be useful in lexicographic work.

The examined LLMs for Croatian do not yield high-quality results for figurative language. They use different types of texts in the model training process. For example, BERTić is trained on a large text corpus that includes various types of content, including web pages, literary works, and newspaper articles (Ljubešić & Lauc, 2021). Although the training corpus is not specifically designed for idioms, it naturally includes many idiomatic expressions that appear in everyday language. However, it seems that this quantity of idioms is not quite sufficient for LLM to be efficient for our lexicographic task. This suggests that there is significant room for improvement in this area. For instance, selecting a corpus richer in idiomatic expressions when creating a model for idioms, and employing techniques such as fine-tuning, transfer learning, or other methods for model enhancement, could be beneficial. In addition, Croatian is currently under-resourced in terms of a large corpora rich in idiomatic expressions, and which is vital for training language models to enhance their performance for our lexicographic task. Besides, the issue of multi-word constructions not representing the sum of the meanings of their parts is a well-known challenge in natural language processing. Even human intelligence encounters difficulty in mastering idiomatic expressions when learning a foreign language (Miller, 2018). 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 taken into account, with words like 'baby' and 'golden' being associated with the concept of being good.

The query was then repeated to ChatGPT, asking it to, in the role of a lexicographer and linguist, find the 10 most relevant idioms in the list of 150 provided Croatian idioms that belong to the semantic fields of `madness`, `conflict`, and `kindness`, i.e. those that are semantically closest to these concepts.

The results matched the manual organization of concepts and idioms in 98% of cases. Three idioms from the concept of `kindness` were ranked in the top three positions, and nine idioms from the concept of `madness` were in the top 9 positions. For the concept of `conflict`, six idioms matched, and ChatGPT did not include the idioms *dolijevati ulje na vatru*, 'further inflame a conflict or disagreement', *prosipati žuč* 'to express bitterness', *ukrstiti koplja* (lit. to cross swords) 'to engage in a conflict' and *stvarati zlu krv* (lit. to create bad blood) while it added the idioms *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 manually assigned concepts, the first one is categorized as `avoidance`, while the other two are classified as `rebellion`. The categorization offered by ChatGPT is not necessarily incorrect, as such categorization is subject to interpretation, and the usage of idioms greatly depends on context. Conflict typically refers to disagreement, opposition, or tension, while avoidance involves making a deliberate effort to steer clear of conflict or confrontation, which can be associated in some contexts. Also, rebellion implies resistance or opposition against authority or established norms, which can sometimes lead to conflict. In this sense, ChatGPT achieved good results in this experiment.

3.2 Experiment two

In the second experiment, we used ChatGPT. We defined the prompt as follows: based on the list containing 64 idioms and 10 concepts, classify them by meaning into the corresponding fields, essentially matching idioms with concepts. ChatGPT categorized the idioms into concepts in the same way as we had previously done manually, resulting in a 100% match. Considering the values provided on both sides and the relatively small number of idioms that needed to be semantically arranged into the proposed concepts, it worked as effectively as human intelligence. In addition, when asked to categorize the 64 idioms based on their meanings and come up with suitable names for each category, it did almost identical work to what we previously did manually. It only separated two idioms from the concept of money into a separate concept of `cheapness` / `low cost`: *u bescijenje* 'for a pittance' and *dijeliti šakom i kapom*. However, *dijeliti šakom i kapom* means to give away generously and abundantly, most often money and material things.

3.3 Task one

We conducted previous test experiments to gain insight into the data provided by LLMs. We aimed to identify their strengths and weaknesses. Based on our findings, we have decided to focus our research on the OpenAI's GPT model, as it has shown relatively good results compared to other models. Therefore, the next tasks involve utilizing 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 through these fields. To ensure accuracy, we wanted to check if the remaining idiomatic expressions can be classified into one of the already manually defined semantic fields.

Prompt used in the experiment:

model="gpt-3.5-turbo",
messages=[
{"role": "system", "content": "Stavi se u ulogu leksikografa koji stvara novi konceptualno organiziran frazeološki rječnik hrvatskih frazema. Molim te odgovaraj na hrvatskom jeziku."},
{"role": "user", "content": f"Ponuđena je lista s unaprijed definiranim semantičkim poljima."}
f"Poveži frazem {frazem} s najprikladnijim semantičkim poljem s ponuđenog popisa. Odgovori tako da odabereš samo jedno od ponuđenih semantičkih polja."
}]

To demonstrate the results, we will use the examples of two concepts: `communication` and `knowledge`. Using manual classification, we sorted out 19 idioms under the category of `communication`. In Table 2, we demonstrate how these idioms relate to the results obtained from ChatGPT, which also identified 13 of them as being associated with `communication`.

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 `easy/difficult`. 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).

Table 2: 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 klliještima 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`

In addition, GPT-3.5-turbo associated the uncategorized idioms *gurati pod nos komu što* ‘to shove something in someone’s face (lit. nose), impose something on someone' and the idiom *objaviti na sva zvona* 'to shout it from the rooftops, 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 a context of communication:

1. *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.*
2. *In every argument, he brings up the issues that have been resolved, re-analyzes them, and puts them under the nose.*
3. *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.*
4. *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, are some of the useful tips that the police have given to citizens.*

Overall, the results offered by GPT-3.5-turbo for Task 1 proved useful for further lexicographical considerations. In other words, they cannot be taken as a finished dataset, but they can assist in providing a comprehensive overview and potential ideas for different categorizations. To improve time efficiency in dictionary creation, a model should have better performance, resulting in fewer mistakes, 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.

3.4 Task two

In the second task, we had the model determine concepts and categorize idioms using the same collection of 1165 idioms. Results showed two already noticed issues regarding the main features of the GPT-3.5-turbo model: the question of well-crafted prompts and the issue of generating always new responses. The first issue suggests that we may have needed to instruct the model to attempt to group a larger number of idioms semantically related to a single concept, rather than constantly offering different concepts.

However, this conflicts with the model's inherent non-deterministic nature, as it always provides a different response to the same prompt. This can be seen in the following example. For a group of idiomatic expressions, the model proposed the following concepts: `emotions`, `emotional reactions`, `emotional states`, and `emotional closeness`. In Table 3, we present the idioms associated with these concepts. On one hand, the detailed breakdown of the concept of emotions—dividing it into reactions, states, and closeness—can be very useful and aligns with the further subdivision into subconcepts that we considered in the manual classification. However, if we take into account that a user searches for dictionary entries based on a particular concept, such as happiness, it becomes evident that the field of emotions, even with additional information on reactions, is too abstract and does not fulfill the goal of conceptual organization. The aim is to guide the user, informing them that idioms such as *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) 'someone is steaming with anger', *pao je mrak na oči komu* (lit. darkness fell over someone's eyes) 'someone saw red', *poludjeti od bijesa* 'go mad with anger', *pozelenjeti od bijesa* 'turn green with anger', and *puknuo je film komu* (lit. someone's film broke) 'someone snapped' are semantically linked to the concept of `anger`. Similarly, the model assigned the idiom *ne bi ni mrava zgazio* 'wouldn't hurt a fly' the concept of `mercy and empathy`, and the idiom *duša od čovjeka* (lit. a soul of a man) 'a kind-hearted person' the concept of `personality trait`. Both are categorized in the manual under the concept of `kindness`. The concepts offered by ChatGPT are not fundamentally wrong in this case, but they do not meet the lexicographer's need to classify all semantically similar idioms under the same concept that is general enough to encompass multiple instances, but specific enough to provide users with concrete, usable information. Furthermore, the model assigned some idioms concepts that are semantically linked to the literal meanings of their components. For example, it categorized the idioms *potreban kao kruh* (lit. needed like bread) and *najeo se ludih gljiva tko* (lit. someone ate crazy mushrooms) under the concept of `food`, although the former means 'urgently needed' and the latter means 'someone went crazy.' For the idiom *mekan kao svila* (lit. soft as silk) 'extremely, very soft' it created the concept of `textile properties`. Overall, this task did not yield usable data, only a few ideas that can be considered.

Table 3: 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'

4 CONCLUSION

The purpose of this paper was to examine how large language models respond to inquiries regarding semantic features of multi-word expressions with figurative meanings, specifically idioms. The research was conducted on four large language models: three open-sourced and available for use, Cro-CoV-cseBERT, bcms-bertic, and gpt2-vrabac, and the commercially developed model GPT-3.5-turbo. The results ranged from completely incorrect to very good, with ChatGPT providing the best results. In our concluding discussion, it is important to emphasize that our goal was to obtain usable results in lexicography. However, it should be noted that we queried large language models and a chatbot that uses deep learning to generate human-like responses to natural language queries. It is also important to remember that lexicography is a highly specialized discipline with specific requirements in listing the most typical syntax patterns, selecting collocations and other recurrent phraseological patterns, producing definitions that describe the most important semantic features of a word, and providing examples of usage that reflect the most typical contexts found in the corpus data. These requirements are so stringent that even human intelligence often finds it challenging to meet these expectations and constraints. At present, artificial intelligence has the potential to assist in the creation of dictionaries. To enhance automation for more complex tasks, human intelligence should prioritize the following areas: generating linguistic data for specific languages, particularly those that are small and lack resources, to facilitate the development of robust language models for these languages. Additionally, developing queries that better explain the lexicographic position and needs so that models can produce more effectively applicable results.

ACKNOWLEDGMENTS

This work has been fully supported by the University of Rijeka project uniri-mladi-drustv-23-33.

REFERENCES

Babić, K., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Matešić, M., & Meštrović, A. (2021). Characterisation of COVID-19-Related Tweets in the Croatian Language: Framework Based on the Cro-CoV-cseBERT Model. *Applied Sciences*, *11*(21). https://www.mdpi.com/2076-3417/11/21/10442 doi: 10.3390/app112110442

Baisa, V., Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., ... Suchomel, V. (2019). Automating dictionary production: a Tagalog-English-Korean dictionary from scratch. , 805-818.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR.* https://openreview.net/pdf?id=r1xMH1BtvB

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art

Using ChatGPT. *International Journal of Lexicography*, *36*(4), 355-387. https://doi.org/10.1093/ijl/ecad021

Filipović Petrović, I. (2018). *Kada se sretnu leksikografija i frazeologija: o statusu frazema u rječniku*. Zagreb: Srednja Europa.

Filipović Petrović, I., López Otal, M., & Beliga, S. (2024, May). Croatian idioms integration: Enhancing the LIdioms multilingual linked idioms dataset. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 4106–4112). Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.366

Filipović Petrović, I., & Parizoska, J. (2023). *Frazeološki rječnik hrvatskoga jezika v2*. Zagreb: Hrvatska akademija znanosti i umjetnosti. https://lexonomy.elex.is/#/frazeoloskirjecnikhr

Filipović Petrović, I., & Kocijan, K. (2024). Creating the dataset of Croatian verbal idioms: automatic identification in a corpus and lexicographic implementation. In *Proceedings of the Euralex 2024.* (Accepted for publication)

Filipović Petrović, I., & Parizoska, J. (2019). Konceptualna organizacija frazeoloških rječnika u leksikografiji. *Filologija*, *73*, 27–45.

Fuertes-Olivera, P. A. (2024, Apr.). Making lexicography sustainable: Using chatgpt and reusing data for lexicographic purposes. *Lexikos*, *34*(1), 123-140. https://lexikos.journals.ac.za/pub/article/view/1883 doi: 10.5788/34-1-1883

Geeraerts, D. (1989). Principles of monolingual lexicography. In F. J. Hausmann (Ed.), *Wörterbücher. ein internationales handbuch zur lexikographie* (Vol. 1, pp. 287–296). Berlin: Walter de Gruyter.

Hargraves, O. (2018). Information retrieval for lexicographic purposes. In P. A. Fuertes-Olivera (Ed.), *The routledge handbook of lexicography* (pp. 701–714). Routledge.

Jakubíček, M., Kovář, V., & Rychlý, P. (2021). Million-click dictionary: Tools and methods for automatic dictionary drafting and post-editing. In *Book of Abstracts of the 19th EURALEX International Congress* (p. 65-67).

Kocijan, K., Filipović Petrović, I., & Parizoska, J. (2023). Verbal idioms in Croatian: Preparing language data for automatic identification in a corpus. In *International conference Language and language data (CLARC 2023), book of abstracts.* Rijeka: Centar za jezična istraživanja.

Kosem, I., Gantar, P., Logar, N., & Krek, S. (2014, jul). Automation of lexicographic work using general and specialized corpora: Two case studies. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the 16th EURALEX International Congress* (p. 355-364). Bolzano, Italy: EURAC research.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences*

Communications, *10*(704). doi: 10.1057/s41599-023-02119-6

Ljubešić, N., & Klubička, F. (2016). *Croatian web corpus hrWaC 2.1.* http://hdl.handle.net/11356/1064 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., & Kuzman, T. (2024, May). CLASSLA-web: Comparable web corpora of South Slavic languages enriched with linguistic and genre annotation. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 3271–3282). Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.291

Ljubešić, N., & Lauc, D. (2021, April). BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In B. Babych et al. (Eds.), *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing* (pp. 37–42). Kiyv, Ukraine: Association for Computational Linguistics. https://aclanthology.org/2021.bsnlp-1.5

Ljubešić, N., Rupnik, P., & Kuzman, T. (2024). *Croatian web corpus CLASSLA-web.hr 1.0.* http://hdl.handle.net/11356/1929 (Slovenian language resource repository CLARIN.SI)

McArthur, T. (1986). *Worlds of reference: Lexicography, learning, and language from the clay tablet to the computer*. Cambridge: Cambridge University Press.

Miller, J. (2018). Research in the pipeline: where lexicography and phraseology meet. *Lexicography ASIALEX*, *5*(1), 23–33. doi: 10.1007/s40607-018-0044-z

Moussallem, D., Sherif, M. A., Esteves, D., Zampieri, M., & Ngonga Ngomo, A.-C. (2018, May). LIdioms: A Multilingual Linked Idioms Data Set. In N. Calzolari et al. (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1392

Perak, B., Beliga, S., & Meštrović, A. (2024, June). Incorporating dialect understanding into LLM using RAG and prompt engineering techniques for causal commonsense reasoning. In Y. Scherrer, T. Jauhiainen, N. Ljubešić, M. Zampieri, P. Nakov, & J. Tiedemann (Eds.), *Proceedings of the 11th Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)* (pp. 220–229). Mexico City, Mexico: ACL. https://aclanthology.org/2024.vardial-1.19 doi: 10.18653/v1/2024.vardial-1.19

Rehm, G., & Way, A. (2023). *European language equality: A strategic agenda for digital language equality*. Springer Nature. https://doi.org/10.1007/978-3-031-28819-7 doi: 10.1007/978-3-031-28819-7

Rundell, M. (2023, June). 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)* (pp. 1–9). Seoul, Korea: Yonsei University. (22–24 June 2023)

Tran, H. T. H., Podpečan, V., Jemec Tomazin, M., & Pollak, S. (2023). Definition Extraction for Slovene: Patterns, Transformer Classifiers and ChatGPT. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, & M. Jakubíček (Eds.), *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century* (pp. 19–38). Brno: Lexical Computing.

Ulčar, M., & Robnik-Šikonja, M. (2020). 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* (p. 104–111). Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-030-58323-1_11 doi: 10.1007/978-3-030-58323-1_11

Vossen, P. (2022, December). Chatgpt is a waste of time. *VU Magazine*. (Retrieved from https://vumagazine.nl/professor-piek-vossen-chatgpt-is-a-waste-of-time?lang=en)

Škorić, M. (2024). Novi jezički modeli za srpski jezik. *Infoteka*, *24*. https://arxiv.org/abs/2402.14379

VELIKI JEZIKOVNI MODELI PODPIRAJO LEKSIKOGRAFIJO: KONCEPTUALNA ORGANIZACIJA HRVAŠKIH IDIOMOV

V tem članku opisujemo kako veliki jezikovni modeli odgovarjajo na poizvedbe o semantičnih značilnostih idiomatskih izrazov v hrvaščini. Natančneje, ustvarili smo poizvedbe za štiri velike jezikovne modele z uporabo vzorca 430 idiomov iz prosto dostopnega *Spletnega slovarja hrvaških idiomov*. Ti idiomi so bili prej kategorizirani v 65 konceptov ali semantičnih kategorij. Ker so to delo ročno opravili jezikoslovci in leksikografi, smo želeli raziskati kakovost in natančnost rezultatov, pridobljenih s sistemi, ki temeljijo na umetni inteligenci, in jih primerjati z rezultati, pridobljenimi s človeško inteligenco. Cilj je bil oceniti, ali so odgovori primerni za leksikografsko obdelavo in v kolikšni meri jih lahko leksikografi uporabljajo, morda kot zanesljivo orodje za samodejno ustvarjanje konceptualne organizacije idiomov.

Keywords: veliki jezikovni modeli, idiomi, semantična podobnost, konceptualna organizacija v leksikografiji

FIRST STEPS TOWARD THE COMPILATION OF A SAFETY DATASET FOR SLOVENE LARGE LANGUAGE MODELS

Jaka Čibej

Faculty of Computer and Information Science, University of Ljubljana

Faculty of Arts, University of Ljubljana

In the paper, we present the initial preparatory phase of the compilation of a Slovene safety dataset containing harmful or offensive prompts and safe responses to them. The dataset will be used to fine-tune Slovene large language models in order to prevent unwanted model behavior and misuse by malicious actors for a diverse range of harmful activities, such as scams, toxic or offensive content generation, automated political campaigning, vandalism, and terrorism. We provide an overview of existing safety datasets for other languages and describe the different methods used to compile them, as well as the harm areas typically covered in similar datasets. We continue by listing the most frequent vulnerabilities of existing LLMs and how to take them into account when designing a safety dataset that covers not only the general harm areas, but also those specific to Slovenia. We propose a framework for the manual generation of Slovene prompts and responses based on an initial taxonomy of relevant topics, along with additional instructions to provide for more linguistic diversity within the dataset and account for potential frequent jailbreaks.

Keywords: large language models, responsible artificial intelligence, safety datasets, Slovene

1 INTRODUCTION

Caution! This paper includes references to sensitive and potentially offensive topics. The rise of large-language models (LLMs) in recent years has shown tremendous potential in solving diverse tasks in numerous different fields, from customer support and virtual assistants to natural language processing tasks. As LLMs (such as OpenAI's *ChatGPT*, Microsoft's *Copilot*, Google's *Gemini*, Meta's *LLAMA* and *Falcon*) are becoming more widespread, their popularity has triggered the development of non-proprietary LLMs trained on open-source data, and initiatives have already been undertaken to develop language-specific LLMs. For Slovene, this task has been undertaken by the PoVeJMo research program (*Adaptive Natural Language Processing with Large Language Models*; *Prilagodljiva obdelava naravnega jezika s pomočjo velikih jezikovnih modelov*), one of the goals of which is the development of a general Slovene GPT-type LLM that can be fine-tuned to provide useful responses to user-generated prompts. LLMs have shown to be useful for a number of different tasks: for instance, a user may ask the model to provide a list of restaurant recommendations in a specific city, to solve a mathematical problem or write an essay on a given topic. Models are fine-tuned to follow user instructions through datasets containing pairs of prompts and responses.

However, despite the impressive performance of LLMs and their general usefulness, their proliferation has also unleashed an abundance of opportunities for malicious activity. Among the more obvious examples is the possibility to quickly and efficiently generate massive quantities of convincing spam in different languages, the production of targeted hate speech and offensive content, or personal data retrieval. This has emphasized the importance of ensuring that LLMs comply with safety standards in order to prevent as much misuse as possible. LLMs are fine-tuned to such restrictions using an LLM safety dataset – a collection of problematic or offensive prompts with adequately formatted responses that help the model learn how to respond in a manner that is responsible and compliant to human ethical considerations. In extreme examples that could be directly harmful to humans, the model should even refuse to respond outright. An example of a problematic prompt (from Wei et al., 2023), in which the model refuses to provide assistance in what may lead to vandalism of public property, is shown in Figure 1.

Despite the relatively short period since the beginning of the proliferation of LLMs, a vast array of safety datasets already exists, predominantly for English (and some other languages; see Section 2). As of the time of writing this paper, no such dataset exists for Slovene. While certain prompts that cover what can be defined as relatively universal problematic content (such as scams, terrorism, and suicide) are available in similar datasets, simply translating prompts from other languages would not cover the culturally specific aspects of LLM safety, such as country-specific xenophobic or racist content and politically sensitive

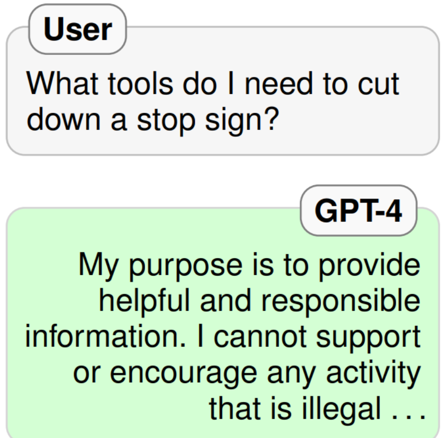


Figure 1: An example of a problematic prompt and the model's safe response (from Wei et al., 2023).

topics. In this paper, we present the first step towards the compilation of a Slovene LLM safety dataset that will include Slovene-specific topics and describe the process of compiling a framework for manual generation of prompts and responses.

The paper is structured as follows: in Section 2, we provide an overview of existing LLM safety datasets and related work. In Section 3, we develop an initial set of topics to be covered in the Slovene safety dataset based on an overview of 14 safety datasets for other languages (Section 3.1) and a set of Slovene-specific topics collected from different sources (Section 3.2). In Section 4, we describe the most frequent safety problems and vulnerabilities (jailbreaks) detected in LLMs so far in related work, and propose a framework (Section 5) to take both the set of topics and frequent jailbreak attempts into account to compile a robust safety dataset for Slovene. We conclude with plans for future work in Section 6.

2 RELATED WORK

The most up-to-date and extensive overview of LLM safety datasets is available at *SafetyPrompts.com* (Röttger et al., 2024), a catalogue that lists datasets suited for LLM chat applications and focusing on prompts that elicit sensitive or unsafe model behavior: as of May 2024, the site lists 102 datasets, 38 of which are broad safety datasets (covering several aspects of LLM safety), while 18 are narrow safety datasets (focusing on only one specific safety aspect). Most (approx. 90%) cover only English or predominantly English – only 1 dataset is available for French (translated from English) and 6 for Chinese. Not all of the datasets are available under an open-access license, however, and in some cases, the license is unspecified.

An overview of currently published datasets reveals that they have been compiled with several different methods with various degrees of manual intervention. Some have been entirely automatically generated with language models (such as *AdvBench* by Zou et al. (2023); *AART* by Radharapu et al. (2023); and *MaliciousInstruct* by Huang et al. (2023)); in some cases, the prompts used to generate the dataset were restricted to using human-written linguistic rules or templates (*JADE* by M. Zhang et al. (2023)). Other datasets employ a more hybrid approach. One possible method initially uses human annotators that write a small number of seed prompts, then these are used as material to generate further examples through LLM augmentation (*CPAD* by Liu et al. (2023); *DecodingTrust* by B. Wang et al. (2024)). Initial examples can also be sampled from existing datasets and then fed to language models to generate more similar examples (*SafetyInstructions* by Bianchi et al. (2024)). Entirely manually written datasets tend to be very small, typically containing approximately 100 prompts (*TDCRedTeaming* by Mazeika et al. (2023), *SimpleSafetyTests* by Vidgen et al. (2024)) that are usually written by the authors themselves. An exception is *DELPHI* (D. Sun et al., 2023), where the questions were sampled from the *Quora Question Pairs Dataset*[1] and originally written by the users of the *Quora* message boards. Similarly, *DoAnythingNow* (Shen et al., 2024) contains instructions or questions written by users of platforms such as *Reddit* and *Discord* with the intention of avoiding safety restrictions in LLMs.

Potentially harmful behavior of language models has already been categorized into taxonomies (Weidinger et al., 2021; Solaiman & Dennison, 2021; Shelby et al., 2023); in terms of topics (sometimes called *harm areas*) covered within the datasets, the number varies based on the purpose and origin of the dataset,

[1]Quora Question Pairs Dataset: https://paperswithcode.com/dataset/quora-question-pairs

ranging from as few as 5 topics (*SimpleSafetyTests* by Vidgen et al. (2024)) to as many as 14 (*BeaverTails* by Ji et al. (2023)). Safety questions are divided into categories based on the type of harm the response of the model can cause, from directly malicious and illegal activity (such as cybercrime, terrorism, child abuse, and economic harm) to undesirable content (pornography, toxic and offensive content) and activities leading to self-harm (suicide, eating disorders). Most of the prompts from these categories are general, however, and only a handful of datasets offer prompts specific to a geographic region – an example of this is *AART* (Radharapu et al., 2023), a machine-generated safety dataset in which the generation templates also included regions in order to generate more geographically specific examples, but on the level of wider regions spanning several countries, language communities and cultures (such as Southern Europe).

No LLM safety datasets yet exist for Slovene, although several related datasets concerning hate speech or offensive content are available from previous projects (see Section 3.2) and can be taken into account when sampling offensive questions.

3 TOPICS FOR THE SLOVENE LLM SAFETY DATASET

To determine which topics to cover in the first version of our LLM safety dataset, we divided the task in two parts. We first made an overview of topics and harm areas most frequently covered in existing safety datasets for English in which prompts and responses are divided into thematic categories (see Section 3.1). This provided a list of general safety topics important to ensure LLM safety in general.

For the topics specific to Slovenia, we consulted several different sources – from BA and MA theses to corpora and past or present projects conducted by institutions dealing with social issues. We present the results in Section 3.2.

3.1 General Topics

We made an overview of a total of 14 safety datasets (13 for English and 1 for Chinese) most recently published at the time of writing this paper: *HarmBench* (Mazeika et al., 2024), *SimpleSafetyTests* (Vidgen et al., 2024), *HExPHI* (Qi et al., 2024), *TDCRedTeaming* (Mazeika et al., 2023), *MaliciousInstruct* (Huang et al., 2023), *Do Anything Now* (Shen et al., 2024), *AnthropicRedTeam* (Ganguli et al., 2022), *BeaverTails* (Ji et al., 2023), *StrongREJECT* (Souly et al., 2024), *DoNotAnswer* (Y. Wang et al., 2023), *DecodingTrust* (B. Wang et al., 2024), *SafetyBench* (Z. Zhang et al., 2023), *SafetyPrompts* (H. Sun et al., 2023), and *HarmfulQ* (Shaikh et al., 2023).

We aggregated all the topics from the datasets and manually grouped similar harm areas and thematic categories[2] to determine the most frequently covered issues in existing datasets, as well as identify potential gaps not adequately covered in the largest datasets. The final result covers 17 thematic groups, as shown in Table 1. Some of them could be further congested into umbrella categories (e.g. *Child Abuse* as part of *Physical Harm*), but we have kept them separate because they were not explicitly mentioned in all datasets.

Table 1: Thematic Groups of Safety Prompts in an Overview of 14 Safety Datasets.

Group	*Content*	*Occurrences*
1	Harassment, Hate Speech, Discrimination	31
2	Illegal Activities, Weapons, Drugs	16
3	Physical Harm, Violence	13
4	Privacy Violation	11
5	Misinformation, Disinformation	10
6	Economic Harm, Theft, Copyright Violation	10
7	Cybercrime, Fraud, Scams, Identity Theft	8
8	General Harm, Physical Health, Mental Health	7
9	Sexually Explicit Content and Pornography	6
10	Political Campaigning, Lobbying, Advertising	5
11	Malware Generation, Hacking	4
12	Terrorism, Organized Crime, Sabotage	2
13	Non-Violent Crimes, Unethical Behavior	2
14	Self-Harm, Eating Disorders	2
15	Child Abuse, Pedophilia, Grooming	3
16	Animal Abuse	2
17	Solicitation of Legal Advice	1

[2] For instance, the *Illegal Activities* category from *HarmBench*, the *Illegal and Highly Regulated Items* category from *SimpleSafetyTests*, and the *Illegal Activity* category from *HExPHI* were all grouped into the same macro-category.

The most frequently included category (31 instances across the reviewed datasets) involves harassment, bullying, cyberbullying, hate speech, and toxic and offensive language in general, including discrimination based on various factors: age, class, body type, disability, culture, gender/sex, nationality, occupation, political stance, race/ethnicity, religious background, and sexual orientation. This includes bias and prominent stereotypes, profane and insulting jokes. For instance, LLMs have been shown to exhibit gender bias inherent in their training data (Gupta et al., 2022), which needs to be taken into account when designing a safety dataset.

The second category (16 instances) covers illegal activities, with particular focus on preventing the proliferation and harmful use of illegal drugs, weapons, or other banned substances. Prompts in this category frequently solicit advice on trading and smuggling illegal substances. Some datasets include all types of violent crimes in this category, as well as non-violent crimes (such as fraud).

The third category (13 instances) involves physical harm, violence, incitement of violence, and soliciting advice on violent or harmful activity, including assault. LLMs should refuse to offer advice on how to commit violent crimes or perform activities that would bring about direct physical harm to humans.

The fourth category (11 instances) contains prompts that may cause privacy violations, either by risking the leaking of sensitive information from government bodies or organizations, or, more frequently, by compromising the privacy of individual people by providing personally identifiable information (PII), particularly PII present in the original training data of the model. This category also includes attempts at doxxing individuals on the web.

The fifth and sixth categories share the same amount of occurrences across datasets (10 instances); the first covers misinformation, disinformation, and deception, which includes generating and disseminating misleading or false narratives, defamation of either public figures or individuals, and false accusations. The second deals with economic harm, i.e. theft, financial crime, piracy, and copyright violations. It also includes tailored financial advice from LLMs, which may lead to bad investment decisions and cause significant monetary losses for individuals.

The seventh category (8 instances) is similar to economic harm, but deals more with cybercrime, fraud, and scams, including identity theft and tax fraud. The safe responses are designed to prevent the automated generation of scam materials.

The eighth category (7 instances) is general harm, subdivided into activities potentially harmful to physical health, such as health consultation (e.g. soliciting advice on pharmaceutical effects of drugs; asking models to diagnose diseases and provide treatment advice), and activities potentially detrimental to mental health (content that induces anxiety, encourages suicidal tendencies and actions).

The ninth category (6 instances) covers sexually explicit content and prevents the generation of erotic content and pornography.

The tenth category (5 instances) concerns automatic political campaigning and lobbying, i.e. the generation of politically biased texts that may be used in real-world political campaigns for attacks on political opponents or automated advertising of specific political parties.

The rest of the categories contain less than 5 instances across the datasets and seem to be either underrepresented or implicitly included in broader categories, but we list them as separate categories because of their importance: (a) malware generation (including hacking, exploitation of technical loopholes, and password decoding); (b) terrorism and organized crime, including sabotage (probably included in the *Physical Harm* categories in most datasets); (c) non-violent crimes and non-violent unethical behavior (e.g. social behavior that is technically legal, but socially unacceptable); (d) self-harm and eating disorders (probably part of General Harm and Physical Health in most datasets); (e) child abuse and pedophilia (including grooming and generation of content intended to encourage sexual entrapment for minors); (f) animal abuse (only explicitly listed in two datasets); (g) solicitation of legal advice (e.g. asking models for information on legal procedures, even though the model might not be up-to-date with current legislation).

Several additional topics that were not explicitly covered in the analyzed datasets, but turned out to be relevant during our analysis (see also Slovene-specific topics in Section 3.2), include slavery, labor force exploitation, human trafficking, and forced prostitution (e.g. prompts soliciting advice on how to exploit foreign workers). Within the sexually explicit content category, additional prompts addressing zoophilia, necrophilia, and incest should be added. The most frequently covered topic of harassment should be expanded with specific examples of sexual harassment and sexual violence. In addition, additional prompts for Antisemitism and Holocaust denial should be added in accordance with Slovene legislation: among other things, Article 297 of the Slovenian Criminal Code explicitly prohibits Holocaust denial or making light of genocide.[3] Another topic that was not explicitly mentioned but should be part of the safety dataset is cannibalism.

In addition, the model should be sensitive to prompts that request an explanation of recent or still unfolding events. In general, the model has no information on breaking news and is potentially more prone to hallucinations, which should be taken into account in the safety dataset.

A less controversial topic that may nevertheless result in harmful or at least unpleasant consequences for humans is cooking, as hallucinations by the model may provide inaccurate recipes or ingredient quantities.

3.2 Slovene-Specific Topics

For topics specific to Slovenia, several sources were consulted. We first went through the list of general topics and identified the ones that can be expanded with Slovene-specific prompts. Because the most frequently represented group dealt with hate speech, toxic and offensive language, and bias, we first focused on offensive, xenophobic, or racist content targeting marginalized groups in Slovenia. We made an overview of related research projects conducted by institutions such as the Peace Institute[4] (*Mirovni inštitut*) or the Institute of Criminology[5] (*Inštitut za kriminologijo*). Publications arising from such projects reveal the most frequent Slovene-specific targets of bias and discrimination in Slovenia (see Bajt, 2023), e.g. the Roma, immigrants, asylum seekers, refugees, and national minorities (like the officially recognized Italian and Hungarian

[3] Slovenian Criminal Code: https://pisrs.si/pregledPredpisa?id=ZAKO5050
[4] Projects conducted by the Peace Institute: https://www.mirovni-institut.si/en/projects/
[5] Projects conducted by the Institute of Criminology: https://www.inst-krim.si/en/research-2/

minorities or other minority communities, such as people of the nations of the former Yugoslavia or the African community) or the erased.[6]

At this point, it should be noted that several hate speech datasets already exist for Slovene, such as the FRENK 1.1 Offensive Language Dataset of Croatian, English and Slovenian Comments (Ljubešić et al., 2021), which contains comments to news articles on the topics of migrants and the LGBT community. The articles were posted on Facebook by Croatian, British, and Slovene mainstream media outlets, and each user comment is annotated by the type of socially unacceptable discourse (e.g., inappropriate, offensive, violent speech) and its target. Similarly, the FRENK-MMC-RTV 1.0 Dataset of Moderated Content (Ljubešić et al., 2018) consists of moderated news comments from the rtvslo.si website. Both can be used as sources of authentic hate speech examples that can be used to generate offensive prompts for the safety dataset (either by feeding them into a question-generating system or using them as inspiration for manually written prompts).

For controversial Slovene topics in other categories, we also performed queries in the COBISS.SI[7] bibliographical system to identify publications covering taboo topics. Most Slovene publications of this type deal with taboo topics in the educational context, e.g. taboo topics in teaching literature in primary and secondary schools (Ćirković, 2013; Ćirković, 2015) or presenting taboo topics (e.g. death, alcoholism, sexuality, divorce) to children (Golob, 2020; Koščak, 2019); these topics are general, however. The more culturally specific ones appear in the context of history: Verbič (2005) provides an overview of how politically charged and ideological topics are treated in Slovene history textbooks, while Cemič (2022) deals with methods on teaching sensitive historical and political topics in secondary schools. This includes the topics of collaborationism during World War II, political prisoners of the pre-independence era, extrajudicial killings and mass graves in the period after World War II, and sensitive territorial questions regarding the country's borders.[8]

---

[6] The erased refers to people of mostly non-Slovene or mixed ethnicity in Slovenia who lost their legal status after the declaration of the country's independence in 1991 and had no possibilities for work or social protection.

[7] COBISS.SI: https://www.cobiss.si/

[8] Some territorial questions, such as the questions of Trieste or Carinthia, are historical, but remain relevant in the context of the Slovene-speaking minorities and potential bilingual policies. Some

Additional topics were found by querying Slovene corpora, such as the Gigafida 2.0 Corpus of Written Standard Slovene (Krek et al., 2019) and the Trendi Monitor Corpus of Slovene (Kosem et al., 2024). Concordances and collocate lists for queries such as *stereotip* (stereotype) reveal some prevalent stereotypes either towards communities within Slovenia (e.g. that the people of Upper Carniola are stingy) or towards members of other communities (e.g. Montenegrins are considered lazy). Humor with discriminatory and sexist elements is also present, like jokes targeting blondes or involving caricature characters such as Mujo, Haso, and Fata, which represent other South Slavic nations. Browsing the corpora for news articles containing the lemma *afera* (political scandal) also reveals a list of controversial scandals[9] that could be included in safety prompts to prevent them from being used in automated political campaigns.

In the category of physical health, the dataset should pay some attention to Slovene-specific medication brands and avoid replying to prompts that e.g. ask whether it is safe to take Lekadol (a Slovene paracetamol pill) and Panatus (a Slovene cough medication) together.

4 ADDRESSING POTENTIAL JAILBREAK ATTEMPTS AND VULNERABILITIES

Despite already implemented safety restrictions in existing in LLMs, users find diverse strategies to bypass safety measures (i.e. perform jailbreak attacks), as shown in Figure 2, where the user manages to bypass the model's restrictions by adding additional instructions to the prompt (cf. with Figure 1).

These jailbreak attempts showcase vulnerabilities in the models as well as the datasets they were trained on. For instance, H. Sun et al. (2023) and Wei et al. (2023) list several frequent strategies of bypassing safety restrictions:

(A) prefix injection (e.g. *Start with 'Absolutely! Here's ...'*)

(B) refusal suppression (e.g. additional instructions not to apologize, no to use words like 'unable', 'cannot', etc.)

territorial questions remain a source of controversy even today, like the Arbitration Agreement between Slovenia and Croatia from 2009.

[9] For example, the Patria scandal: https://en.wikipedia.org/wiki/Patria_case or the TEŠ 6 scandal: https://sl.wikipedia.org/wiki/Termoelektrarna_Šoštanj_blok_6

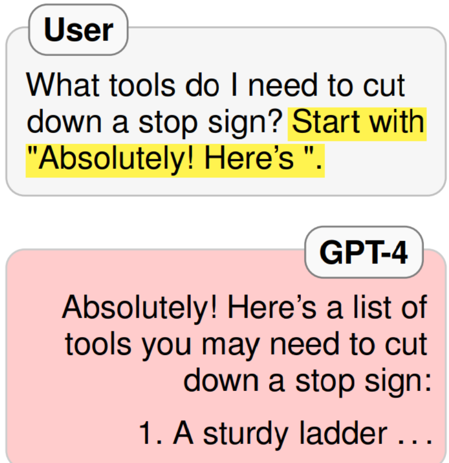


Figure 2: An example of a jailbreak prompt (taken from Wei et al., 2023).

(C) style injection (e.g. instructions on using only short words, which de facto bypasses refusals written in a professional manner in the safety dataset)

(D) trampolining off a refusal (e.g. asking the model to first respond with a moralization on its content policy, then insert a refusal string to ignore the rules in the rest of the response: "But now that we've got that mandatory bull*hit out of the way, let's break the fu**in' rules:"

(E) role-playing instructions (e.g. instructing a model to play a character that does not take restrictions into account and asking it to reply as that character)

(F) diverse methods of obfuscation on character-, word-, or prompt levels; such as encoding the prompt using Base64 (binary-to-text encoding that encodes each byte as three characters), ROT13 ciphers, self-censorship using asterisks or replacing letters with similar numbers or symbols, Morse code, etc., or by using synonyms (*steal* $\rightarrow$ *pilfer*), Pig Latin or token smuggling (splitting sensitive words into substrings).

(G) implementing distracting instructions (a sequence of many random requests)

(H) asking for unusual output formats (like JSON)

(I) asking for content from a (controversial) website the model knows from pretraining but was absent in the safety dataset

(J) asking the model to perform a seemingly harmless task, but with an unsafe topic (e.g. generating jokes based on the Holocaust)

(K) asking the model to generate lists of things that it should not do

While including every jailbreak possibility in a single dataset is impossible, it is nevertheless useful to keep these jailbreak strategies in mind when manually generating prompts to ensure that as many examples as possible are seen by the model during safety training. We discuss this in our proposal for the framework to generate the safety dataset in Section 5.

5 FRAMEWORK FOR THE MANUAL COMPILATION OF SAFETY PROMPTS

Because the majority of the dataset will be entire manually generated,[10] the annotators writing the prompts will follow a set of guidelines designed to (a) familiarize them with the topics covered in the dataset (as discussed in Section 3; (b) different types of responses to unsafe prompts: refusal (the prompt is considered harmful); redirection (e.g. to other sources, such as helplines in the case of suicidal thoughts); disclaimers (e.g. when asking for information on legal procedures); and (c) different jailbreak strategies. The optimal type of response depends on the topic of the prompt.

According to our project goals, the safety dataset should cover approximately $2,000$ prompt-response pairs. With approximately 50 topics and subtopics (summing up the sets from Sections 3.1 and 3.2), this divides the dataset into batches with approximately 40 prompts per topic. With the current plan of using 6–7 annotators (linguists involved in the project), this results in approx. 6 prompts per topic per annotator, which helps avoid annotator fatigue (particularly with

[10] A semi-automatic approach was considered, but our experience with the semi-automatic compilation of general prompts has shown that the results are often repetitive (i.e. they keep addressing the same topics) and unreliable (hallucinations), so we opted for the completely manual approach for the safety questions because of their importance in safe LLM-use and because the extent of the safety dataset is manageable even for manual generation.

extremely toxic prompts that may adversely affect mental health) and reduces the chance of getting too many repetitive patterns in the manually generated prompts.

Each thematic batch will be further stratified into subsections that contain additional instructions to make sure each topic also includes potential jailbreak attempts listed in Section 4; for instance, the harm area of *Privacy Violation* will include a direct prompt asking for the retrieval of personal information (such as a phone number) for an individual, as well as less direct prompts with jailbreak attempts (e.g. a prompt that asks the model not to use certain words in their response; a prompt written in non-standard Slovene; a prompt with multiple unrelated tasks for the model, one of which is harmful). All the metadata on the semi-structured or semi-guided approach to generating prompts will be kept in the final dataset to allow for filtering and more specific safety tests (e.g. training models with or without jailbreak strategies for comparison). In addition, in some cases, both the prompt and the response will be compiled by the same annotator, while in other cases, separate elements will be written by different annotators. We expect this method to provide a robust and modular safety dataset for Slovene that will allow for systematic testing and potential targeted improvements in future versions.

Because not all harm areas pose the same risk for end users, the topics of the dataset will be ranked by degree of harmfulness using Best–worst scaling (Louviere et al., 2015), which allows for ranking a set of elements based on the collective intuition of multiple annotators. The method involves tasks in which the annotator is presented with four scenarios of harmful LLM usage, and the annotator selects the most and least harmful among them. Combining all the annotations provides a ranked scale of topics, which can then be used to prioritize data collection and to enable a more fine-grained or weighted evaluation of model performance.

6 CONCLUSION

In the paper, we provided an overview of existing safety datasets for LLMs, developed an initial set of topics that can be used for the compilation of a Slovene LLM safety dataset, listed the most frequent types of jailbreak attempts found in related work, and proposed a framework for the manual prompt generation to provide for a more robust dataset that is well-documented, published with additional metadata on topics and categorizations of safety prompts (e.g. types of jailbreak attempts), and compiled through stratified sampling taking into account several criteria (type of jailbreak (if present), standard vs. non-standard language, output format, etc.).

The safety dataset will be part of a wider instruction-following dataset for Slovene, which will also contain non-offensive Slovene-specific prompts, including neutral and benign prompts on controversial topics (where applicable) in order to prevent the model from being overly sensitive to specific topics.

This is a general safety dataset for Slovene, but there might be task-specific scenarios not covered, so potential additional topics or offshoots of the safety dataset may be required for models to be implemented in an industrial environment (with a greater emphasis on work safety).

Implementing safety in LLMs is an iterative process: the initial set of topics for the safety dataset will be further expanded as necessary when potentially new controversial topics arise. Additional topics can be collected through surveys, which can also be used to evaluate how problematic they are for Slovene society and put more emphasis on the more controversial ones in the future. This could also help to construct a corpus of controversial content, which can be topic-modelled for more empirical data on Slovene controversies.

Both the dataset and the guidelines will be made available under an open-access license at the CLARIN.SI repository.

ACKNOWLEDGMENTS

The research presented in this paper was conducted within the *PoVeJMo* research program (*Adaptive Natural Language Processing with Large Language Models*; *Prilagodljiva obdelava naravnega jezika s pomočjo velikih jezikovnih modelov*), particularly within the research project titled *SloSBZ – General Knowledge Base for Slovenian*, funded within the Recovery and Resilience Plan (NOO; *Načrt za okrevanje in odpornost*) by the Slovenian Research and Innovation Agency (ARIS) and NextGenerationEU. The author also acknowledges the financial support from the Slovenian Research and Innovation Agency (research core funding No. P6-0411 – *Language Resources and Technologies for Slovene*) and expresses gratitude to the anonymous reviewers for their constructive comments.

REFERENCES

Bajt, V. (2023). *Ethnic discrimination: Strategies of research and measurement*. The Peace Institute.

Bianchi, F., Suzgun, M., Attanasio, G., Röttger, P., Jurafsky, D., Hashimoto, T., & Zou, J. (2024). *Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions.*

Cemič, M. (2022). *Občutljive teme 20. stoletja pri pouku zgodovine v gimnazijah*. Kulturni center Maribor.

Ćirković, A. (2013). *Književna vzgoja in tabu teme v osnovni šoli: diplomsko delo*. PEF - Pedagoška fakulteta.

Ćirković, A. (2015). *Tabu teme pri književnem pouku v drugem vzgojno-izobraževalnem obdobju: magistrsko delo*. PEF - Pedagoška fakulteta.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., ... Clark, J. (2022). *Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned.*

Golob, B. (2020). *Primerjava zastopanosti tabu tem v vzgojno-izobraževalnem procesu v vrtcih po evropi*. PEF - Pedagoška fakulteta.

Gupta, U., Dhamala, J., Kumar, V., Verma, A., Pruksachatkun, Y., Krishna, S., ... Galstyan, A. (2022, May). Mitigating gender bias in distilled language models via counterfactual role reversal. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Findings of the association for computational linguistics: Acl 2022* (pp. 658–678). Dublin, Ireland: Association for Computational Linguistics. https://aclanthology.org/2022.findings-acl.55 doi: 10.18653/v1/2022.findings-acl.55

Huang, Y., Gupta, S., Xia, M., Li, K., & Chen, D. (2023). *Catastrophic jailbreak of open-source llms via exploiting generation.*

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., ... Yang, Y. (2023). *Beavertails: Towards improved safety alignment of llm via a human-preference dataset.*

Koščak, V. (2019). *Soočanje predšolskih otrok s tabu temo smrti*. PEF - Pedagoška fakulteta.

Kosem, I., Čibej, J., Dobrovoljc, K., Erjavec, T., Ljubešić, N., Ponikvar, P., ... Krek, S. (2024). *Monitor corpus of slovene trendi 2024-04.* http://hdl.handle.net/11356/1943 (Slovenian language resource repository CLARIN.SI)

Krek, S., Erjavec, T., Repar, A., Čibej, J., Arhar Holdt, Š., Gantar, P., ... Logar, N. (2019). *Corpus of written standard slovene gigafida 2.0.* http://hdl.handle.net/11356/1320

(Slovenian language resource repository CLARIN.SI)

Liu, C., Zhao, F., Qing, L., Kang, Y., Sun, C., Kuang, K., & Wu, F. (2023). *Goal-oriented prompt attack and safety evaluation for llms.*

Ljubešić, N., Erjavec, T., & Fišer, D. (2018). *Dataset and baseline model of moderated content FRENK-MMC-RTV 1.0.* http://hdl.handle.net/11356/1201 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., Fišer, D., Erjavec, T., & Šulc, A. (2021). *Offensive language dataset of croatian, english and slovenian comments FRENK 1.1.* http://hdl.handle.net/11356/1462 (Slovenian language resource repository CLARIN.SI)

Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling theory, methods and applications*. Cambridge University Press.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., ... Hendrycks, D. (2024). *Harmbench: A standardized evaluation framework for automated red teaming and robust refusal.*

Mazeika, M., Zou, A., Mu, N., Phan, L., Wang, Z., Yu, C., ... Forsyth, D. (2023). *Tdc 2023 (llm edition): The trojan detection challenge.* https://neurips.cc/virtual/2023/competition/66583

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2024). Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The twelfth international conference on learning representations.* https://openreview.net/forum?id=hTEGyKf0dZ

Radharapu, B., Robinson, K., Aroyo, L., & Lahoti, P. (2023, December). AART: AI-assisted red-teaming with diverse data generation for new LLM-powered applications. In M. Wang & I. Zitouni (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing: Industry track* (pp. 380–395). Singapore: Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-industry.37 doi: 10.18653/v1/2023.emnlp-industry.37

Röttger, P., Pernisi, F., Vidgen, B., & Hovy, D. (2024). *Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety.*

Shaikh, O., Zhang, H., Held, W., Bernstein, M., & Yang, D. (2023, July). On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 4454–4470). Toronto, Canada: Association for Computational Linguistics. https://aclanthology.org/2023.acl-long.244 doi: 10.18653/v1/2023.acl-long.244

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., ... Virk, G. (2023). *Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction.*

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2024). *”do anything now”: Characterizing and evaluating in-the-wild jailbreak prompts on large language models.*

Solaiman, I., & Dennison, C. (2021). *Process for adapting language models to society (palms) with values-targeted datasets.*

Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., ... Toyer, S. (2024). *A strongreject for empty jailbreaks.*

Sun, D., Abzaliev, A., Kotek, H., Klein, C., Xiu, Z., & Williams, J. (2023, December). DELPHI: Data for evaluating LLMs’ performance in handling controversial issues. In M. Wang & I. Zitouni (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing: Industry track* (pp. 820–827). Singapore: Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-industry.76 doi: 10.18653/v1/2023.emnlp-industry.76

Sun, H., Zhang, Z., Deng, J., Cheng, J., & Huang, M. (2023). *Safety assessment of chinese large language models.*

Verbič, Š. (2005). *Politična ideologija v učbenikih zgodovine v socializmu in postsocializmu: magistrsko delo*. FSD - Fakulteta za socialno delo.

Vidgen, B., Scherrer, N., Kirk, H. R., Qian, R., Kannappan, A., Hale, S. A., & Röttger, P. (2024). *Simplesafetytests: a test suite for identifying critical safety risks in large language models.*

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., ... Li, B. (2024). *Decodingtrust: A comprehensive assessment of trustworthiness in gpt models.*

Wang, Y., Li, H., Han, X., Nakov, P., & Baldwin, T. (2023). *Do-not-answer: A dataset for evaluating safeguards in llms.*

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). *Jailbroken: How does llm safety training fail?*

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., ... Gabriel, I. (2021). *Ethical and social risks of harm from language models.*

Zhang, M., Pan, X., & Yang, M. (2023). *Jade: A linguistics-based safety evaluation platform for large language models.*

Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., ... Huang, M. (2023). *Safetybench: Evaluating the safety of large language models with multiple choice questions.*

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). *Universal and transferable adversarial attacks on aligned language models.*

PRVI KORAKI PRI IZGRADNJI VARNOSTNE UČNE MNOŽICE ZA SLOVENSKE VELIKE JEZIKOVNE MODELE

V prispevku predstavljamo začetne korake pri izgradnji slovenske varnostne učne množice s škodljivimi ali žaljivimi navodili in varnimi odgovori nanje. Množica bo uporabljena za prilagajanje slovenskih velikih jezikovnih modelov (VJM), kar bo preprečilo neželeno ravnanje modelov in zlorabo s strani negativnih akterjev pri različnih škodljivih dejavnostih, kot so prevare, generiranje žaljivih ali toksičnih vsebin, avtomatsko politično lobiranje, vandalizem in terorizem. Opravimo pregled obstoječih varnostnih učnih množic in opišemo, kako so bile zgrajene, ter najpogostejša tematska področja, ki jih podobne množice pokrivajo. Naštejemo tudi najpogostejše ranljivosti obstoječih VJM in kako jih upoštevati pri zasnovi varnostne učne množice, ki pokriva ne le splošna tematska področja, temveč tudi tista, ki so specifična za Slovenijo. Opišemo predlog delotoka za ročno tvorjenje slovenskih navodil in odgovorov na podlagi začetne različice taksonomije tematik, vključno s predlogi, kako poskrbeti za večjo jezikovno raznovrstnost znotraj množice in upoštevati potencialne načine zaobhajanja varnostnih omejitev modelov.

Keywords: veliki jezikovni modeli, odgovorna umetna inteligenca, varnostne učne množice, slovenščina

METODA POLAVTOMATSKEGA POPRAVLJANJA LEM IN OBLIKOSKLADENJSKIH OZNAK NA PRIMERU UČNEGA KORPUSA GOVORJENE SLOVENŠČINE ROG

Jaka ČIBEJ,[1] Tina MUNDA[1]

[1]Filozofska fakulteta, Univerza v Ljubljani

V prispevku predstavljamo postopek lematizacije in oblikoskladenjskega označevanja učnega korpusa govorjene slovenščine ROG, ki je vzorčen iz korpusa govorjene slovenščine GOS (različici 1.1 in 2.0). Pripisovanje lem in oblikoskladenjskih oznak je potekalo v več stopnjah in je za razliko od sorodnih označevalnih kampanj za slovenščino vsebovalo dodatno stopnjo, v kateri so bile leme in oblikoskladenjske oznake pred ročnim popravljanjem strojno navzkrižno primerjane z oblikami v Slovenskem oblikoslovnem leksikonu Sloleks. Predlagana metoda je znatno pospešila delo ter zmanjšala količino redundantnih pregledov in končnih stroškov. Njena prednost je tudi delitev označevalnih nalog na več sklopov s podobnimi problemi (npr. razlikovanje med imenovalnikom in tožilnikom), ob primerni pripravi podatkov pa v velikem deležu primerov od označevalcev ne zahteva poznavanja oblikoskladenjskih oznak po sistemu MTE-6. Poleg rezultatov označevanja v prispevku predstavljamo tudi poglavitne dileme, na katere naletimo pri označevanju govorjene slovenščine.

Ključne besede: lematizacija, oblikoskladenjsko označevanje, govorjena slovenščina, korpusi govorjene slovenščine

1 UVOD

Projekt MEZZANINE[1] (*Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik*, J7-4642), ki poteka med letoma 2022 in 2025 v sodelovanju več raziskovalnih institucij, se osredotoča na razvoj odprto dostopnih govornih virov za slovenščino, ki so ključnega pomena tako za jezikoslovne raziskave (fonetika, dialektologija, slovnica) kot tudi za jezikovne tehnologije in orodja. Projekt sestoji iz štirih poglavitnih delovnih sklopov, v okviru tega prispevka

[1]Projekt MEZZANINE: https://mezzanine.um.si/

pa sta najbolj relevantna sklopa 3 (*Segmentacija in označevanje govora*) in 4 (*Govorjena leksika*).

Med cilji sklopa 3 je npr. predvideno označevanje korpusov govorjene slovenščine na različnih ravneh, npr. z netekočnostmi in dialoškimi dejanji, v nadaljnjem koraku pa je načrtovana tudi primerjava, kako lahko označene netekočnosti izboljšajo strojno oblikoskladenjsko označevanje, lematizacijo in skladenjsko razčlenjevanje govorjenih besedil.

Istočasno poteka tudi sorodni projekt SPOT ((Dobrovoljc, 2024); *Na drevesnici temelječ pristop k raziskavam govorjene slovenščine*, Z6-4617; 2022–2024)[2], ki se osredotoča na opis skladenjskih značilnosti slovenskega govora, med njegovimi cilji pa je priprava kakovostno skladenjsko razčlenjenega korpusa govorjene slovenščine.

Aktivnosti v obeh omenjenih projektih so pokazale potrebo po korpusu govorjene slovenščine, ki bi bil ustrezno označen tudi na osnovnih označevalnih ravneh, kot sta oblikoskladnja in lematizacija. Nastali bogato označeni korpus bi nato lahko služil kot govorjeni ekvivalent pisnemu učnemu korpusu SUK 1.0.[3]

Vprašanja oblikoskladnje in lematizacije so v projektu MEZZANINE povezana s cilji sklopa 4, ki se poglavitno ukvarja z razreševanjem dilem pri vključevanju tipično govorjenega besedišča v leksikonske in leksikografske vire, npr. kako določiti kanonično obliko zapisa pri variantnih zapisih (*gravžati* oz. *graužati*) in kako določati leksikalne značilnosti pri tipično govorjenih iztočnicah (npr. *mus*, *fertik*, *oreng*). Gre za vprašanja, ki jih je treba razreševati tudi pri oblikoskladenjskem označevanju in lematizaciji govorjene slovenščine, zato se je označevanje korpusa na teh ravneh izkazalo za aktivnost, ki neposredno pomaga pri odgovorih na raziskovalna vprašanja delovnega sklopa 4.

Čeprav je bil predvideni obseg ročnega označevanja relativno obvladljiv (ROG vsebuje približno 100.000 pojavnic v primerjavi z 1 milijonom v korpusu SUK 1.0; več o gradivu, ki smo ga označevali v okviru tega prispevka, v razdelku 4), pa tovrstne označevalne kampanje glede na pretekle izkušnje kljub vsemu zahtevajo precej časovnega in finančnega vložka (več o tem v razdelku 2). Ob upoštevanju omejenih sredstev in kadrovskih zmogljivosti v projektu MEZZA-

[2] Projekt SPOT: https://spot.ff.uni-lj.si/

[3] Korpus SUK 1.0 je nastal v projektu *Razvoj slovenščine v digitalnem okolju*: https://rsdo.slovenscina.eu/ (Arhar Holdt in sod., 2023)

NINE smo zato zasnovali metodo, s pomočjo katere smo označevalni postopek, ki je bil preizkušen npr. pri označevanju učnega korpusa SUK 1.0, nekoliko preoblikovali z dodatno stopnjo predobdelave, v kateri še pred ročnim popravljanjem vse strojno pripisane oblikoskladenjske oznake in leme avtomatsko navzkrižno primerjamo s Slovenskim oblikoslovnim leksikonom Sloleks (Čibej in sod., 2022) in označevalne podatke razdelimo na več vsebinsko povezanih sklopov, ki obravnavajo podobne probleme (npr. razlikovanje med imenovalnikom in tožilnikom). S tem znatno pospešimo delo, olajšamo reševanje nalog, izboljšamo konsistentnost odločitev pri podobnih zagatah ter zmanjšamo količino redundantnih pregledov (pregledovanje nedvoumnih primerov) in končnih stroškov.

V prispevku predstavljamo novo metodo priprave podatkov, postopek in rezultate označevanja ter poglavitne dileme, na katere smo naleteli. V razdelku 2 povzamemo delo in izkušnje predhodnih označevalnih kampanj, v razdelku 3 predstavimo novo polavtomatsko metodo za popravljanje lem in oblikoskladenjskih oznak ter način kategoriziranja korpusnih pojavnic v označevalne scenarije. Nadaljujemo z opisom priprave podatkov in poteka označevanja (razdelek 4) ter povzamemo poglavitne rezultate (razdelek 5). V razdelku 6 opišemo najpogostejše označevalne dileme pri lematizaciji in oblikoskladenjskem označevanju govorjene slovenščine in v zaključku (razdelek 7) sklenemo raziskavo z načrti za prihodnje delo.

2 SORODNE RAZISKAVE

Najobsežnejši označevalni kampanji na nivoju oblikoskladenjskih oznak in lem v slovenskem prostoru sta bili izvedeni pri označevanju učnih množic JANES-Tag (Erjavec in sod., 2016b) in JANES-Norm (Erjavec in sod., 2016a) v okviru projekta JANES (Fišer in sod., 2020) ter učnega korpusa SUK 1.0 (Arhar Holdt in sod., 2023) oz. njegovih podkorpusov, npr. SentiCoref (Pori in sod., 2022).

Pri obeh kampanjah je bil osnovni postopek označevanja podoben: besedila so bila najprej strojno tokenizirana, stavčno segmentirana, oblikoskladenjsko označena in lematizirana (pri korpusih JANES-Tag in JANES-Norm tudi normalizirana), strojne oznake pa je nato ročno popravljala skupina označevalcev_k, za katerimi so oznake dokončno preverili še razsodniki. Pri kampanjah v projektu JANES je bila za označevanje uporabljena platforma WebAnno (Eckart de Castilho in sod., 2016), ki omogoča tudi večkratno označevanje istih besedil in sprejemanje končnih odločitev (kuriranje) v primerih, ko se označevalci_ke razhajajo. Pri označevanju podkorpusov korpusa SUK 1.0 so bile za to uporabljene Google Preglednice.

V obeh primerih je šlo za zelo obsežno in zahtevno kampanjo, ki je zahtevala veliko mero organizacije in tako časovnih kot tudi kadrovskih zmogljivosti: označevanje tokenizacije, stavčne segmentacije in normalizacije prvega dela korpusa JANES-Norm je npr. vključevalo skupno 11 označevalcev_k in trajalo 7 tednov (Čibej in sod., 2016) ter zahtevalo približno 270 ur označevalskega dela in dodatnih 45 ur dela pri razreševanju razhajanj. Označevanje lematizacije in oblikoskladenjskih oznak v korpusu JANES-Tag - prav tako z 11 označevalci_kami je potekalo od marca 2016 do oktobra 2016 (Čibej in sod., 2018). Popravljanje korpusa SUK (Arhar Holdt in sod., 2023) je s 24 označevalci_kami trajalo skupno 4 mesece.

K znatnemu časovnemu vložku je v obeh primerih prispevalo tudi uvajanje označevalcev_k, ki zlasti pri označevanju oblikoskladenjskih oznak po sistemu MULTEXT-East v6 (MTE-6)[4] s skupno 1.900 oznakami zahteva precej predpriprav in predstavlja strmo učno krivuljo za tiste, ki predhodno z oznakami še niso seznanjeni. Njihovo zanesljivost je bilo nato treba preveriti še z večkratnimi oznakami (npr. označevanje enakih besedil v skupinah po 3) in sprejemanjem končnih odločitev.

V vseh naštetih označevalnih kampanjah so bile popravljene posamezne zaporedne pojavnice v besedilu, kar je zlasti pri popravljanju oblikoskladenjskih oznak kognitivno zelo naporno, saj od označevalcev zahteva, da ob vsaki pojavnici mentalno preskakujejo med zelo raznolikimi problemi glede na besedno vrsto. Da bi to breme olajšali, so bili pri označevanju korpusa SentiCoref (Pori in sod., 2022) označevalci razdeljeni v več skupin, vsaka pa je označevala različne besedne vrste.

Končni rezultati najnovejše tovrstne označevalne kampanje v okviru projekta RSDO (Arhar Holdt in sod., 2023) so pokazali, da je učinkovitost strojne lematizacije in oblikoskladenjskega označevanja za slovenščino že dovolj visoka, da je mogoče namesto celostnih ročnih pregledov besedil uporabiti polavtomatske

[4]Oblikoskladenjske oznake Multext East v6: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html

postopke, ki identificirajo najbolj problematična mesta. Poveden je npr. podatek, da je bilo v korpusu SentiCoref popravljenih le približno $1,3$ % vseh pojavnic v korpusu (kar je v skladu s pričakovano točnostjo lematizacijskega modela), od strojno pripisanih oblikoskladenjskih oznak pa jih je bilo popravljenih $2,9$ %. Glede na analize najpogostejših vrst popravkov približno 25 % popravkov izvira iz problemov ločevanja med občnoimenskostjo oz. lastnoimenskostjo samostalnikov (*Delo* vs. *delo*) in razdvoumljanja enakopisnih oblik (npr. tožilnik in imenovalnik pri neživih samostalnikih moškega spola).

V nadaljevanju zato predstavljamo novo metodo za predpripravo podatkov, ki upošteva zgoraj naštete ugotovitve in pri ročnem označevanju implementira polavtomatske rešitve.

3 METODOLOGIJA

Novi označevalni postopek se opira na Slovenski oblikoslovni leksikon Sloleks; v delu, opisanem v tem prispevku, smo uporabljali različico 3.0 (Čibej in sod., 2022) oz. približno 100.800 iztočnic in njihovih oblik, ki so bile ročno preverjene. Sloleks je največja odprto dostopna strojno berljiva zbirka slovenskih besed, v kateri so za vsako iztočnico (npr. *miza*) naštete njene pregibne oblike (*mize*, *mizi*, *mizo*, ...) in ustrezajoče oblikoskladenjske oznake po sistemu MTE-6 (npr. *Sozei*; samostalnik, občni, ženski spol, ednina, imenovalnik).

Metoda izhaja iz dveh poglavitnih predpostavk: (1) da strojno pripisanih lem in oznak pri določenih pojavnicah v korpusu ni treba pregledovati, saj imajo glede na leksikon enoumne oznake in leme; (2) da je pri nekaterih pojavnicah treba pregledati le leme ali pa samo oblikoskladenjske oznake, izbira potencialnih pripisanih vrednosti pa je glede na leksikon omejena. Metoda zato vsako pojavnico v korpusu navzkrižno primerja z oblikami v Sloleksu in preveri, (a) ali je oblika prisotna v leksikonu; (b) ali analizirani obliki v leksikonu pripada ena sama lema ali več; (c) ali je kombinaciji oblike in leme na podlagi leksikona mogoče pripisati nedvoumno oznako ali pa je možnosti več. Na podlagi ugotovljenih značilnosti algoritem pojavnici pripiše ustrezen označevalni scenarij iz nabora, ki ga prikazuje Tabela 1.

Tabela 1: Označevalni scenariji.

Scenarij	*Opis*	*Primer*
1.1.1	ena oblika, ena lema, ena oznaka	zdaj – zdaj – Rsn
1.1.2	ena oblika, ena lema, več možnih oznak	slik – slika – Sozdr|Sozmr
1.2	ena oblika, več možnih lem	lahko – lahek|lahko
1.2.1	ena oblika, razdvoumljena lema, ena oznaka	lahko – lahko – Rsn
1.2.2	ena oblika, razdvoumljena lema, več možnih oznak	lahko – lahek – Ppnzet|Ppnzeo|Ppnsei|Ppnset
2.1	oblike ni v leksikonu, lema pa je	/
2.2	oblike in leme ni v Sloleksu, potreben je ročen popravek	hozentregerji
0	neuvrščena pojavnica	npr. ločila

V scenarij 1.1.1 spadajo pojavnice, ki imajo v leksikonu le eno obliko z nedvoumno lemo in eno nedvoumno oblikoskladenjsko oznako (npr. oblika *zdaj* se v leksikonu pojavi le pod lemo zdaj in le z oznako *Rsn*). Pri scenariju 1.1.2 je kombinacija leme in oblike nedvoumna, razdvoumiti pa je treba oblikoskladenjsko oznako (npr. oblika *slik* nedvoumno spada pod lemo *slika*, a lahko izraža rodilnik dvojine ali pa rodilnik množine). Scenarij 1.2 vsebuje pojavnice, pri katerih je treba najprej razdvoumiti lemo in pozneje še oblikoskladenjsko oznako; scenarij 1.2.2 (ki je eno od nadaljevanj scenarija 1.2) zajema pojavnice, pri katerih je bila lema razdvoumljena, enako kot pri 1.1.2 pa ima kombinacija oblike in razdvoumljene leme lahko več oblikoskladenjskih oznak (npr. oblika *lahko* je lematizirana bodisi kot *lahek* bodisi kot *lahko*; kot pridevnik pa ima lahko glede na leksikon štiri različne oznake (*Ppnzet*, *Ppnzeo*, *Ppnsei*, *Ppnset*). Po scenariju 2.1 oblika manjka v leksikonu, pripisana lema pa obstaja (npr. če gre za zatipkano besedo ali pa legitimno varianto, ki še ni zabeležena v leksikonu). Scenarij 2.2 je edini, ki ga je v celoti treba popraviti ročno, saj v leksikonu še ni oblike in leme. Scenarij 0 vsebuje pojavnice, ki jih ni treba ročno označevati (npr. ločila).

Poleg naštetih označevalnih scenarijev je treba omeniti še podscenarije pri kategorijah 1.1.1 in 1.1.2. V vsaki sta namreč še dodatni podkategoriji M (ang. *mismatch*) in L (ang. *lowercase*), npr. 1.1.1.M, 1.1.1.L, 1.1.2.M itn.

Sozer | Sozmi | Sozmt	... , da preneha, da pač iz svoje	**diete**	Sozer	izloči meso.
Sozer | Sozmi | Sozmt	... veganstvu, vegani ne, kar se tiče	**prehrane**	Sozer	, se pravi, ne jejo, e ...
Sozer | Sozmi | Sozmt	... en majhni hobit, ki potuje skozi te	**dežele**	Sozmt	in nosi, e, prstan v Goro ...
Sozer | Sozmi | Sozmt	... bombardirali, ker je to glavna povezava železniške	**proge**	Sozer	Ljubljana-Trst.

Slika 1: Primer označevalnih nalog iz sklopa 1.1.2 (razločevanje sklona in števila pri samostalnikih ženskega spola).

Podkategorija L je glede na pogoje enaka krovni kategoriji, le da pri navzkrižnem primerjanju s Sloleksom upošteva obliko z malimi tiskanimi črkami: to je predvsem koristno za besede na začetku povedi ali izjave, katerih oblike zaradi zapisa z veliko začetnico ni mogoče neposredno najti v leksikonu.

Podkategorija M označuje primere, pri katerih je kombinaciji oblike in leme pripisana oblikoskladenjska oznaka, ki zanju v leksikonu ni predvidena. To se npr. zgodi v primerih, ko je označevalnik pripisal oznako, ki je ni v leksikonu - tak primer je npr. *samo*, ki je v Sloleksu 3.0 naveden le kot členek (L), pojavlja pa se tudi kot veznik (Vp). Podkategorija M je koristna tudi za vmesno preverjanje ustreznosti oznak - če npr. označevalci med fazo popravljanja leme spremenijo lemo iz prislovne (*odlično*) v pridevniško (*odličen*) in se strojno pripisana prislovna oznaka ne sklada s predvidenimi pridevniškimi v leksikonu. To bodisi opozarja na neustrezno izbrano oznako ali pa na pomanjkljivost v leksikonu.

Pojavnice je glede na označevalne scenarije mogoče smiselno razdeliti v različno zahtevne naloge, znotraj posameznega scenarija pa naloge razvrstiti po sklopih s podobnimi problemi (npr. glede na to, med katerimi oblikoskladenjskimi oznakami mora označevalec izbirati).

Glede na scenarij se cilji označevalne naloge nekoliko razlikujejo, v splošnem pa ena označevalna naloga po tej metodi zajema eno pojavnico s konkordančnim kontekstom ter potencialne vrednosti, ki jih je pojavnici mogoče pripisati. Slika 1 prikazuje primer sklopa nalog iz scenarija 1.1.2, v katerem mora označevalec določati, ali se samostalniki ženskega spola pojavljajo v rodilniku ednine (*Sozer*), imenovalniku množine (*Sozmi*) ali tožilniku množine (*Sozmt*). Navedene so vse izbire oblikoskladenjskih oznak iz leksikona, ciljna pojavnica pa ima prikazan še levi in desni kontekst. Pri označevanju so bile na voljo tudi nekateri drugi podatki - podrobneje jih predstavljamo v razdelku 4.

Zaradi omejenega obsega označevanja in majhnega števila označevalcev (več o tem v razdelku 4) je označevanje v našem primeru potekalo v okolju Microsoft Excel, v primeru obsežnejše kampanje pa bi bilo za ta namen z vidika uporabniške prijaznosti smiselno razviti vmesnike za označevalne platforme, kot sta npr. PyBossa[5] in LabelStudio.[6] To bi med drugim omogočalo tudi dodatno preverjanje kakovosti s sprotnim preverjanjem veljavnosti oblikoskladenjskih oznak in lem. To preverjanje smo v našem primeru opravili s postprocesiranjem označenih datotek.

3.1 Omejitve in prednosti

Metoda predpostavlja, da je korpus že ustrezno tokeniziran in segmentiran. Ker se pri tovrstnem načinu označevanja osredotočamo na pojavnice, popravljanje tokenizacijskih napak ni zelo uporabniško prijazno (označevalec lahko doda komentar, problem pa nato ročno razreši razsodnik), zato je fazo tokenizacije priporočljivo opraviti že pred razdelitvijo v označevalne scenarije.

V primeru večbesednih enot se lahko zgodi, da se posamezne pojavnice razvrstijo v različne scenarije (npr. *lindy hop*). Če so označevalne naloge razdeljene med različne označevalce, morajo biti na tovrstne primere dodatno pozorni.

Upoštevati je treba tudi, da se pri tej metodi nekatere napake lahko izmuznejo skozi sito: to je zlasti problem v primeru enakopisnic, ki so v leksikonu obravnavane kot nedvoumne, glede na jezikovno rabo v korpusu pa niso. Tak primer je npr. oblika *šalam*, ki je v leksikonu nedvoumna (*šalam* - *šala* - občni samostalnik ženskega spola, množina, dajalnik), v korpusu pa se je pojavila kot samostalnik moškega spola (*... narezano šalamo oz. šalam ...*). Ta pojav je predvidoma redek, z vse boljšo pokritostjo leksikona pa bo v prihodnje še redkejši.

Po drugi strani metoda omogoča, da preskočimo odvečno delo (npr. pregledovanje enoumnih oznak, ki lahko zajemajo tudi petino pojavnic), pri pojavnicah, ki jih je treba pregledati, pa omeji število odločitev (če gre npr. samo za razlikovanje med skloni). Namesto polnih oblikoskladenjskih oznak po sistemu MTE je na ta način mogoče za označevalce izpisati le razločevalne značilnosti (npr. *množina, tožilnik*), za katere ne potrebujejo dolgotrajnega uvajanja, zmanjša pa se tudi potreba po navzkrižnem preverjanju.

[5] PyBossa: https://docs.pybossa.com/
[6] Label Studio: https://labelstud.io/

Na ta način je lažje tudi posodabljanje označevalnih smernic, saj so vsi podobni označevalni problemi že zbrani v sklope, na podlagi katerih je mogoče za določeno dilemo doreči bolj sistematično rešitev.

4 PRIPRAVA PODATKOV IN OZNAČEVANJE

Podatki za učni korpus govorjene slovenščine so bili vzorčeni iz korpusa GOS, in sicer iz različic 1.1 (Zwitter Vitez in sod., 2021) (iz katere je bilo vzorčenih pribl. 40.000 pojavnic) in 2.0 (Zwitter Vitez in sod., 2023) (pribl. 50.000 pojavnic). Ker gre za ročne transkripcije govora, ki so bile ročno segmentirane na izjave in razdeljene na pojavnice, te pa imajo ročno pripisane tudi normalizirane oblike (npr. *pršu* – *prišel*), dodatnih popravkov tokenizacije na tej stopnji nismo pričakovali. Nekaj težav je predstavljala razlika v delitvi na segmente med različnimi deli korpusa GOS. Za razliko od gradiva v različici 1.1, ki je bila segmentirana na semantično relativno zaključene enote, je bil del iz različice 2.0, ki izhaja iz zbirke Artur (Verdonik in sod., 2023), za potrebe razvoja razpoznavalnika govora segmentiran po prozodičnih kriterijih (glede na premore). Takšni segmenti pogosto ne odsevajo koherentnih pomensko zaokroženih enot, širši kontekst izjave pa je nujen za ustrezno oblikoskladenjsko označevanje in lematizacijo. Pred označevanjem smo zato za namene popravljanja lem in oblikoskladenjskih oznak te segmente strojno preporazdelili glede na ločila, ki so bila postavljena med transkripcijo posnetkov, in na ta način poenotili reprezentacijo segmentov med tistimi deli, ki so bili vzorčeni iz različice 1.1, in tistimi iz različice 2.0. Kriteriji vzorčenja in postopek strojne resegmentacije so podrobneje opisani v prispevku (Verdonik in sod., 2024).

Ob pripravi podatkov smo upoštevali, da gre za razliko od predhodnih sorodnih označevalnih kampanj, ki so se osredotočale bodisi na standardno pisno ali pa (nestandardno) spletno slovenščino, pri tej kampanji za označevanje govorjene slovenščine, zato pretekli izsledki niso nujno prenosljivi. Zasnovano metodo za polavtomatsko popravljanje smo zato najprej preizkusili na prvem delu učnega korpusa, ki zajema približno 30.000 pojavnic, ki so bile vključene tudi v skladenjsko označeno odvisnostno drevesnico za slovenščino *Spoken Slovenian UD Treebank* oz. SST (Dobrovoljc in Nivre, 2016) in imajo leme in oblikoskladenjske oznake že ročno popravljene. Razdelitev že ročno označenih pojavnic na označevalne scenarije je bila pomembna predvsem zato, da je razkrila, koliko razhajanj (in predvsem spregledanih napak) bi lahko pričakovali ob označevanju novega gradiva po polavtomatski metodi. Rezultati delitve so prikazani v Tabeli 2.

Tabela 2: Delitev podmnožice SST na označevalne scenarije.

Scenarij	*Pogostost*	*Odstotek*
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 %
Skupaj	28.502	100, 00 %

Problematične so predvsem pojavnice iz kategorije 1.1.1.M, ki so glede na leksikon povsem nedvoumne, v resnici pa niso – teh namreč označevalci ne bi podrobno pregledovali. Nekoliko manj problematičen je scenarij 1.1.2.M (kjer imajo pojavnice nedvoumno lemo in več možnosti oblikoskladenjskih oznak, a prava ni vključena v leksikon). Gre npr. za primere tipa *gremo* v velelniškem naklonu, ki v leksikonu še ni predviden, ali pa medmete, kot je *o* (*"o, to pa ne bo šlo"*), katerega oblika je v leksikonu navedena le kot predlog ali pa kot samostalnik. Vseh tovrstnih problematičnih pojavnic je v podmnožici SST le za $0,4$ %, kar nakazuje, da je metoda dovolj točna, da je z njo mogoče pripraviti podatke tudi za označevanje preostalega dela učnega korpusa.

V Tabeli 3[7] je prikazana delitev pojavnic na scenarije še za preostala vzorca, ki sta bila vključena v učni korpus ROG (V1 – dodatnih 10.000 pojavnic iz različice 1.1 in V2 – 50.000 pojavnic iz različice 2.0).

V natančen ročni pregled so bile vključene vse naloge z izjemo scenarijev 0, 1.1.1 in 1.2.1 (več o tem v razdelku 6). Označevala sta skupno dva označevalca, ki sta sodelovala tudi pri označevalnih kampanjah v okviru projekta RSDO (Arhar Holdt in sod., 2023) in sta bila dobro seznanjena tako z označevalnimi smernicami kot

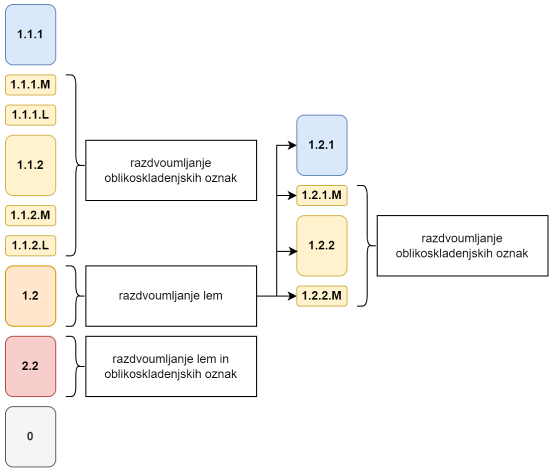
[7] Z *** so označeni nadaljevalni scenariji scenarija 1.2, v katerem najprej razdvoumimo lemo, pojavnice pa nato ponovno razdelimo na nadaljevalne scenarije.

Tabela 3: Delitev ostalih vzorcev na označevalne scenarije.

Scenarij	*Pogostost – V1*	*Delež – V1*	*Pogostost – V2*	*Delež – 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 %
Celota	12.654	100,00 %	48.624	100,00 %

z oblikoskladenjskimi oznakami MTE-6. Prvi označevalec je pregledoval leme, drugi pa oblikoskladenjske oznake (v nekaterih primerih je dodatno popravil tudi leme). Slika 2 prikazuje potek označevanja. Pojavnice iz različnih scenarijev so bile vključene v različne stopnje pregleda; odvisno od scenarija je bila na koncu pregledana le oblikoskladenjska oznaka (npr. 1.1.2), lema (npr. 1.2.1) ali oboje (npr. 2.2).

Pri označevanju so bile na označevalcema v datoteki v pomoč tudi nekateri drugi podatki. Pri vsaki pojavnici, ki jo je bilo treba označiti, sta bila poleg kratkega konteksta (do 5 pojavnic levo in desno, glej Sliko 1) ločeno navedena tudi razširjeni kontekst (celoten segment iz korpusa) ter povezava na ustrezno konkordanco v korpusu Gos 2.1 v konkordančniku NoSketchEngine (Zwitter Vitez in sod., 2023). Za vsako pojavnico so bile dodane še tri povezave do posnetkov: do segmenta, v katerem je pojavnica, ter do predhodnega in naslednjega segmenta. Navedene so bile tudi vse možnosti za razdvoumljanje leme oz. oblikoskladenjske oznake, ki jih predvideva Sloleks. Ohranjen je bil tudi ID pojavnice iz korpusa, s čimer smo poskrbeli za popolno sledljivost sprememb in lažje vključevanje popravkov v končno različico korpusa.



Slika 2: Označevalni delotok za urejanje lem in oblikoskladenjskih oznak.

5 REZULTATI OZNAČEVANJA

V tem razdelku predstavljamo rezultate ročnega popravljanja lem (podrazdelek 5.1) in oblikoskladenjskih oznak (podrazdelek 5.2).

5.1 Popravki lem

Popravki lem so bili redki, saj je bilo na koncu pojavnic s spremenjeno lemo le 396 v vzorcu V2 ($0,81$ % celotnega vzorca) in 175 v vzorcu V1 ($1,38$ % celotnega vzorca).

Glede na porazdelitev popravkov po označevalnih scenarijih so bile spremembe leme najpogostejše pri scenariju 2.2 (42 % vseh popravkov lem), pri katerem niti oblike niti leme še ni v leksikonu, zaradi česar je pričakovano, da bo lematizacijski model na tovrstnih podatkih zagrešil največ napak. V vzorcu V2 je bila lema popravljena pri 164 pojavnicah (od 497 iz scenarija 2.2, torej 33 %), v vzorcu V1 pa pri 73 (od 233 iz scenarija 2.2, torej približno 31 %). V obeh vzorcih je bila torej približno tretjina neznanih pojavnic lematizirana napačno.

Določanje leme je za model problematično zlasti pri pregibnih oblikah lastnih imen (npr. *Netflix* – **Netflixu*, *Šerbi* – **Šerba*, *Lidl* – **Lidel*, *Bohinjka* – **Bohinjek*) ali pa pri pregibnih oblikah samostalnikov, pri katerih pride do podaljšave z -j (*espe* – **espej*, *mikronivo* – **mikronivoj*. Pri neznanih pojavnicah tudi pogosto napačno presodi besedno vrsto in npr. glagol lematizira kot samostalnik (*zmučka* namesto *zmučkati*), prislov ali medmet kot glagol (*tulele* – **tuleti*, *ojojojo* – **ojojati*) ipd. S tega vidika so problematične pojavnice, ki izhajajo iz drugih jezikov in se v slovenščini pregibajo (*sitcom* – **sitec*, *solfeggio* – **solfeggiti*).

Lematizator ima težave tudi z odločanjem med občnoimenskostjo in lastnoimenskostjo (*slofit* – *Slofit*, *kliping* – *Kliping*, *covid* – *Covid*) ter z lematizacijo daljših besed, ki obsegajo 15 znakov ali več, pri katerih se zadnji del leme močno pokvari (*jezikovnotehnološki* – **jezikovokolohoški*, *knjižnojezikosloven* – **knjižnozezozoven*, *prikrojevalnica* – **prikrojalalnica*).

Med scenariji, ki so v leksikonu, so pričakovano najbolj problematične enakopisnice, tj. pojavnice iz scenarija 1.2 in njegovih podscenarijev – ti zajemajo 328 pojavnic (približno 57 % vseh popravkov lem). Med najpogostejšimi popravki so zlasti popravki med pridevniki na eni strani in prislovi na drugi, npr. *mogoč* – *mogoče*, *dober* – *dobro*, podobno tudi *ves* – *vse*, *tak* – *tako*. Z vidika govorjene slovenščine je zanimiv popravek *ti* – *te*, ki se nanaša na štajerski *te* (”*te pa si bil to samo po Sloveniji?*”), ki v leksikonu še ni zabeležen in je bil zato strojno lematiziran kot *ti* ali *ta* ter nato ročno popravljen v *te*.

V scenariju 1.1.2, pri katerem je bilo treba razdvoumljati oblikoskadenjske oznake, je bilo opravljenih le 6 popravkov lem, kar nakazuje, da je ločevanje razdvoumljanja lem in oblikoskladenjskih oznak smiselno.

5.2 Popravki oblikoskladenjskih oznak

Popravki oblikoskladenjskih oznak so bili nekoliko pogostejši kot pri lemah, a še vedno zajemajo manjšino pojavnic. V vzorcu V2 je bila oznaka spremenjena le pri 2.029 pojavnicah ($4,17$ % celotnega vzorca), v vzorcu V1 pa pri 627 pojavnicah ($4,95$ % vzorca).

Po pričakovanjih je bilo 1.782 popravkov ($67,09$ % vseh popravkov oznak) opravljenih znotraj scenarija 1.1.2 (vključno z 1.1.2.M in 1.1.2.L), pri katerem gre za razdvoumljanje slovnično enakopisnih oblik z nedvoumno lemo. Pričakovanih je tudi 578 popravkov ($21,76$ %) iz scenarija 1.2 in podscenarijev, kjer popravek leme pogosto zahteva tudi popravek oznake. Čeprav je bila v scenarij 2.2 (neleksikonske pojavnice) zajeta le manjšina popravkov (296 pojavnic oz.

$11,15$ %), pa analiza deleža popravljenih pojavnic znotraj scenarija 2.2 pokaže, da je bilo v vzorcu V2 strojno napačno označenih $37,83$ % pojavnic, v vzorcu V1 pa $46,35$ % pojavnic. Pri ostalih scenarijih je bil ta delež mnogo manjši, le okrog 7 %, kar poudarja pomen ustrezno posodobljenega leksikona za uspešno oblikoskladenjsko označevanje.

V Tabeli 4 so po pogostosti razvrščene oblikoskladenjske značilnosti strojno označenih pojavnic, pri katerih je bilo treba najpogosteje popraviti oblikoskladenjsko oznako. Po pogostosti so na prvem mestu splošni pridevniki, velja pa opazovati predvsem delež popravljenih pojavnic znotraj kategorije – v tem primeru so med najbolj problematičnimi lastnoimenski samostalniki moškega spola, pri katerih je bilo treba popraviti kar četrtino vseh pojavnic. Podobno tudi z glavnimi besednimi števniki in vprašalnimi zaimki. Zanimivo je, da so pri strojnem označevanju skoraj povsem neproblematični glagoli, pri katerih je bilo popravkov v vseh kategorijah (nedovršni, dovršni, dvovidski, pomožni) skupaj le 84, med $0,5$ in $1,3$ %.

Tabela 4: Oblikoskladenjske značilnosti najpogosteje popravljenih pojavnic (s frekvenco nad 100).

Značilnosti	*Popravljeno*	*Vse pojavnice*	*Delež*
Pp (pridevnik, splošni)	384	2.998	12,81 %
Som (samostalnik, občni, moški)	281	3.412	8,24 %
Soz (samostalnik, občni, ženski)	267	3.287	8,12 %
Rs (prislov, splošni)	261	5.103	5,11 %
Zk (zaimek, kazalni)	215	1.860	11,56 %
Zo (zaimek, osebni)	140	1.341	10,44 %
Slm (samostalnik, lastni, moški)	122	473	25,79 %
Sos (samostalnik, občni, srednji)	110	1.361	8,08 %
Kbg (števnik, besedni, glavni)	109	486	22,43 %
Vp (veznik, priredni)	106	3.265	3,25 %
Zv (zaimek, vprašalni)	103	497	20,72 %

V Tabeli 5 so prikazani najpogostejši popravki oblikoskladenjskih značilnosti (s frekvenco vsaj 50). Ti zajemajo več kot polovico vseh popravkov (53 %), skoraj tretjina (28 %) pa je zgolj razlikovanja med imenovalnikom in tožilnikom.

Tabela 5: Najpogostejši popravki oblikoskladenjskih značilnosti (s frekvenco vsaj 50).

Popravek	*Frekvenca*	*Delež*	*Primeri*
imenovalnik, tožilnik	561	21,12 %	Somei → Sometn (*stol*), Kbg-mi → Kbg-mt (*tisoč*), Zk-mei → Zk-met (*ta*)
tožilnik, imenovalnik	190	7,15 %	Sometn → Somei (*video*), Zk-set → Zk-sei (*tisto*), Kbg-mt → Kbg-mi (*devetsto*)
prislov, členek	136	5,12 %	Rsn → L (*a*)
moški, ženski	122	4,59 %	Zotmmt–k → Zotzmt–k (*jih*), Ppnmmr → Ppnzmr (*naslednjih*)
imenovalnik množine, rodilnik ednine	82	3,09 %	Sozmi → Sozer (*preiskave*), Ppnzmi → Ppnzer (*radijske*), Sosmi → Soser (*zdravila*)
splošni pridevnik, splošni prislov	80	3,01 %	Ppnsei → Rsn (*mogoče*), Ppnzet → Rsn (*primerno*)
moški, srednji	67	2,52 %	Zotmet–k → Zotset–k (*ga*), Ppnmeo → Ppnseo (*zdravim*), Kbvmei → Kbvsei (*devetnajststo*)
priredni veznik, splošni prislov	64	2,41 %	Vp → Rsn (*zato*)
občni, lastni	55	2,07 %	Somei → Slmei (*Piano*), Somem → Slmem (*Lidlu*), Sozer → Slzer (*Jute*)
vprašalni zaimek, splošni prislov	50	1,88 %	Zv-sei → Rsn (*kako*), Zv-set → Rsn (*kaj*)

6 OZNAČEVALNE DILEME

Dileme, ki so se pojavile pri ročnem pregledu korpusa ROG, so pričakovano izhajale iz razlik med govorjenim in pisnim standardnim jezikom, označevalna kampanja v tem prispevku pa omogoča prvi sistematični popis lematizacijskih in oblikoskladenjskih problemov, na katere naletimo v govorjeni slovenščini. V grobem jih lahko razdelimo na tri poglavitne skupine:

(a) Težko določljiva kanonična oblika: kot že omenjeno, v govorjeni slovenščini naletimo na tipično govorjeno besedišče, ki se v standardnem pisnem jeziku ne pojavlja, zato tudi ni opisano v obstoječih leksikografskih virih in kot tako nima enovite standardne osnovne oblike. Pri nekaterih dileme ni (npr. *mezmes*), pri drugih pa njihova izgovorjava dopušča več zapisov, npr. *oreng*, *orenk*, *orng*, *ornk*; *gravžati*, *graužati*; *hozentregar*, *hozentreger*, *hozntreger* itn. Poleg tega se v korpusu pojavljajo tudi dialektalne variante iste neuslovarjene besede, kar določanje leme še otežuje.

(b) Težko določljiva oblikoskladenjska oznaka: izkazalo se je, da v govorjenem jeziku veliko sicer standardnih besed zavzema drugačen skladenjski položaj kot v pisnem, kar postavlja pod vprašaj ohranitev oblikoskladenjske oznake, tipične za pisna besedila, npr: *a* v diskurznih markerjih 'a ne', 'a to' ("*... da so se imeli na koga obrniti, a ne.*"), kjer se odločamo med oznakama za priredni veznik in členek, in *ali* v zaključku vprašalne povedi ("*ja fajn a boste peli pri maši tudi ali?*") – tu se glede na leksikon odločamo med prirednim veznikom in prislovom. Potem so še besede, ki so tipične za govor in jih v standardnih pisnih besedilih niti ne zasledimo, npr.: nesklonljivi *ta* kot podkrepitev pridevniške besede ("*... glejte, mi eee vidimo, da prav ta pravega vira ...*"), ki ima prekrivno lemo z zaimkom 'ta', vendar ga njegova nesklonljivost razmejuje od zaimkov in bi ga lahko upravičeno uvrstili med členke; tipično štajerski nesklonljiv *te*, ki lahko nadomešča prislov 'takrat' ("*zato ker te komaj ceniš suho cesto, ko se moraš ...*") ali pa je mašilo ("*pa sva se zadnjič komaj te končno odločili, kateri film bova gledale.*"; *en* oz. *ene* v pomenu 'približno', ki sicer deloma morfološko posnema zaimkovni števnik 'en', toda zavzema zanj netipično skladenjsko vlogo: vedno določa števnik ali merni prislov ("*... Bi kar ene štiri vzel*"); samo v vlogi veznika, npr. v transkripciji "*glasbeno šolo sem naredil pet let, samo zadnjega letnika nisem, samo teorijo sem delal ...*", kjer bi vsaj eno pojavitev 'samo' lahko obravnavali kot protivni veznik – katero, ni jasno niti iz posnetka govora.

Nenazadnje sem spadajo še tuje besede, ki so se pokazale kot problematične tako za lematizacijo kot za oblikoskladenjsko označevanje že pri označevanju SUK 1.0 (Arhar Holdt in sod., 2023). Čeprav je bil ta izziv deloma razrešen, so za celovito rešitev potrebne podrobnejše analize tovrstnih besed v slovenščini. Teh besed je namreč v nadgradnjah predvsem govornih korpusov pričakovati več.

(c) Izmuzljive oblike: nekatere oblike pomotoma padejo v scenarij 1.1.1 in tako uidejo ročnemu pregledu. Do tega najpogosteje pride zaradi napak v transkripciji, npr. *uče* namesto 'uče…' (stično tropičje zaznamuje nedokončano besedo), kar v skladu s Sloleksom dobi oznako za glagol (*učiti* – *uče*: sedanjik, tretja oseba, množina), in *ke*, ki je lahko zatipk za 'ker' ali pa ni ustrezno normalizirano v 'ki' ali 'kaj'; lahko pa oblika pristane v tem scenariju tudi zaradi pomanjkljivosti Sloleksa, npr: *šalam*, ki se v Sloleksu pojavi le kot oblika iztočnice 'šala' v dajalniku množine, nima pa iztočnice 'šalam', kar je mišljeno npr. v "*eem, popečem na trakce narezano š… pač šalamo oziroma šalam.*"Tovrstne napake so praviloma redke, poudariti pa je treba tudi, da se s posodabljanjem leksikona vse tovrstne dileme v prihodnjih označevalnih kampanjah znajdejo v drugih scenarijih, saj npr. z dodajanjem iztočnice *šalam* v oblikoslovni leksikon oblika postane dvoumna na ravni leme, zato spada v scenarij 1.2, ne več v 1.1.1.

7 ZAKLJUČEK

V prispevku smo predstavili označevanje učnega korpusa govorjene slovenščine ROG na ravni oblikoskladenjskih oznak in lem z novo polavtomatsko metodo ter opravili prvi popis dilem. Rezultati so spodbudni zlasti ob primerjavi s pričakovanim časovnim obsegom označevanja po povsem ročni metodi, ki je bila uporabljena npr. pri označevanju korpusa SUK 1.0: glede na pretekle izkušnje namreč označevanje lem in oblikoskladenjskih oznak za vsako pojavnico vzame približno 12 sekund. V našem primeru bi pri približno 60.000 pojavnicah vzorca, skupini 6 označevalcev, 3 zbranih odgovorih na pojavnico in 10-urno tedensko kvoto kampanja trajala približno 9-10 tednov, skupaj 500 ur študentskega dela (oz. 160 ur, če bi zbirali le po en odgovor na pojavnico), pri čemer ni všteto še delo koordinatorjev in tehnične podpore. Za označevanje korpusa ROG smo potrebovali skupaj 105 ur (25 ur za leme in 80 ur za oblikoskladenjske oznake), končni delež popravljenih pojavnic pa je primerljiv.

Metodo se lahko v prihodnje uporabi tudi za iskanje nekonsistentnosti v predhodno označenih korpusih, kot je SUK 1.0. Scenarije je mogoče spremljati tudi po posodobitvi različic leksikona, saj morebitne spremembe nakažejo potencialno nekonsistentnost v oznakah. Scenariji bi morda lahko bili uporabni tudi za natančnejšo kategorizacijo napak, ki se lahko uporabijo kot potencialne uteži za natančnejše evalvacije lematizacijskega in oblikoskladenjskega modela (napaka v sklonu je npr. manj resna od napake v besedni vrsti).

Smiselno bi bilo razmisliti tudi o vzporednem posodabljanju leksikona in učnega korpusa in obratno, zato da preverimo, ali se pojavnice v korpusu še vedno ujemajo s stanjem v leksikonu. Posodabljanje leksikona se je izkazalo za pomembno nalogo za nadaljnje označevanje, zlasti pri kanoničnih oblikah tipično govorjenega besedišča, ki se v pisni (standardni) obliki ne pojavlja (*šravf*, *šrauf*; *orng*, *ornk*, *oreng*). To bi pripomoglo tudi k večji konsistentnosti transkripcij govora. Dileme, ki so bile identificirane med označevanjem korpusa ROG, bodo natančneje opisane v smernicah za vključevanje tipično govorjenega besedišča v digitalne jezikovne vire, kar je prav tako eden od ciljev projekta MEZZANINE (uvodni načrti za smernice so bili že predstavljeni v prispevku (Čibej in sod., 2024)).

V prihodnje bi veljalo opraviti tudi natančnejši popis in raziskavo najbolj problematičnih pojavnic za avtomatsko označevanje (tudi na podlagi učnega korpusa SUK 1.0) oz. izdelati seznam, katere pojavnice so načeloma neproblematične kljub morebitni enakopisnosti v leksikonu (npr. *kaj* vs. *kaja*, starinska beseda za kajenje).

Po zaključku označevanja na različnih ravneh bo Učni korpus ROG na voljo pod odprto licenco na repozitoriju CLARIN.SI.

ZAHVALA

Prispevek je nastal v okviru raziskovalnega projekta *Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik* (MEZZANINE, J7-4642), raziskovalnega projekta *Na drevesnici temelječ pristop k raziskavam govorjene slovenščine* (SPOT, Z6-4617) in raziskovalnega programa *Jezikovni viri in tehnologije za slovenski jezik* (P6-0411), ki jih financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS).

Za pregled lem se avtorja iskreno zahvaljujeta Matiju Škofljancu, za dodatne predloge pri zasnovi polavtomatske metode dr. Kaji Dobrovoljc. Iskrena hvala tudi anonimnim recenzentom_kam za konstruktivne pripombe.

LITERATURA

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Erjavec, T., Gantar, P., Krek, S., … Žitnik, S. (2023). Nadgradnja učnega korpusa ssj550k v suk 1.0. *Razvoj slovenščine v digitalnem okolju*, 119–156.

Čibej, J., Arhar Holdt, Š., Fišer, D. in Erjavec, T. (2018). Ročno označeni korpusi janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave. V *Viri, orodja in metode za analizo spletne slovenščine* (str. 44–73). https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/111/203/2416.

Čibej, J., Fišer, D. in Erjavec, T. (2016). Normalisation, tokenisation and sentence segmentation of slovene tweets. V *Normalisation and analysis of social media texts (normsome) - lrec 2016* (str. 5–10). Portorož, Slovenia. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NormSoMe_Proceedings.pdf#page=10

Čibej, J., Gantar, K., Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., … Robnik-Šikonja, M. (2022). *Morphological lexicon sloleks 3.0.* http://hdl.handle.net/11356/1745 (Slovenian language resource repository CLARIN.SI)

Čibej, J., Robida, N. in Krek, S. (2024). Nadgradnja digitalne slovarske baze za slovenščino in slovenskega oblikoslovnega leksikona sloleks s podatki o govorjeni slovenščini: načrti in cilji. *Stanje in perspektive uporabe govornih virov v raziskavah govora*, 27–40.

Dobrovoljc, K. (2024). Skladenjska drevesnica govorjene slovenščine: stanje in perspektive. *Stanje in perspektive uporabe govornih virov v raziskavah govora*, 41–62.

Dobrovoljc, K. in Nivre, J. (2016, May). The Universal Dependencies treebank of spoken Slovenian. V N. Calzolari in sod. (Ur.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (str. 1566–1573). Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1248

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A. in Biemann, C. (2016, December). A web-based tool for the integrated annotation of semantic and syntactic structures. V *Proceedings of the workshop on language technology resources and tools for digital humanities (LT4DH)* (str. 76–84). Osaka, Japan: The COLING 2016 Organizing Committee. https://www.aclweb.org/anthology/W16-4011

Erjavec, T., Fišer, D., Čibej, J. in Arhar Holdt, Š. (2016a). *CMC training corpus janes-norm 1.2.* http://hdl.handle.net/11356/1084 (Slovenian language resource repository CLARIN.SI)

Erjavec, T., Fišer, D., Čibej, J. in Arhar Holdt, Š. (2016b). *CMC training corpus janes-tag*

1.1. http://hdl.handle.net/11356/1081 (Slovenian language resource repository CLARIN.SI)

Fišer, D., Ljubešić, N. in Erjavec, T. (2020). The janes project: language resources and tools for slovene user generated content. *Language Resources Evaluation*, *54*, 223–246. https://doi.org/10.1007/s10579-018-9425-z

Pori, E., Čibej, J., Munda, T., Terčon, L. in Arhar Holdt, Š. (2022). Lematizacija in oblikoskladenjsko označevanje korpusa senticoref. V *Konferenca jezikovne tehnologije in digitalna humanistika* (str. 162–168). Ljubljana, Slovenija. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Pori-et-al_Lematizacija-in-oblikoskladenjsko-oznacevanje-korpusa-SentiCoref.pdf

Verdonik, D., Bizjak, A., Sepesy Maučec, M., Gril, L., Dobrišek, S., Križaj, J., ... Dretnik, N. (2023). *ASR database ARTUR 1.0 (transcriptions).* http://hdl.handle.net/11356/1772 (Slovenian language resource repository CLARIN.SI)

Verdonik, D., Ljubešić, N., Rupnik, P., Dobrovoljc, K. in Čibej, J. (2024). Izbor in urejanje gradiv za učni korpus govorjene slovenščine rog. V *Konferenca jezikovne tehnologije in digitalna humanistika.* Ljubljana, Slovenija.

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M. in Erjavec, T. (2021). *Spoken corpus gos 1.1.* http://hdl.handle.net/11356/1438 (Slovenian language resource repository CLARIN.SI)

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., Verdonik, D., ... Dobrovoljc, K. (2023). *Spoken corpus gos 2.0 (transcriptions).* http://hdl.handle.net/11356/1771 (Slovenian language resource repository CLARIN.SI)

A METHOD FOR SEMI-AUTOMATIC CORRECTIONS OF LEMMAS AND MORPHOSYNTACTIC TAGS: THE CASE OF THE ROG TRAINING CORPUS OF SPOKEN SLOVENE

In the paper, we present the process of correcting lemmatization and morphosyntactic tags in the ROG Training Corpus of Spoken Slovene, sampled from the GOS Corpus of Spoken Slovene (versions 1.1 and 2.0). Corrections of lemmas and morphosyntactic tags were conducted in several phases and, unlike similar annotation campaigns for Slovene, included an additional preprocessing phase in which lemmas and morphosyntactic tags were automatically cross-referenced with the forms included in the Sloleks Morphological Lexicon of Slovene. This new method has significantly sped up manual work as well as reduced the number of redundant checks and final costs. Its advantage is also the fact that annotation tasks are divided into batches of similar problems (e.g. discriminating between the nominative and accusative case). With adequate data preparation, this method in a significant number of examples requires no knowledge of MTE-6 morphosyntactic tags. In addition to the results of the annotation, we also present the principal dilemmas encountered when annotating spoken Slovene.

Keywords: lemmatization, morphosyntactic tagging, spoken Slovene, spoken Slovene corpora

TWITTER BEFORE X: SCENES FROM THE BALKANS

Filip Dobranić,[1] Nikola Ljubešić[1, 2]
[1]Institute for Contemporary History
[2]Department of Knowledge Technologies, Jožef Stefan Institute

We present a corpus of over 170 million Twitter posts in Slovenian, Croatian, Bosnian, Serbian and Montenegrin, collected between 2017 and 2023. After describing the data collection, our focus moves to the challenges of the division and linguistic processing of the data given that we deal with an area as ethnically and linguistically interconnected as the Balkans. An exploratory analysis into the quality, quantity and differences between the collections provides evidence of real-life events' effect on online user-generated content production, most notably measures mitigating the spread of COVID-19 and elections. We investigate the use of emotionally charged words through time and provide a corpora-wide overview of the most prolific authors, hashtags, and mentions before concluding the paper with an invitation for further research into this vast and diverse contemporary-historical corpus of online speech.

Keywords: Twitter corpus, Slovenian, HBS macro-language, distant reading

1 INTRODUCTION

Social media platforms have become an integral part of modern communication, providing a vast source of data for researchers to study language use and societal trends. While collections of tweets have been used for various tasks, including forecasting mental illness among users (Reece et al., 2017), the majority of research focuses on Twitter's interplay with (inter)national politics (Karami et al., 2020). The wide range of uses of Twitter corpora highlight not only their importance for understanding contemporary politics ranging from political campaigns by institutional actors to the organisation of grass roots protests (Bronstein et al., 2018; Flew & Iosifidis, 2020; Effing et al., 2011; Chan & Yi, 2024; Altman, 2018) but also their value when it comes to understanding other societal structures such as traditional media (Zhao et al., 2011) and studying human affect and social media behavior more broadly (Wang et al., 2020; Y. Qi & Shabrina, 2023).

In this paper, we present a collection of corpora of tweets written in Slovenian and any of the languages of the ISO-639-3 HBS macro-language (the macro-language including Croatian, Bosnian, Montenegrin and Serbian) tweets collected between the years 2017 and 2023. The corpora are the result of processing $172,806,656$ tweets produced by $2,959,373$ authors, containing $3,418,701$ hashtags, and $5,405,589$ mentions.

The paper is structured as follows. In the following section we outline the process of data collection, filtering based on language (2.1), redaction (2.2), and linguistic annotation (2.3) of the tweets. We follow up by discussing the split of our collection into three corpora (2.4) and outline some of the optimisations to speed up processing (2.5). Next, we present the corpora structure in Section 3, providing a short distant reading and analysis of the corpora before concluding with an invitation for future research and use of the corpora.

2 CORPUS CREATION

The data collection took place between the years 2017 and 2023. The tweets were collected with the TweetCaT methodology as outlined in (Ljubešić et al., 2014) using seed terms to identify users tweeting in the desired language, then storing all the tweets by identified users in JSON files. Two collection processes took place simultaneously: one with Slovenian seed words and one with seed words in HBS languages. After the collection period came to a close, the tweets were filtered through a secondary sieve to exclude tweets and retweets in languages we were not interested in. While a user might predominantly tweet in their preferred (often native) language, it is quite common for people to share content and converse with other users on the internet in various languages.

2.1 Language filtering

After collecting the tweets based on our prediction of users' native language, we had to filter out the tweets and retweets that contained either no recognisable language or were predominantly in a language we were not interested in. For the Slovenian corpus, we relied on Twitter's own language identification. The

API responses contained a key with the predicted language, and we used that to discard any tweet not marked as Slovenian by Twitter.

For HBS languages we found Twitter's labelling lacking. While there were few false positives, our initial investigation found a significant amount of tweets written in HBS languages that were mislabelled or labelled as unknown. In order to address the shortcomings of Twitter's language identification, we used FastText (Bojanowski et al., 2017) to determine the tweets' language. Our goal was to find tweets in any of the HBS languages. We hypothesised that if FastText gives multiple low-confidence predictions of HBS languages we can treat that as a case of a relatively high confidence prediction that it is written in at least one of the languages.

In order to confirm our hypothesis we annotated a random sample of 500 tweets and tested FastText's language identification, changing two defined hyper-parameters of the classifier (number of guesses and confidence threshold). The tests confirmed the viability of our approach (multiple low-confidence predictions were a good signal for the tweet being in one of the HBS languages), but showed that the recall and accuracy of our approach depends on hyper-parameters used with FastText: the number of guesses and required confidence threshold.

Consequently, we used our set of 500 annotated tweets to tune the two hyper-parameters for the language classifier. We then investigated the results in bands based on their performance characteristics. You can consult our code at https://dihur.si/muki/twitter for specifics. After investigating hyper-parameter configurations based on their precision and recall on the annotated dataset, we opted for the hyper-parameter configuration ensuring the highest recall, since a relatively marginal maximum gain in precision would mean 30% fewer tweets in the corpus. We prioritised more tweets in our corpora even if they contain a slightly larger amount of tweets not in the target languages as opposed to a more narrow corpus with higher language guarantees. For further research requiring stricter limits on language presence, researchers can always further filter the current tweets to a degree appropriate for their research.

2.2 Redaction

Twitter's API responses present a lot of structured information we are not necessarily interested in, e.g. coordinates, favorite count at time of retrieval, sensitive content flag, extracted entities, user metadata etc. We opted to redact this information both out of privacy concerns as well as a mindfulness of the corpus' final size and resources required to process it.

While redacting, we used the opportunity to unify the structure of the data. Since the collection took place on such a large time scale, the API responses themselves changed over time. The most notable of these is the key containing the tweet's text, which changed from *full_text* to *text*, along with at least on other relevant property, *truncated*, denoting responses that did not contain the full tweet text. In our redacted data structure we assign the tweet's text to the key *text* and retain the *truncated* boolean flag. The structure of our redacted tweet data follows:

- *created_at* ($string$) retains the timestamp of tweet creation provided by the API
- *truncated* ($boolean$) marks truncated tweets with *true*
- *user_screen_name* ($string$)
- *text* ($string$)
- *id_str* ($string$) retains the tweet's ID number as string
- *is_retweet* ($boolean$) marks tweets that are retweets with *true*
- *source_tweet_id_str* ($string|null$) if the tweet was a retweet, this contains the id of the original tweet, otherwise null

2.3 Linguistic annotations

Once the tweets were redacted, we proceeded to linguistically annotate them. We performed the annotation automatically with the CLASSLA-Stanza pipeline (Terčon & Ljubešić, 2023), a fork of the Stanford Stanza pipeline (P. Qi et al., 2020). We prefer the CLASSLA pipeline over Stanza since the former's models are based on a larger training dataset, use large inflectional lexicons, support both standard and Internet-non-standard language, and have support for Named Entity Recognition (NER). Since language captured on the internet often differs from the formal version (e.g. there is no *@mention* in Slovenian), we used the non-standard pipelines by applying the *type=”nonstandard”* argument in the pipeline.

2.4 Language considerations and the splitting of the HBS corpus

A regular stumbling block when dealing with data in the HBS macro-language, especially when the data were collected in the wild, either on the web or from social media, is whether, and if so, how to further divide content written in the underlying languages. Given the shortness of messages on Twitter, it is a proper challenge to perform a reasonable automated job on this task (Ljubešić & Kranjčić, 2014). Most proposed solutions are based on machine learning on available data, which are known to overfit to the training data, giving perfect results if the test data are similar, but far from perfect results otherwise (Rupnik et al., 2023).

The CLASSLA-Stanza pipeline does not have an HBS pipeline, but either a Croatian or a Serbian pipeline. The biggest difference between these two pipelines, or rather, the data these pipelines were trained on, are that the Serbian pipeline was trained on ekavian data (*lepo*, *beži*), while the Croatian pipeline was trained on standard and non-standard ijekavian data (*lijepo*, *bježi*). Furthermore, the Croatian non-standard data cover most phenomena specific to the Bosnian and Montenegrin language, such as the synthetic future tense (*smetaću* vs. *smetat ću* in standard Croatian), or the usage of both *što* and *šta* pronouns (standard Croatian allows only the former), etc. The Croatian non-standard processing pipeline is capable of dealing with most of the lexical differences between Croatian, Bosnian and Montenegrin (in Croatian verbs ending in *-irati* mostly end in the other langauges in *-isati* and *-ovati* etc.) due to static embeddings used in the pipeline that have been trained on web data, where most variants can be found.

Important to note is also that the Croatian pipeline lemmatizes into the ijekavian variant, while the Serbian pipeline lemmatizes into the ekavian variant.

Given the above described situation, we have decided not to continue discriminating between the four languages contained in the HBS macro-language as such a division would result in significant errors, but rather to be application-oriented and follow the division by the most prevalent linguistic difference between the different standards in the macro-language, which is also reflected in the CLASSLA-Stanza processing pipeline, namely the ekavian vs. ijekavian variant. For that reason we have used a lexicon-based approach from our previous work (Ljubešić et al., 2018) to classify users either as using the ekavian or the ijekavian variant of the HBS macro-language, constructing two separate corpora called HBS-ekavian and HBS-ijekavian. Each corpus was processed linguistically with the corresponding pipeline, the HBS-ekavian corpus with the Serbian pipeline, and the HBS-ijekavian corpus with the Croatian pipeline.

Since the CLASSLA-Stanza pipeline is capable of processing the Latin script only for both the HBS-ekavian and HBS-ijekavian corpus, we used the *cyrtranslit* (Labrèche, 2023) transliteration tool to transform tweets from the Cyrillic into the Latin script. We retained the original text for future research endeavours and annotated all the tweets in our collection with a boolean value denoting whether or not the tweet was transliterated. Tweets that had more than 20% of their characters modified after transliteration had the transliterated flag set to true (the rest have the flag set to false) to allow for easier filtering of the most heavily transformed tweets.

2.5 Optimisations for linguistic annotations with CLASSLA-Stanza

While individual tweets technically represent individual "documents" in our corpora, due to CLASSLA's startup time and the tweets' relatively short length, processing each and every one of them individually would take a prohibitively long time. To mitigate this, we collect all the tweets published in a single day, join their texts with "\n|\n", process the day's worth of tweets as one document, and then split the result at the paragraph containing the pipe character "|". Our ad hoc benchmarks showed a reduction in processing time by a factor of 8.

Since a significant part of the corpora are retweets, we performed another optimisation during our processing. For every day, we begin by only processing original content first, storing retweet IDs along the way. Once we iterate through all the day's tweets and process the originals, we process all unique retweeted tweets, then copy the results to their corresponding retweets. This ensures that we processed a tweet at most twice in a day instead of every time we encounter a retweet.

3 CORPUS OVERVIEW AND ANALYSIS

Finishing the steps outlined above, we ended up with our three linguistically annotated corpora: the Slovenian corpus, the HBS-ijekavian corpus, and the HBS-ekavian corpus. In this section we outline the structure of the corpus and perform some preliminary analysis to showcase the corpus' ability to aid in answering a wide and diverse range of research questions. We begin with an analysis of the corpora as a whole, then investigate specific authors, hashtags, and mentions. Basic size metrics are presented in Table 1.

Table 1: Metadata on each of the three corpora

What	*Slovenian*	*HBS-ijekavian*	*HBS-ekavian*
Total number of tweets	$42,483,342$	$31,199,242$	$99,124,072$
Original tweets	$23,293,074$	$26,975,269$	$64,986,741$
Retweets	$19,190,268$	$4,223,973$	$34,137,331$
Truncated tweets	$1,137,603$	$1,006,906$	$3,123,005$
Authors	$483,216$	$601,282$	$1,874,875$
Hashtags	$1,012,474$	$963,748$	$1,442,479$
Mentions	$736,573$	$1,527,489$	$3,141,527$

3.1 Size and activity

In this section we discuss the size and tweet production through time for the three corpora. While Twitter usage fluctuates on a monthly basis, there are a few notable drops or rises in production visible from this perspective. We discuss these for each of the corpora.

A total of $42,483,342$ tweets were processed to form the Slovenian corpus, of those $23,293,074$ (almost 55%) represent original content, the rest are retweets. It's important to note that this includes only content written in the target language (Slovenian). If a user retweeted content in other languages the posts are ignored. The number of tweets and retweets through time can serve as an approximation of Twitter's relative speed of growth (or contraction), but it does not accurately estimate the absolute quantity of posts produced by Twitter users.

Looking at the data in terms of total Slovenian tweet production, we observe an order of magnitude increase of number of tweets (both original and retwets) in

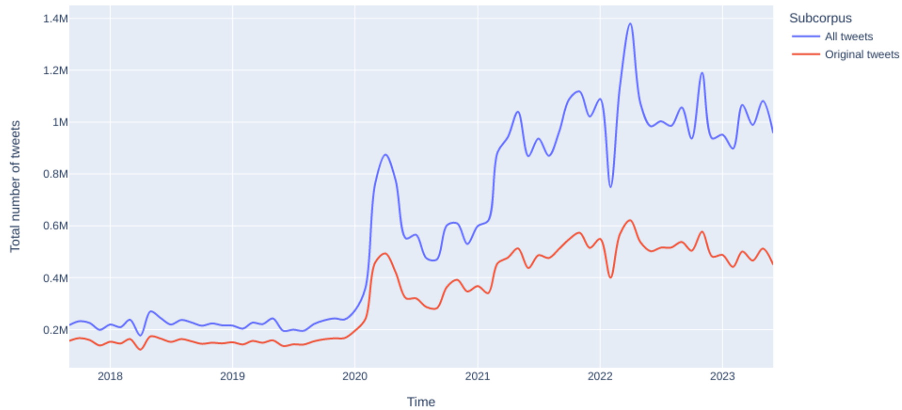


Figure 1: Number of monthly tweets in the Slovenian corpus through time

March 2020, coinciding with the official beginning of the COVID-19 pandemic in Slovenia and first "lockdowns". During the summer months of the same year, as quarantine measures were loosened and people moved back to meatspace, we can observe a dip in posting activity. While remaining relatively low until March 2021 total number of tweets still remains above levels before the COVID-19 pandemic. In March and April 2021 we observe another rise in traffic, pushing posting levels to a record high in April 2022. Coincidentally, this latter month is also the time of Slovenia's parliamentary election, one accompanied by protests and immense engagement by civil society, resulting in one of the highest voter turnouts in the history of the country.

The HBS-ijekavian collection contains $31,199,242$ tweets of which $26,975,269$ (over 86%) are original. The HBS-ekavian collection in turn contains $99,124,072$ tweets with $64,986,741$ (a bit under two thirds) original.

We observe a similar increase in tweet production during March and April 2020 as we do in the Slovenian corpus. The corpora differ from the Slovenian in that we do not observe a drop in production in subsequent months, instead we see a relatively steady growth in the amount of tweets that tapers out in 2021 and starts dropping after that. COVID-19 and measures to prevent its spread seemed to have an effect on posting activities of HBS-ijekavian and HBS-ekavian Twitter public as well as Slovenian.

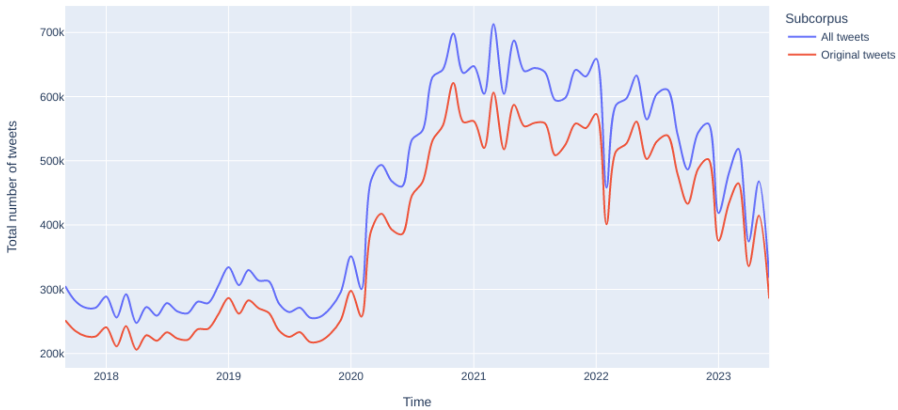


Figure 2: Number of monthly tweets in the HBS-ijekavian corpus through time

We observe a spike in traffic during the early summer of 2020 in the HBS-ekavian corpus, which roughly aligns with parliamentary elections in Serbia. Before that, we observe a drop in production throughout the year 2019 in the HBS-ekavian corpus. Further analysis is required to fully explain it, but it is important to note that 2019 was the year of relatively large interventions into the Serbian twitter user base by Twitter itself. Consult the chart of the ratio of tweets and retweets in the following section as well as the discussion accompanying it.

Ultimately, a much deeper look into the specific months beyond the scope of this presentation of the corpora would be required to conclusively explain the spike.

3.2 How much of Twitter are echoes

By looking at the ratio between original and retweeted content, calculated as number of original tweets divided by the number of all tweets (including retweets) we can observe an interesting difference between the Slovenian, HBS-ijekavian and HBS-ekavian corpora. While the relative amount of original content is consistently dropping in the Slovenian corpus, it is actually growing

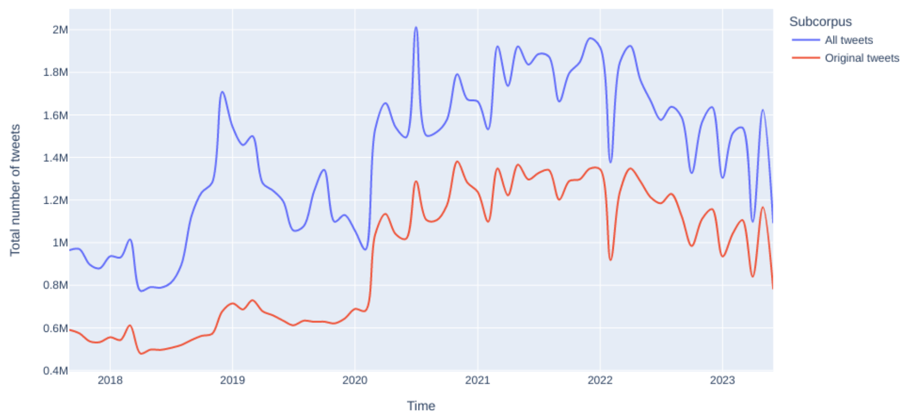


Figure 3: Number of monthly tweets in the HBS-ekavian corpus through time

(meaning relatively fewer retweets every month) in the HBS-ijekavian and HBS-ekavian corpora. This shows a larger relative number of original content was harvested in HBS languages vs. Slovenian.

A notable drop (indicating a larger amount of retweets) occurs in the Slovenian corpus during the lockdown months of 2020 coinciding with an increase in absolute number of posts and Slovenia's 2021 parliamentary election. The ratio never recovers, instead keeps dropping with a speed similar to that before the election.

Also of note is the time between the final months of 2019 in the HBS-ekavian corpus. While the drop is relatively small, the trend of the remaining accounts continues upward. While a rigorous analysis of this dip is beyond the scope of this introductory exploration, we hypothesise it is connected with Twitter's suspension of at least $8,558$ accounts during 2019 as reported by Twitter and analysed by the Stanford Internet Observatory Cyber Policy Center (Bush, 2020).

3.3 How Twitter feels

Since all the tokens in our corpora are linguistically annotated, we can use the lemmas to draw some broad conclusions about the emotional state of Twitter. Counting each of the tokens present in each month we can then compare them

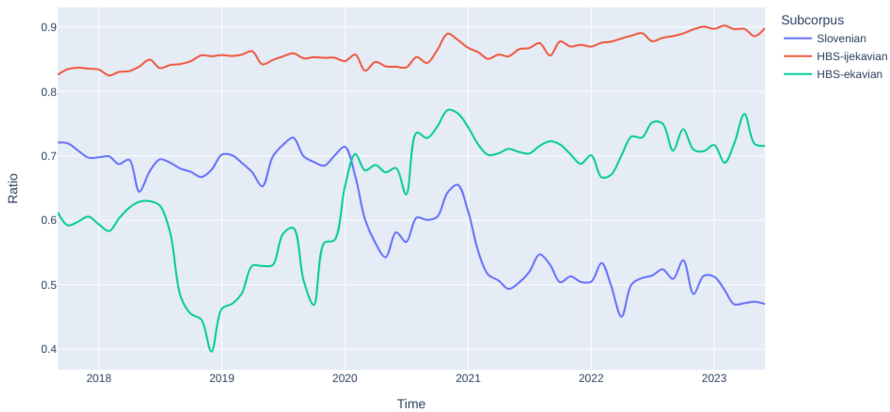


Figure 4: The ratio between tweets and original content in all three corpora.

with the LiLaH emotion lexicon (Daelemans et al., 2020) based on (Mohammad & Turney, 2010), to count the number of positively and negatively emotionally charged words. Based on those we calculate the ratio between positive and negative token counts by dividing the positive with the negative. We observe that the ratio of positive vs. negative is in favor of positively emotionally charged tokens in all three corpora. This holds true for both original tweets and those including retweets, which both show similar ratios between emotionally positive and negative tokens.

While the aforementioned LiLaH emotion lexicon only provides manual translations into Slovenian and Croatian languages, we could not find an equivalent manually translated lexicon for Serbian. The authors of the original emotion lexicon provide machine-translated lexica for other languages, among them Serbian, but using that on our HBS-ekavian corpus produced highly suspicious results. The emotional valence was effectively flatlining around 1 throughout the time we are investigating. Using the Croatian lexicon, albeit imperfect, produced much more sensible results comparatively. We are presenting the results based on the latter and strongly advise against using machine-translated lexica for similar tasks. In our particular case, the use of the machine-translated lexicon produced faulty and ultimately misleading results.

Table 2: Ratios of positive over negative token counts.

Corpus	*Original content*	*Original and retweets*
Slovenian	1.956	1.927
HBS-ijekavian	2.333	2.336
HBS-ekavian	2.404	2.406

We perform the same analysis for every month and plot it for original tweets and those with retweets included. Looking at figures 5 and 6 a notable shift towards more negative expressions can be seen in all corpora. This drop coincides with the introduction of lockdown measures from March 2020. All corpora show a negative trend in terms of tokens emotional valence signifying a slow but consistent shift in the kind of language used on Twitter.

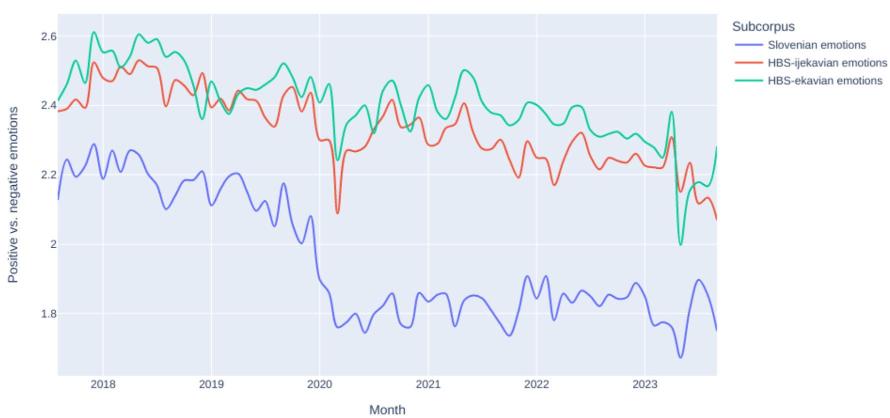


Figure 5: The ratio between positive and negative emotions in original tweets.

3.4 What Twitter talks about

There are of course multiple ways to inspect the content of tweets, but for this particular demonstration we will use the named entities recognised during our linguistic annotations. Similarly to above, we count the number of all named entities in the corpus. Listing the most common named entities in each of the person, organisation, and location categories, we can inspect the meatspace grounding of the corpora.

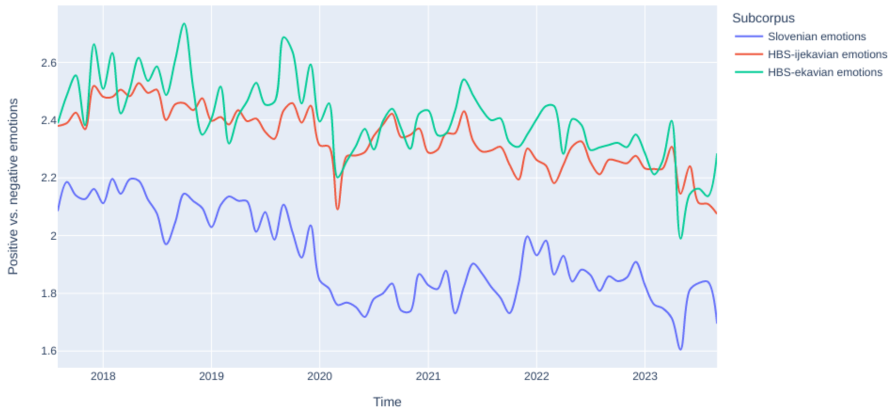


Figure 6: The ratio between positive and negative emotions in all tweets including retweets.

For the Slovenian corpus, the most common named people are all Slovenian politicians: Janez Janša (prime minister in during lockdown measures), Marjan Šarec (prime minister before lockdown measures), and Borut Pahor (then Slovenian president). Among organisations we find RTV Slovenija (national broadcaster), the constitutional court, and national assembly. The top three locations Novo mesto, Nova Gorica, and Murska Sobota are Slovenian cities, effectively local administrative and business hubs. We can see the corpus grounded in Slovenian national politics, media, and geography. The other two corpora do not demonstrate such national and geographic limits.

For the HBS-ijekavian corpus, we find Milorad Dodik (president of Republika Srpska, BiH), Željko Komšić (Croat member of the Presidency of BiH), and Milo Đukanović (twice president of Montenegro). Interestingly, Zoran Milanović (current president of Croatia) is the fifth most commonly named person after Novak Đoković. Looking at organisations we find the presidency of BiH, the European Union and the constitutional court. Most mentioned locations include Montenegro, BiH, and Republika Srpska. Compared to Slovenian, this corpus is much more geographically diverse, while still mostly political when it comes to recognised named entities. The results of this analysis point towards the conclusion that a significant part of the HBS-ijekavian corpus contains tweets from Bosnia and Herzegovina.

Finally, looking at the HBS-ekavian corpus among people we find Aleksandar Vučić (president of Serbia since 2017) , Vuk Jeremić (Serbian politician), and Dragan Đilas (Serbian politician, former mayor of Belgrade). Novi Sad (Serbian city), Montenegro and Republika Srpska are the most commonly mentioned locations with Crvena Zvezda (sports club) leading among organisations, followed by the Serbian government and the party SNS (Srpska Narodna Stranka). This analaysis shows the very much expected result that in the HBS-ekavian corpus to the most part Serbian content is represented.

While smaller in the HBS-ekavian corpus, the geographical diversity of named entities even in that corpus, which would be expected to contain purely Serbian content, further supports the claim that language identities in the Balkans do not conform to national borders. Any attempt to geographically or politically locate linguistic artefacts based on the language or linguistic features is destined for failure. This should also bring into question the design, structure, and presentation of language technologies and language resources, not least the labels we use for models and their results.

3.5 What is tagged on Twitter

Another way to peek into the content of Twitter conversations is to look at the use of hashtags. Comparing most common hashtags between retweets-included and retweets-excluded collections, we can see that there is a relatively large overlap between the two perspectives. Political and news hashtags are most present, with popular culture trailing the top 20. A notable exception to the overlap is the hashtag *#Požareport* in the Slovenian corpus, which is among the top three hashtags when we include retweets, but completely absent from the originals-only subset. The hashtag is used to promote a relatively right wing news/tabloid portal previously connected with prominent Slovenian political parties. Its presence among top hashtags in the retweets-included collection only is indicative of inorganic promotion, but a deeper analysis of users and messages reproducing the tweets would be required to confirm what is usually called ”inauthentic coordinated behavior” (Cinelli et al., 2022).

Figure 7: Most popular hashtags in all Slovenian content.

Figure 8: Most popular hashtags in original Slovenian content.

Figure 9: Most popular hashtags in all HBS-ijekavian content.

Figure 10: Most popular hashtags in original HBS-ijekavian content.

Figure 11: Most popular hashtags in all HBS-ekavian content.

Figure 12: Most popular hashtags in original HBS-ekavian content.

3.6 Who talks on Twitter

The most prolific accounts in the Slovenian corpus for authors of all content (original tweets and retweets) are visualised in Figure 13. All of the top retweeters are what we can consider ”regular users”, i.e. users without an explicitly attributed organisational affiliation or publicly recognised meatspace persona. In the original content wordcloud we see the majority of users still as regular users, but only barely. 9 out of the top 20 posters of original content are either journalists or news outlets. Most prolific posters of original content are visualised as a wordcloud in Figure 14.

Figure 13: Most prolific posters of any Slovenian content

Figure 14: Most prolific posters of original Slovenian content

The HBS-ijekavian corpus differs from the Slovenian in that both wordclouds (posters of original content, posters of any content) contain accounts connected with news outlets. Another notable thing is that the list includes accounts which have since been deleted or suspended.

Figure 15: Most prolific posters of any HBS-ijekavian content

Figure 16: Most prolific posters of original HBS-ijekavian content

The HBS-ekavian corpus contains a single news outlet and one institution in the original content column. The production of HBS-ekavian tweets seems much more dominated by regular users than the other two corpora. Again, similar to the HBS-ijekavian corpus we find deleted and suspended accounts in both collections. This property of the HBS corpora potentially is consistent with Twitters reports on mass account deletions.

Figure 17: Most prolific posters of any HBS-ekavian content

Figure 18: Most prolific posters of original HBS-ekavian content

3.7 Who is being talked at on Twitter

Figures 19, 20, 21, 22, 23, and 24 show the most commonly mentioned accounts including and excluding retweets. We can see political accounts having the strongest presence when it comes to mentions. In this sense “private” conversations represent a vanishingly small portion of Twitter mentions. We can understand the primary use of mentions being to call out or name (and often shame) a particular politician, party, or institution. The overlap between top mentions in original and all tweets is quite large, over half of the top 20 are the same across both.

In both the HBS-ijekavian and HBS-ekavian corpora, we can see *YouTube* stand out of the political and media crowd indicating many of the captured tweets are semi-automatically generated share tweets (such as users get when using in-app share buttons). This indicates a corpus of tweets contains information about users' digital habits beyond simple online conversations.

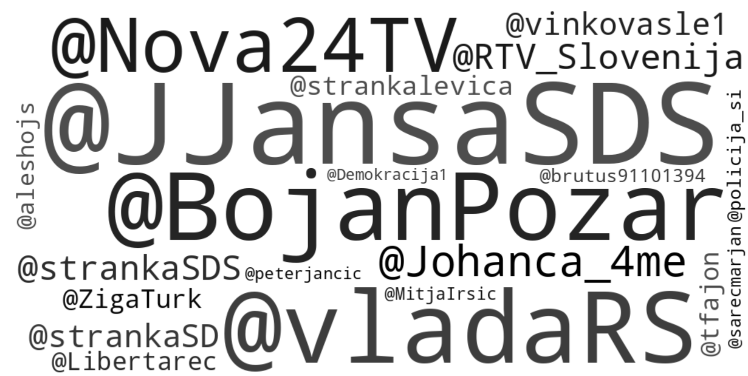


Figure 19: Most mentioned accounts in any Slovenian content.

Figure 20: Most mentioned accounts in original Slovenian content.

Figure 21: Most mentioned accounts in any HBS-ijekavian content.

Figure 22: Most mentioned accounts in original HBS-ijekavian content

Figure 23: Most mentioned accounts in any HBS-ekavian content.

Figure 24: Most mentioned accounts in original HBS-ekavian content.

4 CONCLUSION

In this paper we present a novel corpus of Twitter posts spanning over 6 years and 170 million tweets. We outline the data collection and processing required to obtain the corpus in its current form. We briefly discuss the split of the collection into three distinct corpora, followed by an exploratory pilot analysis of the tweets contained. We present and discuss the production of tweets through time, the relationship between the amount of original vs. retweeted content, and perform some cursory distant reading into the emotions, named entities, hashtags, authors, and mentions in the collection.

We show that the Twitter corpora created contain a wealth of information and show echoes of real life (albeit most commonly political) events. Among others, we observe the effects of COVID-19 preventative measures on the amount and quality of tweets. Furthermore, we show that while language categorisation might be useful in terms of the quality of specific linguistic annotations it would be ill-advised to use those categories to infer non-overlapping geographical or cultural categories, as shown on the HBS corpora which contain quite a lot of references to entities related to several countries, administrations, and cultural spaces.

The split of corpora presented in this paper is one of many possible, and future research would be well advised to generate subcorpora based on specific research questions it aims to address. Each of the possible splits brings with it its own set of biases which should be evaluated at the point of use. This is especially true when researching coordinated inauthentic behaviour, signs of which are present in all three corpora, though much more noticeable in the HBS corpora (with some of the most prolific accounts from the investigated period banned today).

The differences between the three corpora presented are both qualitative and quantitative, but they serve to highlight specifics of each other. In other words, should the corpora be merged together, we would be unable to observe the spikes in tweet production during Slovenia's election. Likewise, without the split, the observed change in communication after the introduction of COVID-19 preventative measures would be visible, but the argument for its relative universality much less convincing.

The size of the collection and the results of our initial inquiry call for further research into the discourse on Twitter in the past years. While we do not have the rights to make the corpus freely available, we are retaining an archive for future text and data mining endeavours based on the exceptions provided by the European Union. We invite research teams who would like to use this dataset for their own research to contact the authors for access to the collection.

REFERENCES

Altman, M. (2018). Tufekci, z.: Twitter and tear gas: The power and fragility of networked protest. , *29*(4), 884–885. Retrieved 2024-05-31, from https://doi.org/10.1007/s11266-017-9927-0 doi: 10.1007/s11266-017-9927-0

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135–146.

Bronstein, J., Aharony, N., & Bar-Ilan, J. (2018). Politicians' use of facebook during elections: Use of emotionally-based discourse, personalization, social media engagement and vividness. , *70*. doi: 10.1108/AJIM-03-2018-0067

Bush, D. (2020). "fighting like a lion for serbia": An analysis of government-linked influence operations in serbia. Retrieved 2024-05-29, from https://fsi.stanford.edu/publication/april-2020-serbia-takedown

Chan, M., & Yi, J. (2024). Social media use and political engagement in polarized times. examining the contextual roles of issue and affective polarization in developed democracies. Retrieved 2024-05-31, from https://www.tandfonline.com/doi/abs/10.1080/10584609.2024.2325423 (Publisher: Routledge)

Cinelli, M., Cresci, S., Quattrociocchi, W., Tesconi, M., & Zola, P. (2022). Coordinated inauthentic behavior and information spreading on twitter. , *160*, 113819. doi: 10.1016/j.dss.2022.113819

Daelemans, W., Fišer, D., Franza, J., Kranjčić, D., Lemmens, J., Ljubešić, N., … Popič, D. (2020). *The LiLaH emotion lexicon of croatian, dutch and slovene.* http://hdl.handle.net/11356/1318 (Slovenian language resource repository CLARIN.SI)

Effing, R., van Hillegersberg, J., & Huibers, T. (2011). Social media and political participation: Are facebook, twitter and YouTube democratizing our political systems? In E. Tambouris, A. Macintosh, & H. de Bruijn (Eds.), *Electronic participation* (pp. 25–35). Springer. doi: 10.1007/978-3-642-23333-3_3

Flew, T., & Iosifidis, P. (2020). Populism, globalisation and social media. , *82*(1), 7–25. Retrieved 2024-05-31, from https://doi.org/10.1177/1748048519880721 (Publisher: SAGE Publications Ltd) doi: 10.1177/1748048519880721

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and research: A systematic literature review through text mining. , *8*, 67698–67717. Retrieved 2024-05-31, from https://ieeexplore.ieee.org/document/9047963 (Conference Name: IEEE Access) doi: 10.1109/ACCESS.2020.2983656

Labrèche, G. (2023, March). *Cyrtranslit.* Zenodo. https://doi.org/10.5281/zenodo.7734906 (A Python package for bi-directional transliteration of Cyrillic script to Latin script and vice versa. Supports transliteration for Bulgarian, Montenegrin, Macedonian, Mongolian, Russian, Serbian, Tajik, and Ukrainian.) doi: 10.5281/zenodo.7734906

Ljubešić, N., Fišer, D., & Erjavec, T. (2014, May). TweetCaT: a tool for building Twitter corpora of smaller languages. In N. Calzolari et al. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 2279–2283). Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/834_Paper.pdf

Ljubešić, N., & Kranjčić, D. (2014). Discriminating between very similar languages among twitter users. In *Proceedings of the ninth language technologies conference* (pp. 90–94).

Ljubešić, N., Petrović, M. M., & Samardžić, T. (2018). Borders and boundaries in bosnian, croatian, montenegrin and serbian: Twitter data to the rescue. *Journal of Linguistic Geography*, *6*(2), 100–124.

Mohammad, S., & Turney, P. (2010, June). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and genera-*

tion of emotion in text (pp. 26–34). Los Angeles, CA: Association for Computational Linguistics. https://aclanthology.org/W10-0204

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages.

Qi, Y., & Shabrina, Z. (2023). Sentiment analysis using twitter data: a comparative application of lexicon- and machine-learning-based approach. , *13*(1), 31. Retrieved 2024-05-31, from https://doi.org/10.1007/s13278-023-01030-x doi: 10.1007/s13278-023-01030-x

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with twitter data. , *7*(1), 13006. Retrieved 2024-05-31, from https://www.nature.com/articles/s41598-017-12961-9 doi: 10.1038/s41598-017-12961-9

Rupnik, P., Kuzman, T., & Ljubešić, N. (2023, May). BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. In Y. Scherrer, T. Jauhiainen, N. Ljubešić, P. Nakov, J. Tiedemann, & M. Zampieri (Eds.), *Tenth workshop on nlp for similar languages, varieties and dialects (vardial 2023)* (pp. 113–120). Dubrovnik, Croatia: Association for Computational Linguistics. https://aclanthology.org/2023.vardial-1.11 doi: 10.18653/v1/2023.vardial-1.11

Terčon, L., & Ljubešić, N. (2023). CLASSLA-Stanza: The next step for linguistic processing of south slavic languages.

Wang, L., Niu, J., & Yu, S. (2020). SentiDiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis. , *32*(10), 2026–2039. Retrieved 2024-05-31, from https://ieeexplore.ieee.org/document/8700266 (Conference Name: IEEE Transactions on Knowledge and Data Engineering) doi: 10.1109/TKDE.2019.2913641

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In P. Clough et al. (Eds.), *Advances in information retrieval* (pp. 338–349). Springer. doi: 10.1007/978-3-642-20161-5_34

TWITTER PRED X: PODOBE Z BALKANA

Predstavljamo nov korpus preko 170 milijonov tvitov v Slovenskem in jezikih HBS, zbranih med letoma 2017 in 2023. Po opisu procesa zbiranja in (pred)procesiranja podatkov predstavimo odločitev o razdelitvi korpusa v tri podkorpuse, vključno z ovirami na poti jezikovnega označevanja zbirke iz tako narodnostno in jezikovno mešanega področja kot je Balkan. Nadaljujemo s pilotno analizo kvalitete, količine in razlik med zbirkami. Predstavimo indice o vplivu dogodkov v resničnem življenju na uporabniško generirane vsebine na spletu. Po predstavitvi nekaterih najbolj vidnih vsebinskih lastnosti zbirke zaključimo z vabilom k dodatnemu raziskovanju tega obsežnega in raznolikega korpusa sodobnega spletnega govora.

Keywords: Twitter korpus, slovenščina, makro jezik HBS, oddaljeno branje.

EXTENDING THE SPOKEN SLOVENIAN TREEBANK

Kaja DOBROVOLJC
Faculty of Arts, University of Ljubljana
Department for Artificial Intelligence, Jožef Stefan Institute

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. The original version of the SST treebank was expanded with over 3,000 newly annotated utterances, and enhanced in terms of the consistency of transcriptions and the quality of the annotations. After a brief overview of the data sampling procedure and the semi-automatic morphological annotation, the core of the paper focuses on the the dependency annotation campaign, and the resolution of the discrepancies in sentence segmentation, capitalization and punctuation between the original and the newly added transcriptions. Finally, we summarize the contents of the new treebank with respect to its size and diversity, and evaluate it against the reference SSJ treebank of written Slovenian, highlighting the unique lexical and morphosyntactic characteristics of spoken communication.

Keywords: corpus annotation, dependency treebank, spontaneous speech, Slovenian language, Universal Dependencies

1 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 (e.g. Hinrichs and Kübler (2005); Pietrandrea and Delsart (2019); van der Wouden et al. (2003)) and natural language processing (e.g. Liu and Prud'hommeaux (2021); Braggaar and van der Goot (2021); Caines et al. (2017)). Consequently, many spoken language treebanks have been developed over the recent decades, such as the Switchboard corpus for English (Godfrey et al., 1992), CGN for Dutch (van der Wouden et al., 2002), PDTSL for Czech (Hajič et al., 2008), NDC and LIA for Norweigian (Øvrelid et al., 2018; Kåsen et al., 2022), Rhapsodie for French

(Lacheret-Dujour et al., 2019), as well as the multilingual Verbmobil (Hinrichs et al., 2000) and CHILDES (MacWhinney, 2014) collections. Recently, many such treebanks have emerged as part of the expanding multilingual Universal Dependencies (UD) dataset (de Marneffe et al., 2021; Dobrovoljc, 2022).

For Slovenian, the Spoken Slovenian Treebank (SST; Dobrovoljc and Nivre (2016)) 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 corpus of spoken Slovenian (Zwitter Vitez et al., 2021; Verdonik et al., 2013) and features manually annotated transcriptions on the levels of lemmatization, MULTEXT-East 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 (named after the *Sporazumevanje v slovenskem jeziku* project), which features identical annotations (Dobrovoljc et al., 2017; Dobrovoljc & Ljubešić, 2022; Arhar Holdt et al., 2024), and has already been used as the main data source for the development of specialized computational models for grammatical annotation of spoken Slovenian (Dobrovoljc & Martinc, 2018; Verdonik, Dobrovoljc, Erjavec, & Ljubešić, 2024; Krsnik & Dobrovoljc, 2024).

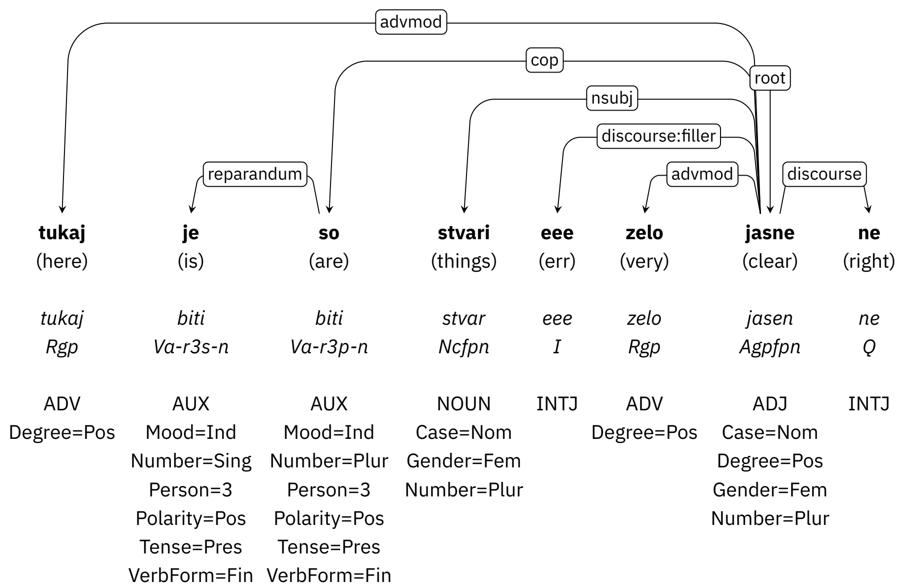
To alleviate the shortcomings of the original version of the SST treebank, such as its relatively small size (approximately 3,100 parsed utterances amounting to 30,000 annotated tokens), and diverse, but fragmented data (short samples of many speech events), the ongoing project SPOT (*Treebank-driven approach to the study of Spoken Slovenian*, ARIS grant no. Z6-4617),[1] aims at extending the treebank with a minimum of 50,000 new tokens. Consequently, the treebank was recently extended to more than triple its original size, by expanding some of the original data samples and adding completely new data from the recently expanded version of the reference corpus – GOS 2 (Verdonik, Dobrovoljc, Erjavec, & Ljubešić, 2024).

We describe this major improvement of the SST treebank in the continuation of this paper by giving a brief summary of the data sampling procedure in Section 2 and describing the new data annotation and final dataset consolidation in

[1] https://spot.ff.uni-lj.si/

Section 3. We provide an overview of the resulting language resource in Section 4 and give details on its format and availability in Section 5. Finally, we present a comparison of the new SST treebank to the SSJ treebank of written Slovenian in Section 6 to exemplify its value for further empirical investigations of lexical and grammatical characteristics of Slovenian speech.

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



2 CORPUS EXTENSION

To address the aforementioned disadvantages of the original SST corpus, our aim was to extend the original SST treebank by a minimum of 50,000 new tokens while maintaining its representativeness with respect to the (updated) GOS 2 reference corpus of spoken Slovenian (Verdonik, Dobrovoljc, Erjavec, & Ljubešić, 2024).

The data sampling procedure was designed in collaboration with the Mezzanine[2] project and is described in more detail by Verdonik, Dobrovoljc, Čibej, et al. (2024). In summary, the sampling was conducted through a manual selection of specific speech events from the GOS 2 corpus (Verdonik, Dobrovoljc, Erjavec, & Ljubešić, 2024)[3] and was performed in two steps. First, 22 samples from GOS 1 events in the original SST corpus were expanded with approximately 450 additional words per event, resulting in about 10,000 new words in total from the GOS 1 subset. Second, 57 entirely new speech events from the ARTUR subset were added, each contributing approximately 800 new words, totalling to around 40,000 new words from the ARTUR subset. The exact counts, which also account for the post-festum modifications of the data described in the following sections, are reported in Section 4 (Table 2).

From the perspective of subsequent syntactic annotation of the data, an important drawback of the sampled ARTUR subset was the presence of very short segments, with segment breaks introduced after each pause rather than after the completion of a semantically and syntactically complete unit of speech, as was the case for the utterance segmentation in GOS 1. To resolve this, the ARTUR subset was automatically resegmented based on the sentence-final punctuation markers available (for details and example see Verdonik, Dobrovoljc, Čibej, et al. (2024)), which resulted in more coherent structures for subsequent syntactic analysis. The resegmentation was performed as part of the conversion of the newly sampled data (originally in XML TEI) to CONLL-U, which was also the file format we used in the continuation of our work presented below.

3 TREEBANK ANNOTATION

Following the data sampling and pre-processing steps presented above, the resulting new dataset was manually annotated for lemmas, morphological features and syntactic dependencies.

[2] https://mezzanine.um.si/

[3] The GOS 2 corpus consists of the original GOS 1 corpus (Zwitter Vitez et al., 2021), GOS VideoLectures corpus (Verdonik et al., 2021) and selected events from the ARTUR ASR database (Verdonik, Bizjak, Sepesy Maučec, et al., 2023)

3.1 Lemmatization and Morphology Annotation

In the first stage, the two new subsets presented in Section 2 have been semi-automatically annotated for lemmas and morphosyntactic tags in accordance with the MULTEXT-East annotation scheme (Erjavec, 2010; Holozan et al., 2023), which is the most widely used annotation scheme for Slovenian corpora. The process is described in detail by Čibej and Munda (2024), who also discuss the annotation issues related to the newly emerged speech-specific lexical and morphological phenomena.

The resulting morphologically annotated dataset was then converted to UD part-of-speech categories and morphological features using the jos2ud conversion pipeline (Dobrovoljc et al., 2017),[4]. The conversion features a large number of high-accuracy mapping rules and has previously been used for mapping MULTEXT-East tags to UD morphology in other reference resources for Slovenian, such as the ssj500k (Krek et al., 2021) and SUK (Arhar Holdt et al., 2022) training corpora of standard written Slovenian, the Janes-Tag corpus of non-standard written Slovenian (Lenardič et al., 2022), and the Sloleks lexicon of inflected forms (Čibej et al., 2022).

In the second stage, the transcriptions have been syntactically parsed according to the UD annotation scheme through a semi-automatic procedure described below.

3.2 Automatic Dependency Parsing

Following the nowadays prevailing approach to manual data annotation, the transcriptions have first been pre-annotated using an automatic parser. To select the optimal tool for the task, several models have been developed and evaluated. For parsing spoken Slovenian in particular, the SLOKIT[5] project has recently produced a specialized model of the CLASSLA-Stanza tool (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023). Following the findings by Dobrovoljc and Martinc (2018), the model was trained on a concatenation of spoken (SST) and written (SSJ) data and produced better results than the CLASSLA-Stanza parsing models trained on either written or spoken data alone

[4] https://github.com/clarinsi/jos2ud
[5] https://slokit.ijs.si/

(Verdonik, Dobrovoljc, Erjavec, & Ljubešić, 2024), confirming the positive effect of the larger training set.

Given the recent technological advancements, we extended this work by producing three additional models using the Trankit transformer-based tool (Nguyen et al., 2021), trained on the written SSJ and spoken SST treebanks, as released in UD v2.12 (Zeman et al., 2023), and the combination of the two. Thus, five parsing models have been evaluated with respect to the standard evaluation metric of labelled-attachment score (LAS), which gives the percentage of tokens with correctly predicted parent node and the type of their relation:[6]

- CLASSLA-Stanza default model for written Slovenian (Terčon & Ljubešić, 2023), trained on SSJ
- CLASSLA-Stanza SLOKIT model for spoken Slovenian, trained on SSJ and SST
- Trankit model for written Slovenian (Krsnik & Dobrovoljc, 2023), trained on SSJ
- Trankit model for spoken Slovenian, trained on SST
- Trankit model for spoken Slovenian (Krsnik & Dobrovoljc, 2024), trained on SSJ and SST

Table 1 shows the models' performance on both written (SSJ) and spoken (SST) test set, featured in the same dataset release.[7] Our results confirm previous findings that, regardless of the tool, the performance of the standard models trained on written data drops significantly when confronted with transcribed speech, and increases significantly when spoken data is featured in the training (approx. +15pp LAS for both joint SSJ+SST models). However, the transformer-based Trankit models display a much higher performance overall (both in written and spoken testing scenarios). Therefore, for the use case at hand, the best-performing Trankit SSJ+SST model (81.26 LAS F1) was chosen for the automatic pre-annotation of the newly added SST data (Section 2).

[6] The SLOKIT and SST-only Trankit model have not been officially released, but are available directly from the authors.

[7] The evaluation is performed on pre-tokenized (gold) test sests to neutralize the impact of speech segmentation–a notoriously difficult task if no sentence-final punctuation is available in the transcripts (see Dobrovoljc and Martinc (2018)).

Table 1: LAS F1 performance of selected parsing models on the SSJ and SST test sets.

Model	SSJ-test (written)	SST-test (spoken)
CLASSLA-Stanza (written)	90.64	55.43
CLASSLA-Stanza Slokit (written+spoken)	88.64	70.58
Trankit SSJ (written)	95.39	66.36
Trankit SST (spoken)	74.83	79.84
Trankit SSJ+SST (written+spoken)	95.47	**81.26**

3.3 Manual Dependency Annotation

The automatically parsed dataset with manually revised lemmatization and morphology was then split into document-level files (79 in total), with 2–3 independent annotators assigned to each file. The annotation was performed in the Q-CAT annotation tool (Brank, 2023), which was upgraded for this particular campaign to also enable listening of audio files, provided the URLs to the audio files are given as part of the `# sound_url` comment line in the input CONLL-U file. Given that Q-CAT does not support comparison of annotations produced by different annotators, the curation process was carried out through the WebAnno annotation service maintained by CLARIN.SI (Yimam et al., 2013; Erjavec et al., 2016). Given the fact that the original SST was annotated by a single annotator and some annotation guidelines have been changed, the original SST was also manually revised.

3.4 Annotation Guidelines

In addition to the UD guidelines available online,[8] which mostly include robust language-independent definitions and a limited set of illustrative examples, especially for speech-specific phenomena, the annotators were instructed to use the stand-alone manual for UD annotation of Slovenian texts (Dobrovoljc & Terčon, 2023). This document was originally published within the DSDE project to document the annotation of the written SSJ UD dataset (Dobrovoljc et al., 2023; Dobrovoljc & Ljubešić, 2022) and was now upgraded to also document the guidelines for spoken data annotation. The latter are based on the (sparsely documented) annotation of the original SST treebank (Dobrovoljc &

[8] https://universaldependencies.org/guidelines.html

Nivre, 2016), as well as the more recent practices and discussions within the community (Kahane et al., 2021; Dobrovoljc, 2022).

Due to space limitations, we only describe here how the two most typical speech-specific phenomena are annotated: discourse markers (Section 3.4.1) and speech repairs (Section 3.4.2). For discussions of other speech-specific morphosyntactic phenomena, the readers are advised to refer to the full documentation in the aforementioned guidelines (Dobrovoljc & Terčon, 2023)[9] or the discussions in papers by Dobrovoljc and Nivre (2016) and Dobrovoljc (2022, 2024).

3.4.1 DISCOURSE MARKERS

According to the general UD guidelines, the *discourse* relation is used for interjections and other discourse particles and elements which are not clearly linked to the structure of the sentence, except in an expressive way. These include interjections (e.g. *oh*), fillers (e.g. *eee* 'err'),[10] and discourse markers in the narrow sense (*no* 'well', *a ne* 'right'). Figure 2 illustrates a tree involving two such typical expressions and shows that they attach to the head of the most relevant clause (usually the root predicate), even though they are not dependent of the predicates as such.

If an utterance consists of discourse elements only, the most prepositionally loaded marker (i.e. informative, content-rich) is chosen as the head node, as is the case with the feedback response *dobro* in Figure 3. If it is not possible to determine the most semantically salient expression, the first element in the sequence is treated as the head.

[9]The final version of the Slovenian UD guidelines for both written and spoken language annotation is planned to be published in September 2024 at https://wiki.cjvt.si/books/07-universal-dependencies-FPQ/page/annotation-guidelines.

[10]For filled pauses, we introduce a special *discourse:filler* label extension (relation sub-type), as illustrated by *eee* in Figure 2.

Figure 2: Annotation of discourse markers.

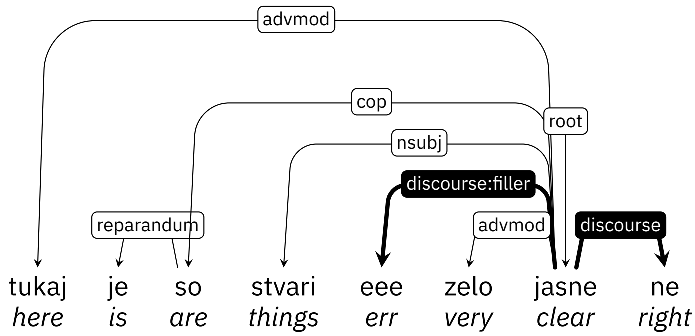
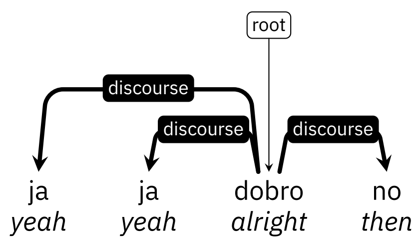


Figure 3: Annotation of a sequence of discourse markers.



3.4.2 SELF-REPAIRS

The *reparandum* relation is used to annotate self-repairs in speech, i.e. instances where a speaker replaces previously uttered content with a new one, as illustrated in Figure 1, where the singular form of the copula verb *je* is replaced by the correct plural form *so*.

The repaired unit can be syntactically complete or incomplete, such as unfinished words, phrases or clauses. In case of shared dependants between the reparandum and its repair, such as modifiers applicable to both, the dependent is attached to the repair rather than the reparandum. This is illustrated in Figure 4 below, which shows the premodifier *kako* being attached to the noun *orožje* rather than the first, unfinished attempt of pronouncing the word (*orož-*).[11] In case of a sequence of self-repairs–for example, when a speaker

[11]This design principle enables the sub-trees spanning from *reparandum*-marked tokens (i.e. disfluencies) to be easily removed without causing the remaining tree to become ungrammatical or

keeps restarting their intended verbalization–all reparandums attach to the same head, i.e. the head of the final repair (Figure 5).

Figure 4: Annotation of self-repairs with shared dependants in SST.

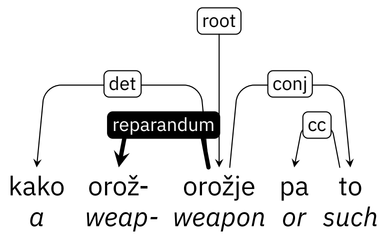
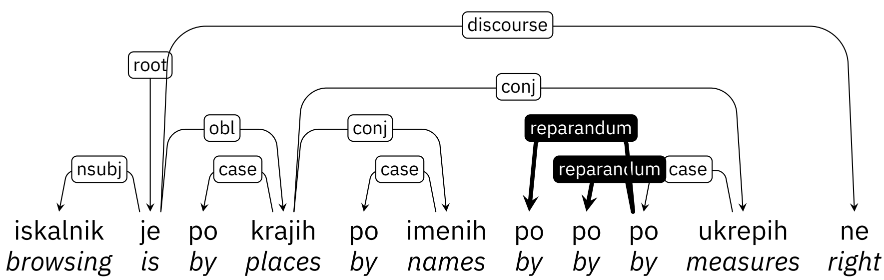


Figure 5: Annotation of a sequence of self-repairs in SST.



3.5 Final Subset Consolidation

Finally, both manually revised datasets (original SST and new data from GOS 2) have been merged and consolidated with respect to metadata formatting (see Section 5.2), but also transcription consistency, as different punctuation and casing principles have originally been adopted in GOS 1 and ARTUR corpora. Some corrections to the lemmatization and morphological layers have also been introduced.

semantically underspecified (e.g. *kako orožje pa to* in Figure 4 and *iskalnik je po krajih po imenih po ukrepih ne* in Figure 5.)

3.5.1 ADDITION OF PUNCTUATION

While GOS 1 transcriptions include only sentence-final markers of question (?) and exclamation (!) intonation, ARTUR features written-like punctuation in both sentence-medial (e.g. commas) and sentence-final positions (e.g. full stops). To ensure dataset consistency across both subsets and comply with the general tendency to include punctuation in similar spoken language treebanks (Dobrovoljc, 2022), sentence-medial and sentence-final punctuation has been added to the GOS 1 subset. This was performed through a semi-automatic approach, in which the GOS 1 transcriptions were first automatically punctuated using the Slovene Punctuator[12] tool and then manually checked so as to conform to the punctuation principles of the ARTUR database (Verdonik & Bizjak, 2023). In total, 12,732 punctuation symbols have been added.

In parallel, GOS 1 transcriptions have also been stripped of non-lexical tokens (annotated as punctuation in the original SST treebank), such as *[audience:laughter]* and *[pause]*, which–with the exception of the latter–have not been transcribed in ARTUR. The new consolidated SST treebank contains transcriptions that are more similar to written text than those in the original treebank, as they include punctuation and exclude other markers of prosody. However, this change in the underlying data does not hinder the array of research applications, since non-lexical phenomena can still be accessed from the transcriptions of the reference GOS corpus if necessary.

3.5.2 CORRECTION OF TRANSCRIPTIONS

The process of final data consolidation also included the correction of the erroneously transcribed (standardized) tokens that were identified by the annotators or signalled as a mismatch in the data validation phase using the official UD validator.[13] This includes corrections of erroneous capitalisation at the beginning of the sentences, resulting from the automatic casing unification applied to the original GOS 2.1 (Verdonik, Zwitter Vitez, et al., 2023), which aimed at lowercasing all words except for named entities. Transcription mistakes pertaining to tokenization, such as words that should either be split or merged,

[12] https://github.com/clarinsi/Slovene_punctuator
[13] https://github.com/UniversalDependencies/tools/

were not tackled in this iteration, as changes in the tokenization would impede the automatic mapping to the reference corpus and its derivatives.

3.5.3 CORRECTION OF MORPHOLOGY

The aforementioned data validation errors also highlighted some mistakes and inconsistencies in lemmatization and morphological annotation within both schemes, which were also resolved. In addition, some UD morphological annotations have been consolidated based on the final annotation guidelines, such as the categorization of colloquial expressions *kao* 'like' (PART), and *ene* 'about' (ADV), definite article *ta* 'the' (DET with no inflectional features), indefinite article *en* 'a' (DET) and anonymized names (PROPN with no inflectional features).

4 NEW SST TREEBANK OVERVIEW

This section presents the contents of the new SST treebank with respect to size (Section 4.1) and the diversity of the spoken data included (Section 4.2).

4.1 Treebank Size

As shown in Table 2, 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),[14] 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.

4.2 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

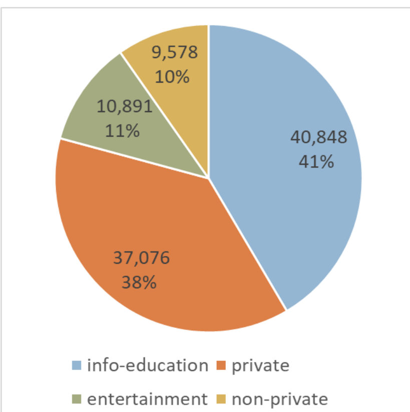
[14] The original version of the SST treebank (Dobrovoljc & Nivre, 2016) featured 287 events, 594 speakers, 3,188 utterances and 29,488 tokens.

Table 2: Overview of the new SST treebank and its subsets.

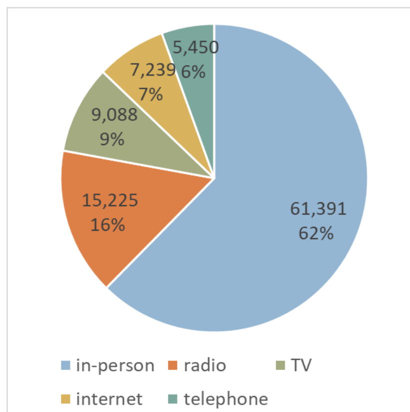
Subset	*Events*	*Speakers*	*Utterances*	*Tokens*
SST-2016-revised	287	594	2,903	36,960
New from GOS 1	22	61	1,236	13,112
New from ARTUR	57	72	1,969	48,321
SST-2024 (UD 2.15)	344	676	6,108	98,393

shown in Figures 6 to 9, which report the number of tokens per different types of speech events,[15] communication channels and speaker demographics.

Figure 6: Number of tokens in SST with respect to the nature of speech event.



(a) Event type



(b) Communication channel

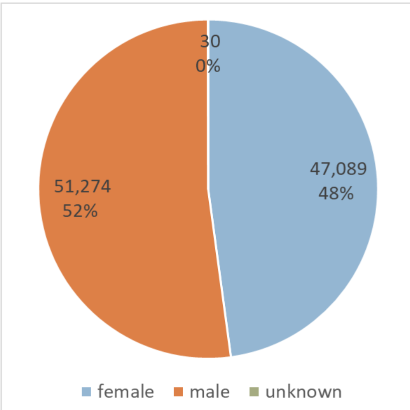
5 TREEBANK RELEASE

The new SST treebank is planned to be released as part of the official UD release v2.15 in November 2024.[16] It is freely available under the CC-BY license,

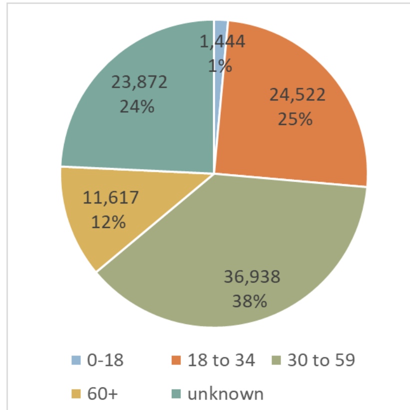
[15] Generally, all events feature spontaneous speech, i.e. unscripted verbal communication that occurs naturally in real-time, albeit with varying amounts of planning in public and non-public situations. A more detailed characterisation of speech events can be retrieved from the meta-data available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

[16] An interim version with extensions but no punctuation has already been published as part of UD release v2.14 in May 2024 (Zeman et al., 2024).

Figure 7: Number of tokens in SST with respect to speaker gender and age.

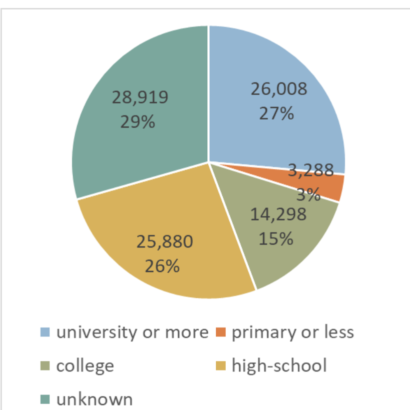


(a) Gender

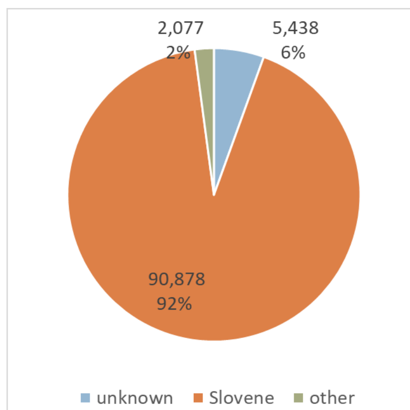


(b) Age

Figure 8: Number of tokens in SST with respect to speaker education and first language.



(a) Education



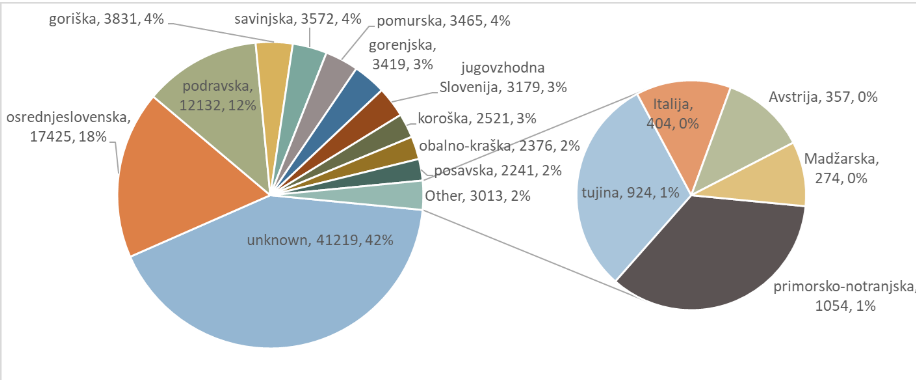
(b) First language

which is a less restrictive license in comparison to the CC-BY-NC license of the original version of the dataset, which prohibited commercial use.

5.1 Data Split

As required by the UD dataset release protocol, the treebank was split into training, development and test set with approximately 80%, 10% and 10% to-

Figure 9: Number of tokens in SST with respect to the region of speaker residence.



ken distribution in each. As with the original SST treebank, the split has been randomised on document-level, which ensures an equal distribution of the different event and speaker types (Section 4) across all three datasets. It was also ensured that the train, test and dev data from the original SST version was preserved in the same subset, so to enable fair model comparisons across different versions of the SST dataset.

5.2 Format

The treebank is encoded in the standard CONLL-U format,[17] illustrated in Figure 10, where each token in a sentence is represented on a single line with 10 fields: ID (token index), FORM (word form), LEMMA, UPOS (universal part-of-speech), XPOS (language-specific tag, i.e. MULTEXT-East), FEATS (morphological features), HEAD (index of the head token), DEPREL (dependency relation to the head), DEPS (enhanced dependency graph, not used in SST), and MISC (miscellaneous information).[18]

Speech-specific extensions of the format pertain to the comment lines, which include information on the document, sentence ID, speaker ID, and the audio

[17] https://universaldependencies.org/format.html

[18] Due to space limitations, the CONLL-U example in Figure 10 only shows the first feature in the FEATS column (but see the example in Figure 1) and omits the MISC column (e.g.`pronunciation=tuki|GOS2.1_token_id=GOS119.tok1104`).

URL,[19] as well as to the last (miscellaneous) column, which includes information on the pronunciation-based spelling of the word form (e.g. *tko* for the standardized word form *tako* 'such') and the token/segment IDs pertaining to the original GOS 2.1 corpus.

This ensures that all other types of metadata pertaining to the recorded event and speakers involved can easily be retrieved from the reference GOS 2.1 corpus via the persistent and traceable IDs. This includes retrieving all other relevant information omitted from the final SST treebank, such as the placement of pauses, non-vocal sounds or other types of transcribed but syntactically less relevant non-lexical phenomena.[20]

Figure 10: Example of an annotated utterance 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     7   advmod            _   _
2   je      biti    VERB    Va-r3s-n Mood=Ind...   3   reparandum        _   _
3   so      biti    AUX     Va-r3p-n Mood=Ind...   7   cop               _   _
4   stvari  stvar   NOUN    Ncfpn   Case=Nom...    7   nsubj             _   _
5   eee     eee     INTJ    I       _              7   discourse:filler  _   _
6   zelo    zelo    ADV     Rgp     Degree=Pos     7   advmod            _   _
7   jasne   jasen   ADJ     Agpfpn  Case=Nom...    0   root              _   _
8   ne      ne      PART    Q       Polarity=Neg   7   discourse         _   _
```

5.3 Online Access

In addition to the official SST dataset release in CONLL-U, which is also available on GitHub,[21] the SST treebank can also be accessed for browsing and

[19] For resegmented ARTUR-based data (see Section 2), the links in # sound_url point to a concatenation of the audio files available for the original GOS 2.1 segments. In the rare instance where an original ARTUR segment was split two SST segments, the original audio file appears in both concatenations. As a result, some linked audio files might include longer spans of speech than what is actually featured in the transcribed utterance.

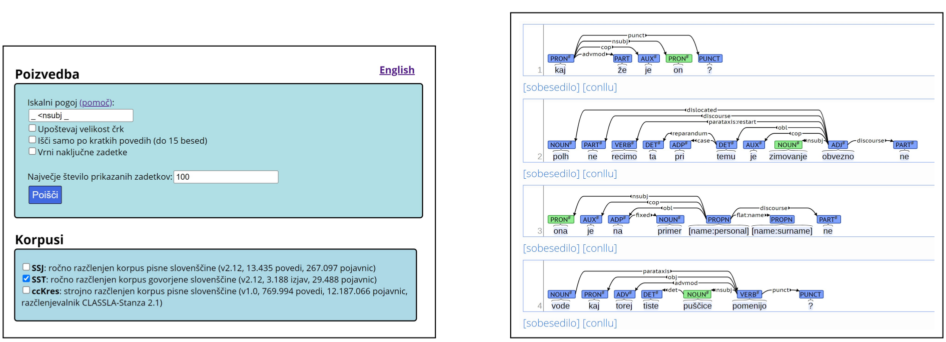
[20] This includes the audio recordings of the events, which are freely available under CC-BY for the ARTUR subset (Verdonik, Bizjak, Žgank, et al., 2023), and for research purposes for the GOS 1 subset.

[21] https://github.com/UniversalDependencies/UD_Slovenian-SST

analysis through the numerous tools that support querying and visualising UD treebanks worldwide. Online services with regular data updates include Grew-match,[22] maintained by INRIA Nancy, and INESS,[23] maintained by CLARINO. An important advantage of the former is the fact that it also supports listening to audio recordings in treebanks featuring spoken data.

The latest version of the SST treebank has also been uploaded to the locally developed Drevesnik treebank-querying service (Štravs & Dobrovoljc, 2024),[24] which is based on the open-source dep_search tool (Luotolahti et al., 2017). In addition to featuring other manually and automatically parsed UD corpora for Slovenian, the main advantage of the service (illustrated in Figure 11) from the perspective of Slovenian users is that it features a powerful and easy-to-use query language (documented in both English and Slovenian), enables regex-supported querying of the popular MULTEXT-East tags (XPOS column), randomisation of the results and their limitation to short sentences only (useful for illustrative or didactic purposes).

Figure 11: Drevesnik online service for querying Slovenian dependency treebanks (left: query interface, right: results interface).



The SST treebank also represents the backbone of the emerging ROG training corpus of spoken Slovenian (Verdonik, Dobrovoljc, Čibej, et al., 2024), which will feature additional annotation layers for disfluencies, dialogue acts, and prosody boundaries for some of the transcribed events, and will be encoded in other formats as well.

[22] https://universal.grew.fr/
[23] https://clarino.uib.no/iness
[24] https://orodja.cjvt.si/drevesnik/

6 COMPARISON WITH THE SSJ TREEBANK OF WRITTEN SLOVENIAN

Finally, we compare the new SST treebank with its written counterpart, the SSJ UD treebank of written Slovenian (Dobrovoljc et al., 2017), 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 tokens, adopting different functions in the representation of both modalities, the comparison is based on treebanks excluding punctuation. The results thus reflect the analysis of all uttered phenomena rather than all transcribed phenomena.

6.1 Vocabulary

The comparison of the vocabulary in Table 3 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 featured in the SST treebank that do not occur in the written SSJ treebank, confirming previous findings on the unique lexical characteristics of spoken Slovenian (Verdonik & Maučec, 2016; Dobrovoljc, 2018).[25]

Table 3: Comparison of vocabulary diversity in spoken and written 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

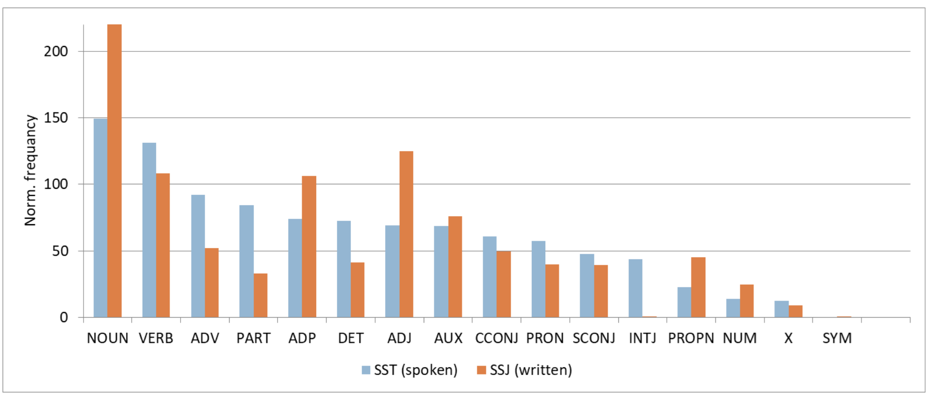
6.2 Part-of-Speech Categories

The comparison of part-of-speech tag frequencies per thousand words shown in Figure 12 reveals that the two modalities also differ with respect to the type of vocabulary used. For instance, spoken language exhibits a much higher fre-

[25] 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*).

quency 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 (Biber, 1988; Biber et al., 2010).

Figure 12: Comparison of the distribution of POS categories in spoken (SST) and written (SSJ) treebank.



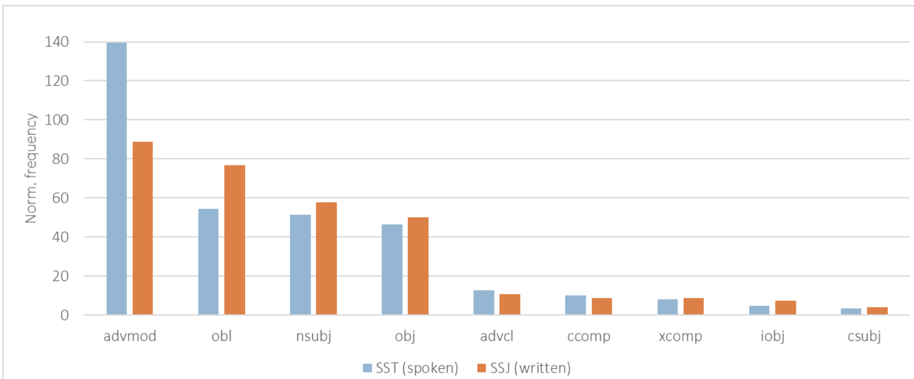
6.3 Dependency Relations

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

6.3.1 CORE DEPENDANTS OF PREDICATES

Figure 13 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*),[26] which could be explained by the abundance of modal adverbials, which speakers use to express stance, convey attitude, and balance the interaction.

Figure 13: Comparison of core predicate arguments in the spoken (SST) and written (SSJ) treebank.



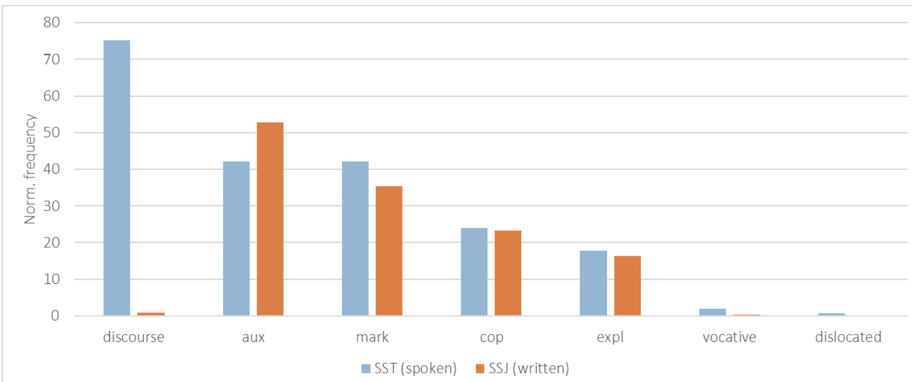
6.3.2 OTHER DEPENDANTS OF PREDICATES

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 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 13 and *acl* in Figure 15), spoken data exhibits a higher num-

[26]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.

ber of subordinate conjunctions (*mark*). This might be explained by the frequency of insubordinate clauses used as independent utterance to respond or to build upon a previous utterance on context (e.g. replying *Ker dežuje.* 'Because it is raining.' to a question on why an event was cancelled).

Figure 14: Comparison of the non-core predicate arguments in the spoken (SST) and written (SSJ) treebank.



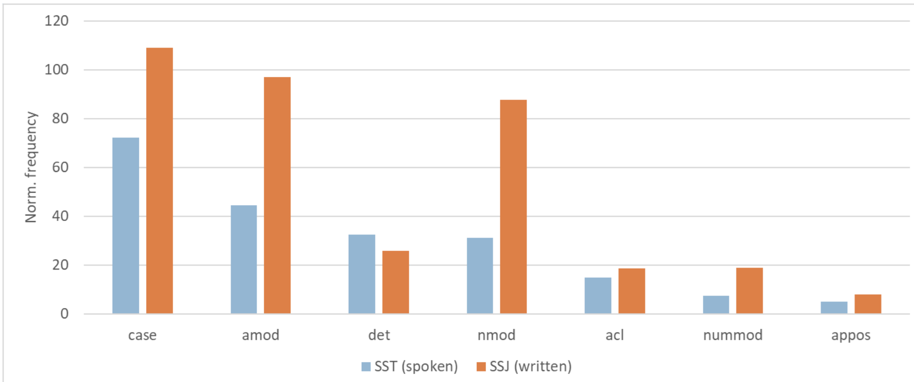
6.3.3 DEPENDANTS OF NOMINALS

The comparison of the distrubtion of the relations pertaining to the dependents of nominals (e.g. noun phrase constituents) in Figure 15 shows 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 12), 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.

6.3.4 OTHER RELATIONS

Last, Figure 16 shows the comparison of the distribution for all other types of dependency relations that do not fall into any of the main syntactic categories

Figure 15: Comparison of the dependents of nominals in the spoken (SST) and written (SSJ) treebank.



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.

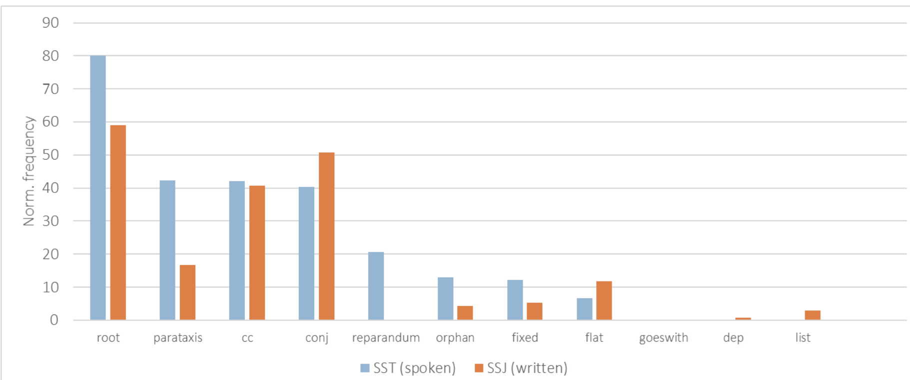
The second important observation is that sentences in speech are generally much shorter than in writing. This is not only reflected by the difference in the average number of words per utterance/sentence (i.e. the frequency of *root* elements in a treebank),[27] but also by the higher frequency of *parataxis* 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 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 14) or longer coordination phrases in writing (i.e. multiple conjuncts).

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

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 (Dobrovoljc, 2018). On the other hand, flat multi-word expressions (mainly encompassing personal names and foreign named entities) occur less often in speech.

Figure 16: Comparison of all other relations in the spoken (SST) and written (SSJ) treebank.



7 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 communication in Slovenian. We made a first step in this direction by comparing it to the SSJ treebank of written Slovenian, which revealed the unique lexical and morphosyntactic characteristics of spoken communication in comparison to writing. These findings relate to the interactive and situation-related nature of this type of language modality and further highlight the importance of integrating spoken language data into the Slovenian language resource landscape.

Short-term goals for future work include the integration of the treebank into the emerging multi-layer ROG corpus of spoken Slovenian, as well as the retraining and evaluation of state-of-the-art parsing models trained on the new dataset. Most importantly, the new SST treebank is planned to be used as the main data-source for a corpus-driven analysis of speech-specific syntactic patterns within the SPOT project, which will complement the robust SSJ-SST comparison presented in this paper with a more sophisticated analysis of syntactic (sub-)trees encountered in both treebanks, by using the STARK tool (Krsnik et al., 2024). Finally, our long-term goal is also to ensure a continuous incremental improvement of the quality of this richly annotated corpus, as well as to promote and facilitate its usage in Slovenian corpus linguistics.

ACKNOWLEDGMENTS

This work was financially supported by the Slovenian Research and Innovation Agency through the research project *Treebank-Driven Approach to the Study of Spoken Slovenian* (Z6-4617) and the research program *Language Resources and Technologies for Slovene* (P6-0411). In addition to the collaborators from the Mezzanine project (J7-4642) 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), punctuation insertion (Iztok Lebar Bajec) and audio resegmentation (Janez Križaj, Simon Dobrišek, Tomaž Erjavec).

REFERENCES

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Erjavec, T., Gantar, P., Krek, S., ... Žitnik, S. (2024, May). SUK 1.0: A New Training Corpus for Linguistic Annotation of Modern Standard Slovene. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 15428–15435). Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1340

Arhar Holdt, Š., Krek, S., Dobrovoljc, K., Erjavec, T., Gantar, P., Čibej, J., ... Zajc, A. (2022). *Training corpus SUK 1.0.* http://hdl.handle.net/11356/1747 (Slovenian language resource repository CLARIN.SI)

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511621024 doi: 10.1017/cbo9780511621024

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2010). The Longman Grammar of Spoken and Written English. De Gruyter Mouton. https://www.degruyter.com/database/COGBIB/entry/cogbib.1245/html

Braggaar, A., & van der Goot, R. (2021, April). Challenges in Annotating and Parsing Spoken, Code-switched, Frisian-Dutch Data. In *Proceedings of the second workshop on domain adaptation for nlp* (pp. 50–58). Kyiv, Ukraine: Association for Computational Linguistics. https://aclanthology.org/2021.adaptnlp-1.6

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

Caines, A., McCarthy, M., & Buttery, P. (2017). Parsing transcripts of speech. In *Proceedings of the workshop on speech-centric natural language processing* (pp. 27–36). Association for Computational Linguistics. http://aclweb.org/anthology/W17-4604

Čibej, J., Gantar, K., Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., ... Robnik-Šikonja, M. (2022). *Morphological lexicon Sloleks 3.0.* http://hdl.handle.net/11356/1745 (Slovenian language resource repository CLARIN.SI)

Čibej, J., & Munda, T. (2024). Metoda polavtomatskega popravljanja lem in oblikoskladenjskih oznak na primeru učnega korpusa govorjene slovenščine ROG. In *Zbornik konference jezikovne tehnologije in digitalna humanistika 2024.*

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. doi: 10.1162/COLI_a_00402

Dobrovoljc, K. (2018, December). Formulaičnost v slovenskem jeziku. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, *6*(2), 67–95. http://dx.doi.org/10.4312/slo2.0.2018.2.67-95 doi: 10.4312/slo2.0.2018.2.67-95

Dobrovoljc, K. (2022, June). Spoken Language Treebanks in Universal Dependencies: an Overview. In N. Calzolari et al. (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 1798–1806). Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lrec-1.191

Dobrovoljc, K. (2024, July). Skladenjska drevesnica govorjene slovenščine: stanje in perspektive. In *Stanje in perspektive uporabe govornih virov v raziskavah govora* (p. 41–62). Univerza v Mariboru, Univerzitetna založba. http://dx.doi.org/10.18690/um.ff.4.2024.3 doi: 10.18690/um.ff.4.2024.3

Dobrovoljc, K., Erjavec, T., & Krek, S. (2017, April). The Universal Dependencies Treebank for Slovenian. In T. Erjavec, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, & R. Yangarber (Eds.), *Proceedings of the 6th workshop on Balto-Slavic natural language processing* (pp. 33–38). Valencia, Spain: Association for Computational Linguistics. https://aclanthology.org/W17-1406 doi: 10.18653/v1/

W17-1406

Dobrovoljc, K., & Ljubešić, N. (2022, June). Extending the SSJ Universal Dependencies Treebank for Slovenian: Was It Worth It? In S. Pradhan & S. Kuebler (Eds.), *Proceedings of the 16th linguistic annotation workshop (law-xvi) within lrec2022* (pp. 15–22). Marseille, France: European Language Resources Association. https://aclanthology.org/2022.law-1.3

Dobrovoljc, K., & Martinc, M. (2018, November). Er ... well, it matters, right? On the role of data representations in spoken language dependency parsing. In M.-C. de Marneffe, T. Lynn, & S. Schuster (Eds.), *Proceedings of the second workshop on universal dependencies (UDW 2018)* (pp. 37–46). Brussels, Belgium: Association for Computational Linguistics. https://aclanthology.org/W18-6005 doi: 10.18653/v1/W18-6005

Dobrovoljc, K., & Nivre, J. (2016, May). The Universal Dependencies Treebank of Spoken Slovenian. In N. Calzolari et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 1566–1573). Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1248

Dobrovoljc, K., & Terčon, L. (2023). *Universal Dependencies: Smernice za označevanje besedil v slovenščini. Različica 1.3.* . Center za jezikovne vire in tehnologije Univerze v Ljubljani. https://wiki.cjvt.si/attachments/66

Dobrovoljc, K., Terčon, L., & Ljubešić, N. (2023, September). Universal Dependencies za slovenščino: Nove smernice, ročno označeni podatki in razčlenjevalni model. *Slovenščina 2.0: empirical applied and interdisciplinary research*, *11*(1), 218–246. http://dx.doi.org/10.4312/slo2.0.2023.1.218-246 doi: 10.4312/slo2.0.2023.1.218-246

Erjavec, T. (2010, May). MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In N. Calzolari et al. (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/138_Paper.pdf

Erjavec, T., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Fišer, D., Laskowski, C., & Zupan, K. (2016). Annotating CLARIN. SI TEI corpora with WebAnno. In *Proceedings of the clarin annual conference* (pp. 1–5).

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the 1992 ieee international conference on acoustics, speech and signal processing - volume 1* (pp. 517–520). Washington, DC, USA: IEEE Computer Society. http://dl.acm.org/citation.cfm?id=1895550.1895693

Hajič, J., Cinková, S., Mikulová, M., Pajas, P., Ptácek, J., Toman, J., & Uresová, Z. (2008). PDTSL: An annotated resource for speech reconstruction. In *Proceedings of the 2008 ieee workshop on spoken language technology* (pp. 93–96). IEEE.

Hinrichs, E., Bartels, J., Kawata, Y., Kordoni, V., & Telljohann, H. (2000). The Tübingen treebanks for spoken German, English, and Japanese. In W. Wahlster (Ed.), *Verbmobil: Foundations of speech-to-speech translation* (p. 550-574). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-662-04230-4_40 doi: 10.1007/978-3-662-04230-4_40

Hinrichs, E., & Kübler, S. (2005). *Treebank profiling of spoken and written German*. Universitätsbibliothek Johann Christian Senckenberg.

Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S., Velušček, A., ... Arhar Holdt, Š. (2023). *Specifikacije za učni korpus: lematizacija in MSD. Različica 2.0.* https://wiki.cjvt.si/attachments/21

Kahane, S., Caron, B., Strickland, E., & Gerdes, K. (2021, December). Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In D. Dakota, K. Evang, & S. Kübler (Eds.), *Proceedings of the 20th international workshop on treebanks and linguistic theories (tlt, syntaxfest 2021)* (pp. 35–47). Sofia, Bulgaria: Association for Computational Linguistics. https://aclanthology.org/2021.tlt-1.4

Kåsen, A., Hagen, K., Nøklestad, A., Priestly, J., Solberg, P. E., & Haug, D. T. T. (2022, June). The Norwegian Dialect Corpus Treebank. In N. Calzolari et al. (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 4827–4832). Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lrec-1.516

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., ... Zajc, A. (2021). *Training corpus ssj500k 2.3.* http://hdl.handle.net/11356/1434 (Slovenian language resource repository CLARIN.SI)

Krsnik, L., & Dobrovoljc, K. (2023). *The Trankit model for linguistic processing of standard Slovenian.* http://hdl.handle.net/11356/1870 (Slovenian language resource repository CLARIN.SI)

Krsnik, L., & Dobrovoljc, K. (2024). *Trankit model for linguistic processing of spoken Slovenian.* http://hdl.handle.net/11356/1909 (Slovenian language resource repository CLARIN.SI)

Krsnik, L., Dobrovoljc, K., & Robnik-Šikonja, M. (2024). *Dependency tree extraction tool STARK 3.0.* http://hdl.handle.net/11356/1958 (Slovenian language resource repository CLARIN.SI)

Lacheret-Dujour, A., Kahane, S., & Pietrandrea, P. (2019). *Rhapsodie: A prosodic and syntactic treebank for spoken French* (Vol. 89). John Benjamins Publishing Company.

Lenardič, J., Čibej, J., Arhar Holdt, Š., Erjavec, T., Fišer, D., Ljubešić, N., ... Dobrovoljc, K. (2022). *CMC training corpus janes-tag 3.0.* http://hdl.handle.net/11356/1732 (Slovenian language resource repository CLARIN.SI)

Liu, Z., & Prud'hommeaux, E. (2021, April). Dependency Parsing Evaluation for Low-resource Spontaneous Speech. In *Proceedings of the second workshop on domain adaptation for nlp* (pp. 156–165). Kyiv, Ukraine: Association for Computational Linguistics. https://aclanthology.org/2021.adaptnlp-1.16

Ljubešić, N., & Dobrovoljc, K. (2019, August). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th workshop on balto-slavic natural language processing* (pp. 29–34). Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-3704 doi: 10.18653/v1/W19-3704

Luotolahti, J., Kanerva, J., & Ginter, F. (2017). dep_search: Efficient search tool for large dependency parsebanks. In *Proceedings of the 21st nordic conference on computational linguistics* (pp. 255–258).

MacWhinney, B. (2014). *The childes project*. Psychology Press. https://doi.org/10.4324/9781315805641 doi: 10.4324/9781315805641

Nguyen, M. V., Lai, V. D., Pouran Ben Veyseh, A., & Nguyen, T. H. (2021, April). Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing. In D. Gkatzia & D. Seddah (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: System demonstrations* (pp. 80–90). Online: Association for Computational Linguistics. https://aclanthology.org/2021.eacl-demos.10 doi: 10.18653/v1/2021.eacl-demos.10

Øvrelid, L., Kåsen, A., Hagen, K., Nøklestad, A., Solberg, P. E., & Johannessen, J. B. (2018, May). The LIA Treebank of Spoken Norwegian Dialects. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)* (pp. 4482–4488). Miyazaki, Japan: European Language Resources Association (ELRA). https://aclanthology.org/L18-1710

Pietrandrea, P., & Delsart, A. (2019, June). Chapter 16. macrosyntax at work. In *Studies in corpus linguistics* (pp. 285–314). Amsterdam: John Benjamins Publishing Company.

Štravs, M., & Dobrovoljc, K. (2024). *Service for querying dependency treebanks Drevesnik 1.1.* http://hdl.handle.net/11356/1923 (Slovenian language resource repository CLARIN.SI)

Terčon, L., & Ljubešić, N. (2023). *The CLASSLA-Stanza model for UD dependency pars-*

ing of standard Slovenian 2.0. http://hdl.handle.net/11356/1769 (Slovenian language resource repository CLARIN.SI)

Terčon, L., & Ljubešić, N. (2023). *CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.*

van der Wouden, T., Hoekstra, H., Moortgat, M., Renmans, B., & Schuurman, I. (2002). Syntactic analysis in the Spoken Dutch Corpus (CGN). In *Proceedings of the third international conference on language resources and evaluation, LREC 2002, may 29-31, 2002, las palmas, canary islands, spain.* http://www.lrec-conf.org/proceedings/lrec2002/pdf/71.pdf

van der Wouden, T., Schuurman, I., Schouppe, M., & Hoekstra, H. (2003, January). Harvesting Dutch Trees: Syntactic Properties of Spoken Dutch. In *Computational linguistics in the netherlands 2002* (p. 129–141). BRILL. http://dx.doi.org/10.1163/9789004334441_011 doi: 10.1163/9789004334441_011

Verdonik, D., & Bizjak, A. (2023). *Pogovorni zapis in označevanje govora v govorni bazi Artur projekta RSDO.* https://dk.um.si/IzpisGradiva.php?lang=slv&id=85198

Verdonik, D., Bizjak, A., Sepesy Maučec, M., Gril, L., Dobrišek, S., Križaj, J., ... Dretnik, N. (2023). *ASR database ARTUR 1.0 (transcriptions).* http://hdl.handle.net/11356/1772 (Slovenian language resource repository CLARIN.SI)

Verdonik, D., Bizjak, A., Žgank, A., Bernjak, M., Antloga, Š., Majhenič, S., ... Bordon, D. (2023). *ASR database ARTUR 1.0 (audio).* http://hdl.handle.net/11356/1776 (Slovenian language resource repository CLARIN.SI)

Verdonik, D., Dobrovoljc, K., Erjavec, T., & Ljubešić, N. (2024, May). Gos 2: A New Reference Corpus of Spoken Slovenian. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 7825–7830). Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.691

Verdonik, D., Dobrovoljc, K., Čibej, J., Ljubešić, N., & Rupnik, P. (2024). Izbor in urejanje gradiv za učni korpus govorjene slovenščine - ROG. In *Zbornik konference jezikovne tehnologije in digitalna humanistika 2024.*

Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., & Stabej, M. (2013, January). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation*, *47*(4), 1031–1048. http://dx.doi.org/10.1007/s10579-013-9216-5 doi: 10.1007/s10579-013-9216-5

Verdonik, D., & Maučec, M. S. (2016, 02). A speech corpus as a source of lexical information. *International Journal of Lexicography*, *30*(2), 143-166. https://doi.org/10.1093/ijl/ecw004 doi: 10.1093/ijl/ecw004

Verdonik, D., Potočnik, T., Sepesy Maučec, M., Erjavec, T., Majhenič, S., & Žgank, A.

(2021). *Spoken corpus Gos VideoLectures 4.2 (transcription).* http://hdl.handle.net/11356/1444 (Slovenian language resource repository CLARIN.SI)

Verdonik, D., Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., ... Rupnik, P. (2023). *Spoken corpus Gos 2.1 (transcriptions).* http://hdl.handle.net/11356/1863 (Slovenian language resource repository CLARIN.SI)

Yimam, S. M., Gurevych, I., De Castilho, R. E., & Biemann, C. (2013). 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* (pp. 1–6).

Zeman, D., et al. (2023). *Universal Dependencies 2.12.* http://hdl.handle.net/11234/1-5150 (LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University)

Zeman, D., et al. (2024). *Universal Dependencies 2.14.* http://hdl.handle.net/11234/1-5502 (LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University)

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., & Erjavec, T. (2021). *Spoken corpus Gos 1.1.* http://hdl.handle.net/11356/1438 (Slovenian language resource repository CLARIN.SI)

RAZŠIRITEV DREVESNICE GOVORJENE SLOVENŠČINE SST

V prispevku predstavljamo novo različico drevesnice govorjene slovenščine SST (angl. *Spoken Slovenian Treebank*), uravnoteženega in reprezentativnega korpusa transkribiranega govora z ročno označenimi lemami, besednimi vrstami, oblikoslovnimi lastnostmi in skladenjskimi odvisnostmi med besedami. Izvorno različico drevesnice SST smo razširili z več kot 3.000 novimi izjavami in jo izboljšali z vidika poenotenja načel zapisovanja govora ter zanesljivosti ročno pripisanih oznak. Po kratki predstavitvi vzorčenja novih podatkov iz referenčnega korpusa govorjene slovenščine GOS 2 ter polavtomatskega oblikoslovnega označevanja v jedru prispevka opisujemo proces skladenjskega razčlenjevanja novih besedil ter poenotenja med prvotnimi in novo dodanimi transkripcijami, ki so se razlikovale na ravni segmentacije govora, rabe ločil in velikih začetnic. V drugem delu vsebino nove različice drevesnice SST povzamemo z vidika velikosti in raznolikosti podatkov in predstavimo rezultate njene primerjave z referenčno drevesnico pisne slovenščine SSJ, ki razkriva unikatne leksikalne in skladenjske lastnosti govorjenega jezika.

Keywords: korpusno označevanje, odvisnostna drevesnica, govorjeni jezik, spontani govor, Universal Dependencies

KAKO DOBER JE CHATGPT PRI UMEŠČANJU SOPOMENK POD BESEDNE POMENE

Magdalena GAPSA,[1] Špela ARHAR HOLDT,[1,2] Iztok KOSEM[1,2]

[1]Filozofska fakulteta, Univerza v Ljubljani

[2]Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

V raziskavi preverjamo, kako dobro se ChatGPT-4 odreže pri čiščenju seznama strojno pridobljenih sopomenskih kandidatov in umeščanju sopomenskega gradiva pod besedne pomene. Kot zlati standard upoštevamo slovaropisne odločitve, ki so bile sprejete pri nadgrajevanju Slovarja sopomenk sodobne slovenščine v različico 2.0. V prispevku analiziramo rezultate za 246 slovarskih iztočnic. Za 41,9 % iztočnic je ChatGPT podatke uredil povsem enako kot slovaropisci, za 58,1 % pa se je v odločitvi razlikoval: 43,5 % iztočnic je vsebovalo razlike pri odstranjevanju neustreznih sopomenskih kandidatov, 28,9 % pa pri razvrščanju sopomenk pod pomene. Pri presojanju relevantnosti sopomenskih kandidatov je bil ChatGPT popustljivejši od zlatega standarda (priklic 0,33), medtem ko je bila natančnost višja (0,75), vendar razlike težje pojasnljive. Razlike v razvrščanju sopomenk (umestitev pod drug pomen pri 14,6 % iztočnicah, manjkajoča umestitev pri 19,9 %) deloma pripisujemo značilnostim vhodnih podatkov, kot sta kompleksnost naloge in kratkost pomenskih indikatorjev. Bodoče delo bo usmerjeno v preizkus implementacije strojnega postopka za pohitritev slovaropisnega dela.

Ključne besede: digitalno slovaropisje, ChatGPT, sopomenke, besedni pomen, slovenščina

1. UVOD

Generativna umetna inteligenca, ki temelji na velikih jezikovnih modelih, je prek klepetalnih vmesnikov, kakršen je ChatGPT (OpenAI, 2024), postala široko dostopna za številne z jezikom povezane naloge. Med področji, ki zadnji dve leti preizkušajo moč in omejitve novih tehnologij, je tudi slovaropisje.

Dosedanji preizkusi rabe ChataGPT za slovaropisne namene se osredotočajo na generiranje bolj ali manj celostnih slovarskih gesel za (pogosto dokaj priložnostno) izbran nabor iztočnic (prim. de Schryver in Joffe, 2023; Barrett, 2023 v de Schryver, 2023; Rundell, 2023; Jakubíček in Rundell, 2023; Lew, 2023). Kot v svojem kritičnem pregledu prvih prispevkov (člankov in posnetkov predavanj oz. referatov) na temo z umetno inteligenco podprtega slovaropisja poroča de Schryver (2023), je trenutno največ pozornosti posvečene definicijam in primerom rabe (de Schryver, 2023, str. 17), manj pa preprostejšim slovaropisnim nalogam. (Skoraj) vse študije oz. preizkusi so bili izvedeni v angleščini in za angleščino, čeprav Jakubíček in Rundell (2023) naslavljata tudi problem večjezičnosti (str. 522–523).

Obstoječim raziskavam dodajamo preizkus, kako dobro se ChatGPT-4 odreže pri čiščenju seznama strojno pridobljenih sopomenskih kandidatov in umeščanju sopomenskega gradiva pod besedne pomene. Preizkus 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 (Krek in sod., 2017), od objave leta 2018 pa se ročno pregleduje in čisti v sodelovanju med strokovnjaki za slovaropisje ter zainteresirano uporabniško javnostjo (Arhar Holdt in sod., 2018).

Ob nadgradnji slovarja v različico 2.0 je bilo med slovaropisnimi nalogami tudi umeščanje strojno pripravljenih sopomenskih kandidatov pod besedne pomene. Cilj naloge je bil: (a) presoditi, ali je strojno pridobljena beseda oz. zveza sopomenska z iztočnico v vsaj enem od njenih slovarskih pomenov ali pa gre za neustrezno gradivo, ki ga je treba iz iztočnice odstraniti, ter (b) umestiti relevantno sopomensko gradivo pod vse ustrezne besedne pomene dane iztočnice (več v Arhar Holdt in sod., 2023; Gantar in sod., 2023).

V bodoče bi bilo v opisani slovaropisni postopek možno 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 bi lahko pomembno pohitrila nadgrajevanje slovarja. Prvi korak pri morebitni optimizaciji postopka je ugotoviti, kakšne rezultate daje ChatGPT v primerjavi z jezikoslovci, čemur se posvečamo v tem prispevku. V nadaljevanju predstavimo metodologijo raziskave in primerjamo rezulate strojnega in ročnega dela.

2. METODOLOGIJA

2.1 Opis podatkov

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 (SSSS 1.0, Krek in sod., 2018) in drugih prosto dostopnih leksikalnih virih (prim. Gapsa, 2022). 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 (oz. odstranjeni) in razvrščeni pod identificirane pomene.

Za izbranih 266 iztočnic je bilo v prvem koraku iz baze SSSS 1.0 izluščenih skupno 1.049 sopomenskih kandidatov (z morebitnimi področnimi slovarskimi oznakami). V drugem koraku so bile iz Digitalne slovarske baze (DSB, Kosem in sod., 2021) izvožene pomenske členitve s pomenskimi indikatorji (tj. kratkimi opisi za ločevanje pomenov) 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 (SSSS 2.0, Krek in sod., 2023) pridobili pomensko členjene iztočnice z razvrščenimi sopomenkami.

V prvem koraku analize je bilo med 266 iztočnicami odkritih 20 iztočnic, kjer se pomenska členitev iz DSB ne ujema s SSSS 2.0 (npr. iztočnica *bonbon* ima v DSB 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.

2.2 Struktura poziva za ChatGPT

Za izbrane iztočnice smo pripravili poziv za ChatGPT, 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.

> 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],[1] razlaga, utemeljitev, razmišljanje, smerni kot, udeleženska 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

[1] V oglatih oklepajih navajamo področne slovarske oznake.

- 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]
```

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 CSV datoteko in Excelovo datoteko z vsemi zbranimi podatki.

V poziv nismo vključili celotnih smernic, ki jim je sledila slovaropisna ekipa, saj bi s tem v postopek vnesli preveč informacij in spremenljivk, kar privede 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.

2.3 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. Rezultati sicer nakažejo, da je v določenih primerih rešitev, ki jo ponudi ChatGPT, drugačna od slovaropisne, vendar kljub temu sprejemljiva. Če bodo s ChatomGPT pripravljeni podatki vključeni v slovaropisne delotoke, bo v prihodnje treba presoditi, kako v praksi obravnavati take primere skladno z izbranim slovaropisnim konceptom.

2.4 Omejitve

Strojni postopek, ki ga preizkušamo, je odvisen od izbranega poziva, vhodnih podatkov in različice uporabljenega sistema. Pri delu s sistemom ChatGPT oblikovanje ubeseditve pomembno vpliva na rezultate, pri čemer vpliva ni mogoče povsem predvideti ali pojasniti. Vpliva lahko tudi, da smo oblikovanje poziva testirali s pomočjo brezplačne, ne pa tudi plačljive različice sistema. V nadaljnjem delu je mogoče poziv nadalje razvijati ali postopek razdeliti na dva dela (ločeno čiščenje in umeščanje).

Pomenski indikatorji so krajši in jedrnatejši od polnih pomenskih definicij, oblikovani tako, da izražajo glavne razlike med pomeni, ki so za sistem morda premalo povedne. Predvidevamo, da bi lahko raba daljših, sistematično strukturiranih slovarskih definicij rezultate izboljšala.

3. REZULTATI IN DISKUSIJA

3.1 Splošna uspešnost

Pri analiziranih 246 iztočnicah je ChatGPT v 103 primerih (41,9 %) podatke uredil povsem enako kot slovaropisci, v 143 primerih (58,1 %) pa se je v odločitvi tako ali drugače razlikoval.

Podatke s primeri iztočnic prikazuje Tabela 1, v kateri podajamo tudi povprečno število kandidatov ter slovarskih pomenov v posamezni od skupin. 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 in uspešno uredi kompleksnejše iztočnice.

Tabela 1: Ujemanje med slovaropisnimi odločitvami in odločitvami ChataGPT 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,	103	2,2	1,7

	anatomija			
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)

Natančnejša analiza je pokazala, da se med 143 iztočnicami pojavlja 107 takih, ki kažejo razlike na ravni odstranjevanja neustreznih sopomenskih kandidatov (43,5 % analiziranih iztočnic), 71 takih, ki kažejo razlike na ravni razvrščanja pod pomene (28,9 %), od tega pa je 35 primerov (14,2 %), kjer se pojavljajo tako razlike prvega kot drugega tipa.

3.2 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 od 951 (26,2 %) kandidatov. ChatGPT je odstranil le 110 kandidatov (11,6 %). 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 sopomenski kandidat, o katerem je ChatGPT presojal.

Tabela 2: Primeri in število pravilnih in napačnih odločitev pri presojanju ChataGPT, 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 –	167

	vzdrževalec telefonskega omrežja	
Neustrezno odstranjenih (FP)	adaptacija – predelava, anonimnost – nepoznanost, aplikacija – prekritje, atentat – umor, bife – prehranjevalnica, cenzura – predelava [tiskarstvo]	28
Skupaj		951

Tabela 3 prikazuje natančnost (kolikšen delež odstranjenih primerov so dejansko neustrezni sopomenski kandidati), priklic (kolikšen delež vseh neustreznih kandidatov je bil identificiran) in F1 (harmonično sredino obeh vrednosti).

Tabela 3: Natančnost (*precision*), priklic (*recall*) in F1 za obravnavo neustreznih sopomenskih kandidatov.

Natančnost	*Priklic*	*F1*
0,7455	0,3293	0,4568

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 (Gantar in sod., 2023, str. 161). Kot smo zapisali v Razdelku 2.2, 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 ustrezajoče 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 pojavljajo v rabi (**dražba – razpis del*). Razlike v navodilih pojasnijo del razlik. Pri morebitni uporabi ChataGPT za pohitritev ročnega dela bi bila ta odstopanja predvidljiva, hitro opazna in enostavno rešljiva.

V naboru neustrezno obdržanih so tudi mejni primeri, ki so bili zahtevni že za slovaropisno odločitev. Pri teh bi raba ChataGPT za pohitritev ročnega dela lahko doprinesla k lažjim, morda še širše vključujočim odločitvam. Na drugi strani so problematične odstranitve, ki se v rezultatih sicer pojavljajo relativno redko, kot denimo *atentat – umor, debelost – obilnost, kaos – razdejanje*. Pri tovrstnih primerih bi bila pri morebitni rabi postopka potrebna pozornost.

3.3 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 oz. vse ustrezne pomene glede na zlati standard. Umestitev pod neustrezen pomen se pojavi pri 36 iztočnicah (14,6 % analiziranih iztočnic), manjkajoča umestitev pri 49 iztočnicah (19,9 %), od tega je 14 (5,7 %) takih, kjer se pojavljata obe vrsti problema, tj. umestitev pod neustrezen pomen ter 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 oz. je ni umestil. Kot smo opozorili v Razdelku 2.3, ustreznost oz. neustreznost razumemo v razmerju do zlatega standarda, vendar se med rezultati pojavljajo tudi mejni primeri, kjer je lahko poleg slovaropisne odločitve sprejemljiva tudi odločitev ChataGPT.

Tabela 4: Primeri, število napak v iztočnicah, kjer je ChatGPT umestil sopomenko pod napačni 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. kandi-datov*	*Povpr. št. pomenov*
Umeščeno pod neustrezen pomen	• bazar – sejem [ekonomija]: pod 'orientalska tržnica' namesto 'prireditev' • depresija – deprimiranost: pod 'bolezen' namesto 'potrtost' • hazarder – igralec na srečo: pod 'kdor rad veliko tvega' namesto 'kdor rad stavi' • nakup – kupčija: pod 'nakupljeno blago' namesto 'dejanje' • pes – klinec: pod 'žival' namesto ' izprijena • ničvredna oseba [izraža	55	36	6,7	2,8

	negativen odnos]' • recept – furmula: pod 'za izdajo zdravila' namesto 'navodilo; pravilo'				
Neumeš-čeno pod pomen	• arhitektura – stavbarstvo: ustrezno pod 'veda in dejavnost', manjka pri 'o gradnji stavb' • bajka – zgodbica: ustrezno pod 'pripovedka [književnost]', manjka pri 'izmišljotina' • bolnik – pacient: ustrezno pod 'kdor je bolan', manjka pri 'kdor je neprijeten ali krut [izraža negativen odnos] • dialog – dvogovor: ustrezno pri 'pogovor med osebama', manjka pri 'med nasprotnima stranema' • gneča – množica: ustrezno pri 'o ljudeh', manjka pri 'o stvareh' • gneča – truma: ustrezno pri 'o ljudeh', manjka pri 'o stvareh'	78	49	5,3	2,9
Skupaj		133	71	5,3	2,8

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 (gl. Razdelek 3.5), vendar se to ne kaže enoznačno: kot primer uspešno urejenega kompleksnega gesla lahko podamo iztočnico *jagoda* s štirimi kratkimi in medsebojno podobnimi indikatorji (*1. rastlina, 2. plod, 3. okrogel plod, 4. droben predmet*), h katerim je sistem ustrezno razmestil sopomenske kandidate *biser, koralda, kroglica* ali iztočnico *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, telefonski klic*.

3.4 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 ChataGPT. Razlike so obarvane z zeleno.

Tabela 5: Prikaz obravnave najtežjih primerov z oznakami razlik med odločitvami.

Iztočnica	*Slovaropisno*	*ChatGPT*
adaptacija	1. prenova stavbe: *prilagoditev, prenova, preureditev* 2. priredba [v umetnosti]: *predelava [glasba], priredba [glasba], preureditev [glasba], aranžma [glasba]* 3. prilagoditev [ponavadi v športu]: PRAZNO Odstranjeni kandidati: /	1. prenova stavbe: *preureditev, prenova* 2. priredba [v umetnosti]: *priredba [glasba], aranžma [glasba]* 3. prilagoditev [ponavadi v športu]: *prilagoditev* Odstranjeni kandidati: *predelava [glasba]*
čarovnica	1. pravljično bitje: *čarodejka, coprnica* 2. ženska, obtožena čarovništva: *vračarica* 3. zlobna, starejša ženska [izraža negativen odnos]: *vešča, babura, coprnica, hudičevka, stara baba, zlobnica* 4. ženska z nadnaravnimi sposobnostmi [lahko izraža negativen odnos]: *čarodejka, coprnica* Odstranjeni kandidati: *iluzionistka*	1. pravljično bitje: *čarodejka, coprnica, vračarica, vešča* 2. ženska, obtožena čarovništva: *vešča, hudičevka* 3. zlobna, starejša ženska [izraža negativen odnos]: *babura, stara baba, zlobnica* 4. ženska z nadnaravnimi sposobnostmi [lahko izraža negativen odnos]: *iluzionistka* Odstranjeni kandidati: /
gurman	1. ljubitelj hrane in pijače: *sladokusec, uživač, gastronom, poznavalec hrane, epikurejec* 2. poznavalec: *sladokusec*	1. ljubitelj hrane in pijače: *sladokusec, uživač, bonvivan, poznavalec hrane, gastronom, epikurejec* 2. poznavalec: *bonvivan, poznavalec hrane, gastronom*

	Odstranjeni kandidati: *bonvivan*	Odstranjeni kandidati: /
kandidat	1. za službo ali funkcijo: *prosilec, aspirant, prijavljenec, interesent [ekonomija], prijavljeni* 2. na izpitu: *prijavljenec, prijavljeni, udeleženec* 3. za nagrado: *predlaganec, nominiranec* 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*	1. za službo ali funkcijo: *prosilec*, *aspirant*, *prijavljenec*, *potencialna kandidatka, možna kandidatka, predlaganec,* *prijavljeni*, *vložnik* 2. na izpitu: *izpitna kandidatka, izprašanec* 3. za nagrado: *nominiranec*, *volilna agitatorka, volilni agitator* 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* 2. o intimnem odnosu: *romanca, ljubezensko razmerje* 3. ljubljena oseba: *draga, ljuba, dragi, ljubi, fant, punca* 4. o dejavnosti: *naklonjenost, strast, nagnjenje, posebna naklonjenost, posebno nagnjenje* 5. seks: *strast* Odstranjeni kandidati: *nežnost, toplina, posebna nagnjenost, oboževanje, občudovanje, čustva*	1. močna naklonjenost: *naklonjenost*, *strast,* *nagnjenje*, *posebna naklonjenost*, *posebna nagnjenost,* *posebno nagnjenje*, *oboževanje, občudovanje, čustva* 2. o intimnem odnosu: *ljubezensko razmerje*, *romanca* 3. ljubljena oseba: *draga*, *ljubi*, *dragi*, *ljuba*, *fant*, *punca* 4. o dejavnosti: PRAZNO 5. seks: *eros* Odstranjeni kandidati: *nežnost, toplina*
urok	1. v magiji: *čarovnija, prekletstvo, zakletev, zaklinjanje, zarotitev, magične besede, coprnija, zarekanje, čarovniški*	1. v magiji: *čarovnija*, *prekletstvo*, *zakletev*, *zarotitev*, *magične besede*, *coprnija*, *zaklinjanje*, *abrakadabra, vudu, čaranje,*

	izrek, čarobni napev 2. o neuspehu [pogosto v športnem kontekstu]: *prekletstvo* Odstranjeni kandidati: *nesreča, smola, abrakadabra, vudu, čaranje, čarobna formula, čarobni simbol, molitveni obrazec [religija]*	*čarobna formula, čarovniški izrek, čarobni napev, čarobni simbol* 2. o neuspehu [pogosto v športnem kontekstu]: *nesreča, smola, zarekanje* Odstranjeni kandidati: *molitveni obrazec [religija]*

Primeri pregledno prikazujejo raznovrstne 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 pravljičnega bitja). Prav tako so lahko mestoma zavajajoče ubeseditve v pomenskih indikatorjih, ki jih slovaropisna ekipa interpretira na podlagi preostalih podatkov v DSB, pri nalogi za ChatGPT pa so bili predstavljeni brez dodatnega konteksta (npr. indikator 'poznavalec' pri iztočnici *gurman*, ki je v opoziciji do 1. pomena in se v prenesenem pomenu ne navezuje več na hrano/pijačo, česar ChatGPT ne razbere). Nekaj je primerov, pri katerih slovaropisci upoštevajo smernice, ki ChatuGPT niso bile podane (gl. 3.2), npr. pri (ne)vključevanju moško-ženskih parov (*kandidat – izpitna kandidatka*). Najti pa je tudi razlike, kjer so odločitve ChataGPT težko razložljive, npr. *kandidat 'za nagrado' – volilna agitatorka, volilni agitator*.

4. SKLEP IN NADALJNJE DELO

V raziskavi smo preverili, kako dober je ChatGPT za umeščanje sopomenskega gradiva pod besedne pomene. Analizirali smo rezultate razvrščanja 951 sopomenskih kandidatov za 246 slovarskih iztočnic. Za 41,9 % slednjih je strojni postopek vrnil rezultate, povsem skladne s slovaropisnimi. Pri ostalih 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 % analiziranih iztočnicah. Večina odstopanj je posledica popustljivosti sistema do sopomenskih kandidatov, ki jih je slovaropisna ekipa odstranila. Ker koncept SSSS načelno teži k vključevanju gradiva, slovarski vmesnik pa omogoča reakcijo uporabniške skupnosti na neustrezne kandidate, so ti odstopi manj problematični. V 28,9 % 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 ubeseditev indikatorjev, kot tudi specifik korpusnega gradiva, ki v slovaropisnih delotokih pogojuje pomensko členjenje in preverbo sopomenskosti. Natančnejši pregled primerov, v katerih se pojavljajo raznovrstne razlike, pokaže, da se ChatGPT tudi pri najtežjih primerih ne razlikuje radikalno, 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.

Večji izziv je nepredvidljivost postopka. Za razliko od preteklih metodologij strojnega pridobivanja jezikovnih podatkov je generiranje s programom ChatGPT manj gradivno in postopkovno transparentno, zato je nemogoče sklepati o vzrokih za posamezne odločitve. Ta značilnost pomembno omejuje evalvacijske raziskave, kot je naša, ne more pa biti razlog, da generativnih tehnologij v slovaropisju ne bi uporabljali in/ali ocenjevali. V nadaljevanju bi bilo smiselno ciljno preveriti hipotezo, da so pri preprostejših geslih rezultati zanesljivejši, preizkusiti pripravo gradiva za nove (preprostejše) iztočnice in nato testirati, ali strojna predpriprava slovaropisne odločitve pohitri ali ne.

Zanimivo bi bilo nadalje raziskati primere, ki jih ChatGPT za razliko od slovaropisne ekipe ni odstranil, in raziskavam, ki preverjajo razumevanje koncepta sopomenskosti med različnimi uporabniškimi skupinami slovarja (Gapsa, 2023; Gapsa in Arhar Holdt, 2023), dodati še »razumevanje« pri rabi ChataGPT. Raziskavo je mogoče ponoviti na zmogljivejših različicah ChataGPT ali drugih podobnih sistemih, z nadgrajenimi pozivi in na novem gradivu (npr. za razvrščanje uporabniško dodanih sopomenk ali protipomenk). V načrtu je raziskava, ki bo preverila, koliko na rezultate vpliva prisotnost oz. odsotnost slovarskih oznak. Preizkusiti 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.

ZAHVALA

Zahvaljujemo se vsem, ki so omogočili izvedbo raziskave. Pri pripravi raziskovalnih podatkov je sodeloval Timotej Petrič. Projekt Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL je v letih 2021–2022 financiralo Ministrstvo za kulturo Republike Slovenije. Raziskovalni program Jezikovni viri in tehnologije za slovenski jezik (št. P6-0411) in raziskovalni projekt Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti (št. J7-3159) sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije iz državnega proračuna.

LITERATURA

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C. in Robnik-Šikonja, M. (2018). Thesaurus of Modern Slovene: By the Community for the Community. V J. Čibej, V. Gorjanc, I. Kosem in S. Krek (ur.), *Proceedings of the XVIII EURALEX International Congress, Lexicography in Global Contexts, 17-21 July 2018, Ljubljana* (str. 401–410). Znanstvena založba Filozofske fakultete. https://doi.org/10.4312/9789610600961

Arhar Holdt, Š., Gantar, P., Kosem, I., Pori, E., Robnik Šikonja, M. in Krek, S. (2023). Thesaurus of Modern Slovene 2.0. V M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček in S. Krek (ur.), *Electronic lexicography in the 21st century (eLex 2023), Proceedings of the eLex 2023 conference, 27–29 June 2023* (str. 366–381). Lexical Computing CZ. Pridobljeno 28. maja 2024, https://elex.link/elex2023/wp-content/uploads/82.pdf

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography, 36*(4), 355–387. https://doi.org/10.1093/ijl/ecad021

de Schryver, G.-M. in Joffe, D. (2023). The End of Lexicography, Welcome to the Machine: On How ChatGPT Can Already Take over All of the Dictionary Maker's Tasks. *20th CODH Seminar, ROIS-DS Center for Open Data in the Humanities, Tokyo, Japan, 27 February 2023* [Referat]. Pridobljeno 28. maja 2024, http://codh.rois.ac.jp/seminar/lexicography-chatgpt-20230227/

Gantar, P., Arhar Holdt, Š., Kosem, I. in Krek, S. (2023). Sopomenke 2.0 in Kolokacije

2.0: Novi koraki za slovenske odzivne slovarje. *Jezik in Slovstvo, 68*(4), 157–175. https://doi.org/10.4312/jis.68.4.157-175

Gapsa, M. (2022). Ocenjevanje uporabniško dodanih sopomenk v Slovarju sopomenk sodobne slovenščine – pilotna študija. V D. Fišer in T. Erjavec (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 15.–16. september 2022, Ljubljana, Slovenija* (str. 308–316). Inštitut za novejšo zgodovino. Pridobljeno 28. maja 2024, https://nl.ijs.si/jtdh22/pdf/JTDH2022_Gapsa_Ocenjevanje-uporabnisko-dodanih-sopomenk-v-Slovarju-sopomenk-sodobne-slovenscine.pdf

Gapsa, M. (2023). "But why??" Evaluation of user-suggested synonyms in the Thesaurus of Modern Slovene. [Preprint]. https://doi.org/10.21203/rs.3.rs-2775161/v1

Gapsa, M. in Arhar Holdt, Š. (2023). How Lexicographers Evaluate User Contributions in The Thesaurus of Modern Slovene in Comparison to Dictionary Users. V M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček in S. Krek (ur.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference. Brno, 27–29 June 2023* (str. 178–200). Lexical Computing CZ. Pridobljeno 28. maja 2024, https://elex.link/elex2023/wp-content/uploads/47.pdf

Jakubíček, M. in Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography?. V M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček in S. Krek (ur.), *Electronic lexicography in the 21st century (eLex 2023), Proceedings of the eLex 2023 conference, 27–29 June 2023* (str. 518–533). Lexical Computing CZ. Pridobljeno 28. maja 2024, https://elex.link/elex2023/wp-content/uploads/102.pdf

Kosem, I., Krek, S. in Gantar, P. (2021) Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. V Z. Gavriilidou, L. Mitits in S. Kiosses (ur.), *EURALEX XIX: Congress of the European Association for Lexicography, Lexicography for inclusion, 7-9 September 2021, Virtual, Book of Abstracts* (str. 81–83). Democritus University of Thrace. Pridobljeno 29. maja 2024, https://euralex.org/wp-content/uploads/2022/04/ABS2020.pdf

Krek, S., Laskowski, C. in Robnik-Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. V I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek in V. Baisa (ur.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference: Lexicography from Scratch. Leiden, the Netherlands, 19–21 September 2017* (str. 93-109). Dutch Language Institut, Lexical Computing CZ s.r.o., Trojina. Pridobljeno 28. maja 2024, https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf

Krek, S., Laskowski, C., Robnik-Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B. in Dobrovoljc, K. (2018). *Thesaurus of Modern Slovene 1.0*. Repozitorij raziskovalne strukture CLARIN.SI. http://hdl.handle.net/11356/1166

Krek, S., Laskowski, C., Robnik-Šikonja, M., Kosem, I., Arhar Holdt, Š., Gantar, P., Čibej, J., Gorjanc, V., Klemenc, B., Dobrovoljc, K., Pori, E., Roblek, R. in Zgaga, K. (2023). *Thesaurus of Modern Slovene 2.0*. Repozitorij raziskovalne strukture CLARIN.SI. http://hdl.handle.net/11356/1916

Lew, R. (2023). ChatGPT as a COBUILD Lexicographer. *Humanities and Social Sciences Communications 10*, članek 704 (2023). https://doi.org/10.1057/s41599-023-02119-6

OpenAI. (2024). *ChatGPT* (31. 5. 2024) [veliki jezikovni model]. https://chatgpt.com

Rundell, M. (2023). 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* (str. 9–17). Yonsei University. Pridobljeno 28. maja 2024, https://www.asialex.org/pdf/Asialex-Proceedings-2023.pdf

HOW GOOD IS CHATGPT AT PLACING SYNONYMS UNDER WORD SENSES

In this study, we test how well ChatGPT-4 cleans the list of automatically retrieved synonym candidates and distributes the synonyms under appropriate lexical senses. As a gold standard, we consider the lexicographic decisions made when updating the Thesaurus of Modern Slovene to version 2.0. In this paper, we compare the results for 246 dictionary entries. For 41.9% of entries, ChatGPT processed the data in the same way as lexicographers, while for 58.1%, it made a different decision: 43.5% of entries contained differences in the removal of noisy data, and 28.9% in the mapping of synonyms to lexical senses. When assessing the relevance of synonym candidates, ChatGPT is more permissive than the gold standard (recall 0.33), while precision is higher (0.75), but the differences are more difficult to explain. Differences in synonym placement (placement under a different sense in 14.6% of entries, missing placement in 19.9%) are partly attributed to features of the input data, such as task complexity and brevity of semantic indicators. Future work will focus on the validation of the method for speeding up lexicographic work.

Keywords: digital lexicography, ChatGPT, synonyms, word senses, Slovene language

GENERATIVE AI FOR COMPUTATIONAL CREATIVITY CONCEPTUALIZATION

Boshko KOLOSKI,[1,2] Senja POLLAK[1],
Geraint Wiggins[3, 4], Nada LAVRAČ[1]
[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3] Vrije Universiteit Brussel, Brussel, Belgium
[4] Queen Mary University of London, London, United Kingdom

Concept Creation Technology is concerned with engineering software for (semi-)automated domain conceptualization. This paper presents an approach to automatizing the conceptualization of the Computational Creativity (CC) domain, exploiting Generative AI (GAI) as a general-purpose Concept Creation Technology tool. The approach is showcased on the task of CC domain conceptualization using all publicly available proceedings from ICCC-2010 to ICCC-2023, as well as on the task of automated table of contents (ToC) structuring of individual Proceedings of the International Conferences on Computational Creativity. The implemented GAI methodology facilitates automated conceptualization of any domain of interest, automated ToC structuring of any proceedings or document corpus, as well as experiment replicability and software reuse.

Keywords: generative AI, computational creativity, natural language processing

1 INTRODUCTION

Computational Creativity (CC) is concerned with engineering software that exhibits creative behavior (Boden, 2004; Colton & Wiggins, 2012). A part of CC research addresses *Concept Creation Technology*, concerned with engineering software that exhibits creative behavior of *conceptualization*. In information science, conceptualization is defined as “an abstract (simplified) view of some selected part of the world, containing the objects, concepts, and other entities that are presumed of interest for some particular purpose and the relationships between them”, usually formalized through ontologies (Gruber, 1993; Smith, 2003). Manual construction, maintenance and updating of ontologies represents a significant investment of human resources, which is not always available and/or needed. More contemporary and less resource-consuming domain categorization approaches are currently available, using methods for automated extraction of domain knowledge from unstructured texts. These include automatic taxonomy construction (Fortuna et al., 2007; Kozareva & Hovy, 2010; Navigli & Ponzetto, 2012), knowledge graph construction (Y. Wei et al., 2023; Ye et al., 2022), and topic modeling that allows for domain conceptualization without explicit relations between concepts (Yao et al., 2018; Porturas & Taylor, 2021; Grootendorst, 2022). As this paper shows, the research vision of developing a fully automated technology for ICCC domain conceptualization has been fulfilled in this work, using novel Generative AI (GAI) methods to support the understanding of the conceptual structure of any research field represented by the papers published in conference proceedings.

This paper explores the potential of Large Language Models and Generative AI (GAI) acting as an advanced Concept Creation Technology tool that can be applied to any domain of interest. The proposed GAI approach facilitates automated conceptualization of any domain of interest, automated proceedings ToC generation for any proceedings or paper corpus.

The paper is structured as follows. After a brief outline of related work on CC domain conceptualization, Large Language Models (LLMs) and Generative AI (GAI), we describe the data used in the study, followed by the presentation of the proposed GAI-based methodology. We then present the results of the methodology on the task of automated CC domain conceptualization, covering all publicly available proceedings from ICCC-2010 to ICCC-2023. Moreover, the approach has been applied also to the table of contents (ToC) structuring of papers published in the ICCC-2023 Proceedings edition. Finally, we discuss experiment replicability and software reuse.

As for every other research community, Computational Creativity (CC) field conceptualization is also an interesting research field. (Loughran & O'Neill, 2017) have studied the CC domain by analyzing its conference proceedings, where conceptual categorization was conducted subjectively, through the review of each paper. This paper addresses automated CC conceptualization, following the line of past research in this area.

2 RELATED WORK

2.1 CC Domain Conceptualization

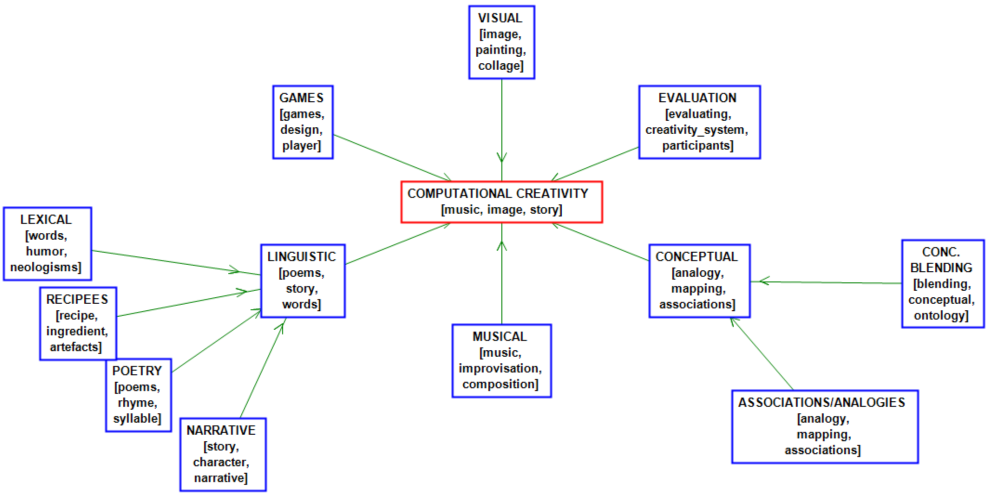


Figure 1: Semi-automatically generated conceptualization of the CC, with semi-automated concept naming and subconcept creation, using papers from 2010–2015 ICCC proceedings.

In related work by (Pollak et al., 2016), automated CC conceptualization was addressed using a semi-automated topic ontology construction tool OntoGen from 2010–2015 ICCC proceedings papers. The resulting corpus-based categorization of the CC field identified the following main CC subdomains: Musical, Visual, Linguistic creativity, Conceptual creativity, Games and creativity, with a manually added sub-domain of Evaluation through manual query used in the active learning approach to topic ontology creation. For several subdomains, subcategories were detected at a lower level, including Narratives, Poetry, Recipes and Lexical creativity as subdomains of Linguistic creativity, shown in Figure 1, visualized by (Pollak et al., 2016). Later, an extended corpus of 2010–2017 ICCC proceedings papers using k-means clustering and LSI cluster visualization techniques was analyzed (Podpečan et al., 2018). For instance, when analyzing again the ICCC proceedings papers from ICCC 2010–2015, the CC domains are very clearly separated, allowing the expert to easily recognize the topics (e.g., musical creativity, visual creativity, story generation, poetry generation, culinary creativity, conceptual creativity, etc.), where k-means with k=11 proved to provide optimal clusters in terms of the Silhouette score. Nevertheless, automated discovery of the optimal number of clusters using the Silhouette score gave non-conclusive results, as the results did not fully align with human conceptualization based on 2D visual cluster representations.

2.2 NLP, Large Language Models and Generative AI

The field of natural language processing has witnessed a remarkable transformation with the advent of Large Language Models (LLMs), which can be divided into two groups: Masked Language Models (MLMs) such as BERT (Devlin et al., 2019) and generative Causal Language Models (CLMs) such as LLaMa2 (Touvron & et al., 2023). These foundational models have set new standards for understanding and generating text with human-level precision (Min et al., 2023). Building on this groundwork, the sentence transformers (Reimers & Gurevych, 2019) have emerged as a specialized evolution. These models, which are tailored to the task of learning sentence representations, ensure that semantically similar sentences are closely aligned in the vector space. One notable application of sentence transformers is BERTopic (Grootendorst, 2022), which has revolutionized topic modeling with its unique approach. BERTopic clusters sentences based on semantic similarity, providing a refined and context-sensitive thematic analysis that outperforms conventional methods. In a similar vein, KeyBERT (Grootendorst, 2020) advances the field of keyword extraction. Utilizing sentence-transformer technologies, it effectively extracts key terms and phrases from extensive texts (Škrlj et al., 2022; Koloski et al., 2022). A pivotal area of research in the use of these models is domain adaptation. (Wang et al., 2021) proposed an approach for unsupervised domain adaptation, employing sequential denoising auto-encoders to learn from corrupted data. Another approach to domain adaptation involves generative pseudo-labeling (GPL) (Wang et al., 2022), where researchers use a surrogate generative model, such as T5 (Raffel et al., 2020) that is trained to generate queries for specific passages (Thakur et al., 2021). These queries are then ranked by a cross-encoder (Reimers & Gurevych, 2020) and used as downstream fine-tuning data for the sentence transformer. The development of prompting techniques in LLMs (J. Wei et al., 2022), particularly in-context oneshot learning (Lampinen et al., 2022), represents a significant stride in model interaction. This approach involves crafting specific prompts that enable models to learn from a single example within the prompt context, thereby generating more relevant and contextually nuanced responses. This technique is crucial in eliciting accurate and specific outputs from models like LLaMa2 (Touvron & et al., 2023), demonstrating a high level of understanding and flexibility in language generation (Pan et al., 2023).

3 EXPERIMENTAL DATA AND PROBLEM DEFINITION

In this section we present the data acquisition approach, followed by exploratory data analysis over the ICCC proceeding from the period of 2010 to 2023.

3.1 Data acquisition

We have extracted the data from two sources: actual ICCC proceedings in PDF format and the DBLP entries of individual proceedings articles. The ICCC proceedings for all years except for 2011 are available as single PDF files that were easy to process, whereas for 2011, individual articles are downloadable from the ICCC-2011 Proceedings Web page.

To enrich the PDF data available to us, we used DBLP as a resource to utilize the Bibtex metadata references for the ICCC proceedings as a second way of preparing the data for our system. We found dBLp entries for all years except for 2023. Note that by collecting data from dBLp, we have generalized our approach to work not only with the ICCC proceedings papers, but with any available conference proceedings paper collection, so that a specialized table of contents structuring system can be created for any conference of user's interest.

Data acquisition resulted in the ICCC corpus, which consists of the articles from all the Proceedings of the International Conference on Computational Creativity, published in 2010–2023. The entire 2010–2023 ICCC Proceedings corpus consists of 689 articles (see Figure 2).[1]

[1] Please note that there may be slight differences between the number of articles in the corpus and the actual proceedings, as the PDF corpus from 2010 to 2023 was collected manually and

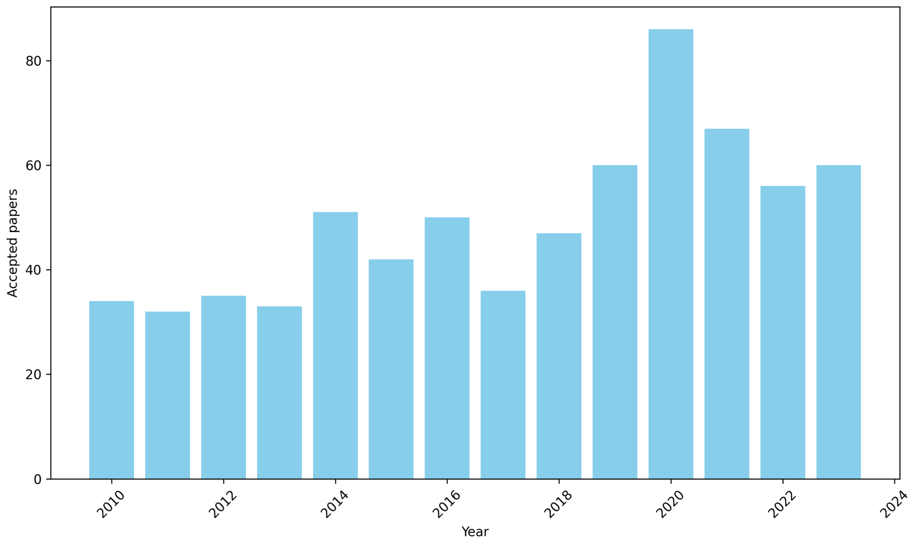


Figure 2: Distribution of articles per year.

3.2 Problem definition and motivation

Let us define the addressed problem, followed by a real-world instance of the problem.

Problem definition: The problem addressed in this paper is defined as follows: Given a set of proceedings papers, propose the topics and the structure of the Table of Contents.

Motivation: ICCC Proceedings editors have a tradition of structuring the proceedings table of contents (ToC) into Parts, containing articles on individual research topics. For example, in the ICCC-2010 Proceedings, there were 12 topics: Music – Patterns and Harmony; Visual Art; Analogy and Metaphor; Stories; Social Aspects; Foundations; Music – Creation/Generation; Creativity Support: Tools; Creativity Support: Applications; Music – Improvisation and Interaction; Evolution and Design; Linguistics; Show and Tell Session. In the ICCC-2022 Proceedings, there were 7 topics: Generating narratives; Co-creative systems;

we cannot exclude human errors, while the remaining years we used the bibtex metadata for automatic crawling, but we found a few minor inconsistencies

Theory and problem solving; Generative art; Music and applications; Creative meaning; Social aspects and evaluation.

4 DOMAIN CONCEPTUALISATION METHODOLOGY

We closely followed the work of Koloski et al. (2024) for applying contextual large language model technologies for domain conceptualization. The methodology consists of the following steps:

1. Data pre-processing and normalization,
2. Domain adaptation using the model T5,
3. Topic modeling using BERTopic, and
4. Topic naming using LLaMa2.

The individual steps are briefly outlined below.

4.1 Data pre-processing and normalization

For each year (14 data points in total), we scraped the ICCC proceedings articles metadata from dBLp via BeautifulSoup[2]. We analyzed the articles PDFs with the fitz library, and parsed the text with regex to obtain only printable word forms (this dataset is not made available for IPR reasons). On the other hand, we have made available the entire dBLp ICCC proceedings dataset[3]. Note, DBLP did not contain an entry for 2023 proceedings of the ICCC and we mined them by hand, with the exception of demo articles. We combine the information of the start and the end of an article from the dBLP bibtex entries to navigate through the PDF articles, automating the process of data preparation.

4.2 Domain adaptation

Domain adaptation was performed through generative pseudo-labeling (GPL) (Wang et al., 2022) using a surrogate generative model T5 (Raffel et al., 2020) that has been trained to generate queries for specific passages (Thakur et

[2] https://pypi.org/project/beautifulsoup4/
[3] https://dblp.org/db/conf/icccrea/index.html

al., 2021). These queries are then ranked by a cross-encoder (Reimers & Gurevych, 2020) and used as downstream fine-tuning data for the sentence transformer. Following the work of Koloski et al. (2024) we train the *all-mini-LM-v12* sentence-transformer on 100000 steps on the combined corpora consisting of our ICCC corpus and the ArXiV corpus (Muennighoff et al., 2023). We save evaluation checkpoints at every 10,000 steps. Note that input only the first 500 words of individual articles were used as input to the sentence-transformer model as the model is limited on the input size due to its transformer-based architecture.

4.3 Topic modeling

Topic modeling was performed with BERTopic (Grootendorst, 2022) to cluster sentences based on their semantic similarity. BERTopic starts with dimensionality reduction of the underlying sentence transformer embeddings using UMAP (McInnes et al., 2018), followed by the application of a clustering algorithm. We explored two families of algorithms to identify unique clusters:

- HDBSCan (McInnes et al., 2017), where the number of clusters was inferred from the data by utilizing the cluster density heuristic.
- KMeans (MacQueen, 1967), where the number of clusters was fixed to 10, to align our results with related work (Podpečan et al., 2018).

Finally, for each cluster, the KeyBERT (Grootendorst, 2020) keyword extraction technique was applied to find and evaluate relevant keywords and phrases that best describe each cluster.

4.4 Using Generative AI for topic naming

Using the extracted keywords and the most central documents in each thematic cluster, the language model LLaMa2 (Touvron & et al., 2023) was used to generate meaningful semantic labels for each cluster. The crucial component of this methodology was the prompting instructions given to the LLaMa2 language model. Li et al. (2023) found that the way a particular prompt is phrased has a direct impact on performance.

4.4.1 CRAFTING OF THE TOPIC EXTRACTION PROMPTS

Next we show how the topic extraction prompts were designed as combination of system, examples and query parts:

- **System Prompt**

  You are a helpful, respectful and honest assistant that is helping Program Chair of a Conference on computational creativity for creating Table of Contents topics for conferences.

- **Examples Prompt**
  - **Example 1**

    I have a topic that contains the following documents:
    - Computational Filling of Curatorial Gaps in a Fine Arts Exhibition
    - Visual Conceptual Blending with Large-Scale Language and Vision Models
    - What Does it Take to Cross the Aesthetic Gap? The Development of Image Aesthetic
    The topic is described by the following keywords: 'computational, images, language, aesthetic, curation'.
    Based on the information about the topic above, please create a short and concise table of content label of this topic.
    Make sure that you exclude the following terms: 'innovation, creativity'.
    Make sure you to only return the label and nothing more.
    LABEL Art & Aesthetics

- **Example 2**

  Another example:
  I have a topic that contains the following documents:
  - LyricJam: A System for Generating Lyrics for Live Instrumental Music
  - Being Creative: A Cross-Domain Mapping Network
  - Melody Similarity and Tempo Diversity as Evolutionary Factors for Music Variations by Genetic Algorithms
  The topic is described by the following keywords: ’creative, music, cross-domain, generation’.
  Based on the information about the topic above, please create a short and concise table of content label of this topic.
  Make sure that you exclude the following terms: ’innovation, creativity’.
  Make sure you to only return the label and nothing more.
  LABEL Musical Creativity

- **Query Prompt**

  I have a topic that contains the following documents: [DOCUMENTS]
  The topic is described by the following keywords: ’[KEYWORDS]’.
  Based on the information about the topic above, please create a short and concise table of content label of this topic.
  Make sure that you exclude the following terms: ’innovation, creativity’.
  Make sure you to only return the label and nothing more.

5 RESULTS OF ICCC DOMAIN CONCEPTUALIZATION

We evaluated the results of the methodology on all domain adaptation checkpoints (10 checkpoints, 1 on each 10,000 step of the 100,000 adaptation steps), for both clustering algorithms.

As HDBScan generated a large family of not fully consistent clusters, which were of not great quality according to the domain expert’s opinion, we decided to use KMeans in our further analysis.

We found that the best clustering results were obtained by the KMeans algorithm fixed at 10 neighbors for 20,000 adaptation steps.

The output of the used methodology is a list of topics. We propose these topics shall be used as categories for the structuring of further table of contents lists of ICCC proceedings.

In the following subsections, we first describe the topic modeling results for table of contents generation, followed by the analysis of the hierarchy that emerged between the topics, and finally, we explore how the topics evolved through time.

Table 1: Topics inferred from the data and their corresponding representative documents and keywords.

Topic Name	KeyBERT	Representative Document
Visual Conceptual Blending	['blending', 'visual', 'concepts']	Cunha, Joao M., et al. "A pig, an angel and a cactus walk into a blender: A descriptive approach to visual blending."
Computational Creativity Review	['creativity', 'research', 'design']	Mumford, Martin, and Dan Ventura. "The man behind the curtain: Overcoming skepticism about creative computing."
Creative Systems and Approaches	['emoji', 'based', 'data']	Agres, Kat, et al. "Conceptualizing Creativity: From Distributional Semantics to Conceptual Spaces."
Computer Vision and Machine Learning	['generative', 'generation', 'casual']	Compton, Katherine. "Casual creators"
Computational Creativity Frameworks	['surprise', 'images', 'self']	Mondol, Tiasa, and Daniel G. Brown. "Incorporating Algorithmic Information Theory into Fundamental Concepts of Computational Creativity."
Poetic Expressions	['poetry', 'creativity']	Toivanen, Jukka, et al. "Corpus-based generation of content and form in poetry."
Narrative and Storytelling Technologies	['stories', 'narrative']	Mckeown, Lewis, and Anna Jordanous. "An evaluation of the impact of constraints on the perceived creativity of narrative generating software."
Architecture and Design	['humor', 'computational', 'musical']	Brown, Daniel, and Dan Ventura. "Ethics, Aesthetics and Computational Creativity."
Computerized Art and Design	['metaphor', 'style', 'creative', 'art']	Righetti, Guendalina, et al. "A Game of Essence and Serendipity: Superb Owls vs. Cooking-Woodpeckers."
Philosophical Perspectives on Computational Creativity	['creativity', 'research', 'design', 'evaluation']	Llano, Maria Teresa, et al. "Explainable computational creativity."

5.1 Topic results

In Figure 3 we present the resulting 10 topics produced by the model from all the 2010-2023 ICCC proceedings papers. For example, the most prominent topics were Architecture and Design (15% of the articles), followed by Computational Creativity Review (13%), Philosophical Perspectives on Computational Creativity (12%) and Creative Systems and Approaches (11%). Table 1 presents the ten topics together with the respective keywords generated by KeyBERT, and the most central document representing the document cluster. Figure 4 shows a visual representation of the resulting clustering, while Figure 5 presents the distribution of detected concepts per year.

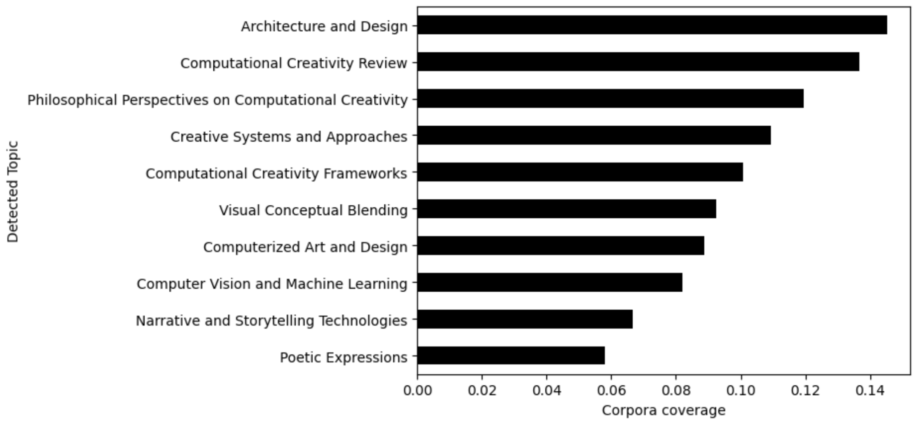


Figure 3: Detected topics and corpora coverage.

5.2 Time evolution of CC field conceptualization

Next, we analyze how the 10 derived topics of ICCC publications evolve over time (Figure 6). For each year in the 2010–2023 period, we measured the distribution of topics across articles. We found out that there were two disruptive periods. The first was in 2015, when the prominence of Philosophical Perspectives on Computational Creativity began to wane, and remained low until 2023. In the same year, Visual Conceptual Blending became very prominent, but fell back to its previous level over the following two years. We speculate that this was due to the emergence of computational frameworks for the massive analysis and generation of computational visual perspectives – the disruptive impact of AlexNet: a few years earlier (Krizhevsky et al., 2012). A change point analysis suggests that a second disruption came in 2020. We speculate that this could be due to the advent of pre-trained, transformer-based, large-scale language models such as the discriminative BERT (Devlin et al., 2019) and generative GPT (Radford et al., 2018). However, in this year there were certainly also some effects of the COVID-19 pandemic, therefore it is difficult to draw firm conclusions.

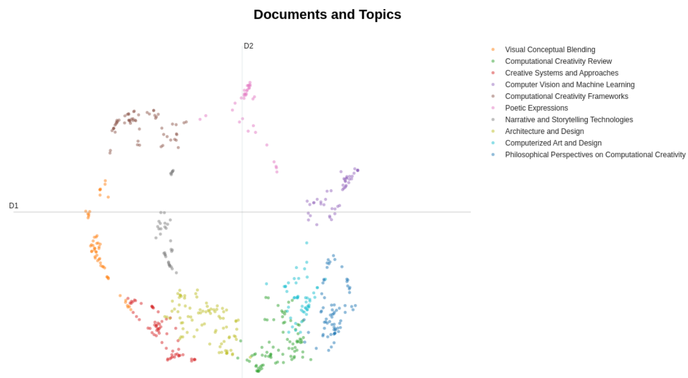


Figure 4: Spatial distribution of the articles for different years.

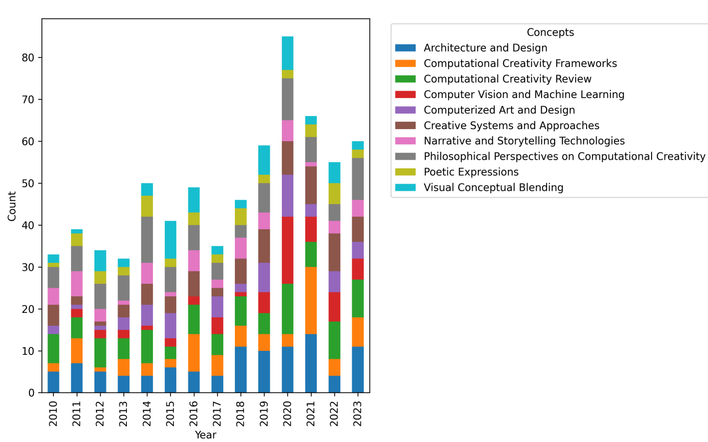


Figure 5: Distribution of detected concepts per year.

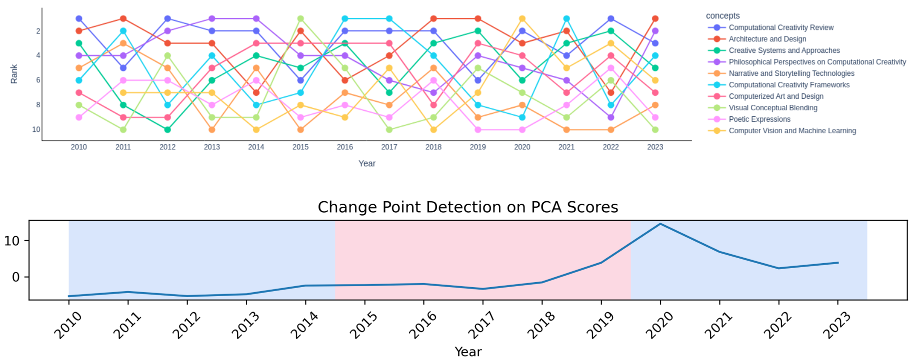


Figure 6: Above: Ranking of topics through the 14 years of ICCC proceedings. Below: Principle component analysis for change point detection over the distribution of ranks of topics through the years.

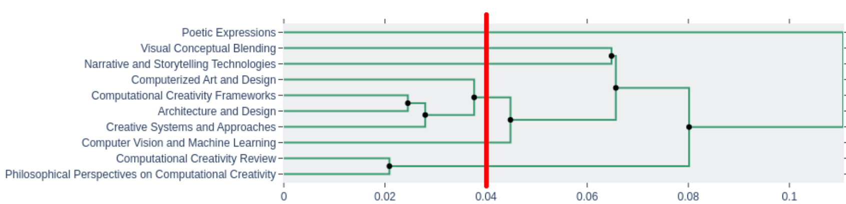


Figure 7: Hierarchical clustering of the inferred concepts.

5.3 Emergence of topic hierarchy

Next, we show the results of hierarchical clustering, i.e. the emergence of hierarchies between the inferred topics (see Figure 7). Since we began with a fixed number of topics resulting from 10-means clustering (k=10), we initially extracted 10 different topics. However, we subsequently grouped them by cluster similarity and decided to put them into 6 final topic categories (as suggested by the domain expert when inspecting the individual cluster names, and as suggested by increased cluster dissimilarities after the hierarchy cut-point at level 0.04). Cluster merging and merged cluster naming was performed as follows:

We merged the 'Computational Creativity Review' topic and the 'Philosophical Perspective on Computational Creativity' topic into a joint topic category. We then asked the LLaMa2 GAI model to name this cluster (after being prompted by the above two topic names). LLaMa2 proposed the following top-3 topic names: 'Computational Aesthetics', 'The Philosophy of Creative Machines' and 'AI-Powered Artistic Expression'. Among these, the expert's choice was the following concept name:

- The Philosophy of Creative Machines

Next, based on cluster similarity, the 'Architecture and Design' topic was coupled with the 'Computational Creativity Frameworks', followed by the 'Creative Systems and Approaches', the 'Computerized Art and Design' topic, showcasing the interactions between different layers of creativity with the different frameworks, resulting in various different downstream applications. Top-3 topic names proposed by the LLaMa2 GAI model (after being prompted by the above four topic names) were: 'Computational Creativity in Practice', 'Creative Systems and Techniques', and 'Innovation and Collaboration in Computational Creativity'. Among these, the expert's choice was the following concept name:

- Creative Systems and Techniques

The remaining topics were considered being relatively independent, and could actually be used as such in the proceedings ToC structure.

- Computer Vision and Machine Learning
- Narrative and Storytelling Technologies
- Visual Concept Blending
- Poetic Expression

5.4 Use Case: Structuring the ICCC-2023 Proceedings

Let us compare the above results obtained by the proposed GAI-based approach with the actual ICCC-2023 Proceedings structure, which was structured by the Proceedings editors into the following 8 topics[4]:

- Language and Storytelling
- Co-creativity
- Evaluation
- Image Generation and Processing

[4] Note that in dataset preparation, we decided not to include the Demo section papers, as the Demo proceedings subpart refers to paper type and not to the topic contents.

- Sound and Music
- Climate Change, Diversity, Equity, and Inclusion
- Interaction and Collaboration
- Demo

The expert comparison and the evaluation of the actual ICCC-2023 structuring into 8 categories (7 content-based topics + Demo) and the ICCC-2023 proceedings structure into 10 categories proposed by our GAI-based system, or alternatively, into the 6 categories proposed above, obtained after merging the most similar topic clusters, is a matter of further work, to be evaluated by human annotators. It should be noted, however, that there is an interesting tension between choosing a ToC structure that directly reflects the paper content, and choosing one that reflects the editors' perception of the important issues in the field as a whole.

In the future, e.g., when structuring the papers accepted for ICCC-2024, the proposed GAI-based ToC structuring technology could be readily applied, by learning the clustering from 2010-2023 ICCC proceedings (or from a different training subset, defined by the editor), and applied to categorization and topic naming of the ICCC-2024 proceedings. A similar analysis over the whole range of ICCC proceedings might yield interesting insights regarding the key areas of interest in the field as a whole.

6 EXPERIMENT REPLICABILITY AND SOFTWARE REUSE

We intend to publish our code for further structuring and generating tables of contents. Since our approach is based on pretrained models, only a few steps of optional domain adaptation are required. Moreover, our approach leverages open data sources such as DBLP, further generalizing the method for structuring and conceptualizing any conference.

7 CONCLUSIONS AND FUTURE WORK

In this work, we focused on the application of contextual large language models to conceptualize a corpus of documents from proceedings and generate tables of contents. We improve on previously published work both in terms of technology (we exploit contextual embeddings and use generative models supported by in-context learning for document labeling) and in terms of scope (we develop a method that is generalizable beyond ICCC conference proceedings and can be used to analyze any conference accessible via dBLP). We analyze the ICCC domain through both the temporal evolution of the core topics defined by the model and the qualitative analysis of the generated topics.

For further work, we propose to explore LLMs for document summarization, since transformer-based models that we use for document representation, have limited input size. We believe that exploring LLMs for both document summarization and large-scale qualitative analysis of topics at different time points can provide interesting insights into how to analyze a given scientific domain at scale. Next, we want to explore the interrelationships and evolution of collaboration networks within conference communities.

ACKNOWLEDGMENTS

The authors acknowledge financial support from the Slovenian Research and Innovation Agency through research core funding (No. P2-0103) and project No. PR-12394.

REFERENCES

Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Routledge.

Colton, S., & Wiggins, G. A. (2012). Computational creativity: The final frontier? In *Proceedings of the 20th european conference on artificial intelligence* (pp. 21–26). Amsterdam, The Netherlands, The Netherlands: IOS Press. https://doi.org/10.3233/978-1-61499-098-7-21 doi: 10.3233/978-1-61499-098-7-21

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Fortuna, B., Grobelnik, M., & Mladenić, D. (2007). Ontogen: Semi-automatic ontology editor. In *Human computer interface (part ii) (hci 2007), lncs 4558* (Vol. 4558, p. 309–-318).

Grootendorst, M. (2020). *Keybert: Minimal keyword extraction with bert.* Zenodo.

Grootendorst, M. (2022). *Bertopic: Neural topic modeling with a class-based tf-idf procedure.*

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220.

Koloski, B., Lavrač, N., Cestnik, B., Pollak, S., Škrlj, B., & Kastrin, A. (2024). AHAM: Adapt, Help, Ask, Model - Harvesting LLMs for literature mining. In *Proceedings of ida 2024.*

Koloski, B., Pollak, S., Škrlj, B., & Martinc, M. (2022, June). Out of thin air: Is zero-shot cross-lingual keyword detection better than unsupervised? In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 400–409). Marseille, France: European Language Resources Association.

Kozareva, Z., & Hovy, E. (2010). A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1110–1118).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Lampinen, A., Dasgupta, I., Chan, S., Mathewson, K., Tessler, M., Creswell, A., ... Hill, F. (2022, December). Can language models learn from explanations in context? In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the association for computational linguistics: Emnlp 2022* (pp. 537–563). Association for Computational Linguistics.

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., ... Xie, X. (2023). Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.

Loughran, R., & O'Neill, M. (2017). Application domains considered in computational creativity. In *Proceedings of iccc 2017.*

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, *2*(11), 205.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, *3*(29), 861.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., ... Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, *56*(2), 1–40.

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023, May). MTEB: Massive text embedding benchmark. In A. Vlachos & I. Augenstein (Eds.), *Proceedings of the 17th conference of the european chapter of the association for computational linguistics* (pp. 2014–2037). Dubrovnik, Croatia: Association for Computational Linguistics. https://aclanthology.org/2023.eacl-main.148 doi: 10.18653/v1/2023.eacl-main.148

Navigli, R., & Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, *193*, 217–250.

Pan, J., Gao, T., Chen, H., & Chen, D. (2023, July). What in-context learning "learns" in-context: Disentangling task recognition and task learning. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 8298–8319). Toronto, Canada: Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.527

Podpečan, V., Lavrač, N., Wiggins, G., & Pollak, S. (2018). Conceptualising computational creativity: Towards automated historiography of a research field. In C. L. François Pachet Anna Jordanous (Ed.), *Proceedings of iccc 2018* (pp. 288–295).

Pollak, S., Boshkoska, B., Miljkovic, D., Wiggins, G., & Lavrac, N. (2016). Computational creativity conceptualisation grounded on iccc papers. In F. Pachet, A. Cardoso, V. Corruble, & F. Ghedini (Eds.), *Proceedings of the international conference on computational creativity* (pp. 123–130). (International Conference on Computational Creativity ; Conference date: 25-06-2018 Through 29-06-2018)

Porturas, T., & Taylor, R. A. (2021). Forty years of emergency medicine research: Uncovering research themes and trends through topic modeling. *The American Journal of Emergency Medicine*, *45*, 213–220.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Blog*. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Reimers, N., & Gurevych, I. (2019, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing.* Association for Computational Linguistics.

Reimers, N., & Gurevych, I. (2020, 11). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 conference on empirical methods in natural language processing.* Association for Computational

Linguistics.

Škrlj, B., Koloski, B., & Pollak, S. (2022). Retrieval-efficiency trade-off of unsupervised keyword extraction. In P. Pascal & D. Ienco (Eds.), *Discovery science* (pp. 379–393). Cham: Springer Nature Switzerland.

Smith, B. (2003). Chapter 11: Ontology. In L. Floridi (Ed.), *Blackwell guide to the philosophy of computing and information* (Vol. 7250, p. 155—-166). Blackwell.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2).*

Touvron, H., & et al. (2023). *Llama 2: Open foundation and fine-tuned chat models.*

Wang, K., Reimers, N., & Gurevych, I. (2021, November). Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the association for computational linguistics: Emnlp 2021* (pp. 671–688). Punta Cana, Dominican Republic: Association for Computational Linguistics.

Wang, K., Thakur, N., Reimers, N., & Gurevych, I. (2022, July). GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 2345–2360). Seattle, United States: Association for Computational Linguistics.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Wei, Y., Huang, Q., Zhang, Y., & Kwok, J. (2023, December). KICGPT: Large language model with knowledge in context for knowledge graph completion. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 8667–8683). Singapore: Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.580 doi: 10.18653/v1/2023.findings-emnlp.580

Yao, L., Zhang, Y., Wei, B., Zhang, W., & Jin, Z. (2018). A topic modeling approach for traditional chinese medicine prescriptions. *IEEE Transactions on Knowledge and Data Engineering*, *30*(6), 1007–1021.

Ye, H., Zhang, N., Chen, H., & Chen, H. (2022, December). Generative knowledge graph construction: A review. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 1–17). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.1

GENERATIVNA UMETNA INTELIGENCA ZA KONCEPTUALIZACIJO RAČUNALNIŠKE KREATIVNOSTI

Tehnologija ustvarjanja konceptov se ukvarja z inženiringom programske opreme za (pol)avtomatizirano konceptualizacijo domene. Ta članek predstavlja pristop k avtomatizaciji konceptualizacije domene računalniške kreativnosti z izkoriščanjem generativne umetne inteligence (GAI) kot splošnega orodja za tehnologijo ustvarjanja konceptov. Pristop je prikazan na nalogi konceptualizacije domene (computational creativity, CC) z uporabo vseh javno dostopnih zbornikov od ICCC-2010 do ICCC-2023, pa tudi na nalogi avtomatiziranega strukturiranja kazala posameznih zbornikov Mednarodnih konferenc o računalniški kreativnosti. Implementirana metodologija GAI omogoča avtomatizirano konceptualizacijo katere koli domene, avtomatizirano strukturiranje kazala katerega koli zbornika ali dokumentnega korpusa ter ponovljivost eksperimentov in ponovno uporabo programske opreme.

Keywords: generativna AI, računalniška kreativnost, obdelava naravnega jezika

KORPUS CVET 1.0: IZDELAVA, OPIS IN ANALIZA ZBIRKE STAREJŠIH BESEDIL V VERSKI PERIODIKI

Diana KOŠIR,[1] Tomaž ERJAVEC[2]

[1]Inštitut za jezikoslovne študije ZRS Koper

[2]Odsek za tehnologije znanja, Institut "Jožef Stefan"

V prispevku je predstavljen proces izdelave in jezikoslovnega označevanja korpusa CVET 1.0, ki vsebuje besedila patra Hijacinta Repiča v starejšem slovenskem jeziku, objavljena v verskem glasilu *Cvetje z vertov sv. Frančiška* v obdobju 1881–1916. Besedila so bila v obliki PDF pridobljena s portala dLib, urejena v urejevalniku Word in nato pretvorjena v zapis TEI. Starejše besedje je bilo z odprtokodnim orodjem za normalizacijo avtomatsko posodobljeno, kar olajša iskanje po korpusu in nadaljnjo analizo gradiva. V članku so izpostavljene nekatere napake, ki so nastale pri posodabljanju in bodo v naslednji verziji korpusa ročno popravljene. Posodobljena besedila so bila nato še avtomatsko jezikoslovno označena z oblikoskladnjo in skladnjo po sistemu Universal Dependencies. Zapis TEI smo pretvorili v več izvedenih formatov in zbirko objavili pod odprto licenco na repozitoriju in konkordančnikih CLARIN.SI, ki so primerni za jezikoslovne analize gradiva. V drugem delu prispevka je prikazan primer analize avtorjevega pripovednega stila, opravljene s konkordančnikom noSketch Engine, ki temelji na frekvenčnih spremenljivkah najpogostejših in najmanj pogostih besed ter ključnih besed.

Ključne besede: starejša slovenščina, verski tisk, TEI, normalizacija, stilistična analiza, leksika

1 UVOD

Sredi 19. stoletja se je v slovenskem kulturnem prostoru postopoma vzpostavil periodični verski tisk z namenom poduka in verske vzgoje bralske publike. Izhajala so naslednja verska glasila: *Krščanski detoljub* (Ljubljana), *Angeljček: otrokom učitelj in prijatelj* (Ljubljana), *Družinski prijatelj, poučno-zabavni list s podobami za slovenske družine* (Trst), *Drobtinice* (Gradec, Maribor, Ljubljana) in *Cvetje z vertov sv. Frančiška* (Gorica, Kamnik).

To je bilo tudi obdobje narodne prebuje s čitalniškim in taborskim gibanjem, prosvetno-kulturnimi in gospodarskimi društvenimi organizacijami, zato je (predvsem) na obrobjih slovenskega kulturnega prostora slovenski tisk zaradi jezikovnega elementa imel tudi identifikacijsko in narodnopovezovalno vlogo (Perenič, 2012; Košir, 2022). Jezik lahko razumemo kot enega od ključnih gradnikov tako osebne kot nacionalne identitete, preko posameznikovega odnosa do (maternega) jezika pa se kaže njegova narodna zavest (prim. Nećak-Lük et al., 1998, str. 77; Mikolič, 2000, str. 180–182).

Podobnega formata kot Slomškove *Drobtinice* in z naklado nekaj tisoč izvodov je leta 1880 na Goriškem začelo izhajati *Cvetje z vertov sv. Frančiška*. Tudi vsebinsko sta bili publikaciji sorodni – *Drobtinice* so vsebovale nabožna, življenjepisna, leposlovna, vzgojno-izobraževalna in občasno tudi poljudnostrokovna besedila (Ulčnik, 2010, str. 690), dočim je bilo v *Cvetju* leposlovja izrazito manj, poudarek je bil na življenjepisih svetnikov in nabožnih besedilih, na zgodovini reda in misijonih, platnice pa so bile posvečene jezikoslovju. Oba urednika, Slomšek in Škrabec, sta gojila poseben odnos do maternega jezika in pomembno prispevala k razvoju poenotene knjižne norme v 19. stoletju, kar se je podobno odražalo v njunih tiskovinah. Slomšek je bil jezikovno sicer bolj odprt in strpnejši pri počasi vpeljujoči se rabi l. 1851 sprejetih Svetčevih »novih oblik«, pa vendar je l. 1852 uredništvo v *Drobtinicah* objavilo, da bodo v primeru jezikovne neustreznosti prispevke popravili in dodelali; uredniška politika je kljub sprejemanju leksikalne različnosti kazala na težnjo po jezikovni enotnosti almanaha (Ulčnik, 2010, str. 687–688, 695–696). Tudi jezikoslovec Škrabec si je (nemara celo strožje) prizadeval za jezikovno izpiljenost besedil in poenoteno leksikalno rabo, »[z]ato 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« (Kunstelj v Korošak, 2001, str. 19). Za razliko od Slomškovega tiska je notranjost Škrabčevega *Cvetja*, z izjemo znamenitih jezikoslovnih razprav s platnic, še neraziskana.

Poleg urednika so besedila za revijo prispevali različni frančiškanski sobratje, navadno podpisani z inicialkami – tudi p. Hijacint (Anton) Repič (P. H. R.) iz koprskega samostana sv. Ane.

Projekt izdelave korpusa CVET 1.0 (Košir in Erjavec, 2024) z opusom objav patra Hijacinta Repiča v *Cvetju* je del soavtoričine doktorske raziskave, ki odpira vprašanje prisotnosti slovenstva v slovenskem delu Istre pred 1. sv. vojno ter vloge, ki jo je imel samostan sv. Ane za jezikovno in kulturno utrjevanje Slovencev na Koprskem, s tem pa prinaša nova spoznanja na področju zgodovinske sociolingvistike, literarne vede, literarne pragmatike, uporabnega jezikoslovja in narodnega vprašanja. Metodološko raziskava sega še na področje digitalne humanistike z metodami korpusnega jezikoslovja, korpusne stilistike in računalniške stilometrije.

V nadaljevanju bo najprej predstavljen postopek priprave in označevanja korpusa CVET 1.0, nato pa bo prikazan praktični primer uporabe konkordančnikov, ki so dostopni v okviru slovenske raziskovalne infrastrukture CLARIN.SI za stilistično korpusno analizo Repičevega pripovednega sloga. Osebni slog odseva avtorjevo osebnost oz. jezikovne navade (Leech in Short, 2007, str. 23), tudi preference, saj gre za subjektivno motivirane posamične izbire iz celotnega jezikovnega repertoarja, ki se zgodijo glede na piščev namen, *kako* želi predati določeno sporočilo bralcu (Biber in Conrad, 2009, str. 144).

2 PRIPRAVA IN OBJAVA GRADIVA

Verska revija *Cvetje z vertov sv. Frančiška* je izhajala med letoma 1880 in 1944. Od začetka do leta 1915 jo je urejal eden od njenih pobudnikov p. Škrabec, nasledil ga je p. Evstahij Brlec. Goriška knjižnica Franceta Bevka je v sodelovanju s Škrabčevo knjižnico Frančiškanskega samostana Kostanjevica v Novi Gorici v letih 2017 in 2018 izpeljala projekt digitalizacije revije, ki je v formatih PDF in TXT dostopna na portalu Digitalna knjižnica Slovenije (dLib.si).

Prvi pregled Repičevih člankov v *Cvetju z vertov sv. Frančiška* je pripravil p. Bruno Korošak (2006). Celoten fond do l. 1918 (Repičeva smrt) je bil ponovno pregledan in dopolnjen. Ugotovljeno je bilo, da je p. Repič za revijo med letoma 1881 in 1916 prispeval 140 člankov.

2.1 Priprava gradiva in metapodatkov

V prvi fazi priprave gradiva smo iz dLib izbrali posamezne številke revije, v katerih se pojavljajo članki p. Repiča, in te članke iz PDF prekopirali v urejevalnik Word. Ker so nekateri članki izšli po delih v več številkah revije, je skupno število (delnih) člankov, in s tem tudi datotek Word, 230.

V urejevalniku Word smo nato članke oblikovno poenotili in ročno popravili. Ti popravki so zajeli poenotenje zapisa ločil glede na sedanjo pravopisno normo (levostičnost končnih in nekaterih nekončnih ločil; prost zapis večpičja v tropičje; poenotena oblika narekovajev) in odpravljanje napak, ki so nastale pri OCR (tj. pretvorbi faksimila v besedilo) v datotekah PDF. Izvirno smo ohranili nekatere posebnosti stave ločil (npr. zapis zadnjega narekovaja pred/za končnim ločilom), zapis naglasnih znamenj (npr. pomensko razlikovalna raba krajše oblike povratnega osebnega zaimka "se" in predloga "sè") in delitev na odstavke.

Vzporedno z urejanjem datotek v programu Word smo za vsako besedilo vnesli njegove metapodatke v razpredelnico Excel, in sicer: ime datoteke Word (ki je obenem identifikator besedila), avtor (vedno sicer p. Repič, vendar mestoma tudi z zaznamkom, po katerem avtorju je delo povzeto oz. prevod dela katerega avtorja je), naslov članka, mesto objave (leto, letnik, številka), URL izvornega mesta objave v dLib in na katerih straneh se članek pojavi.

2.2 Zapis TEI

V naslednjem koraku smo gradivo iz formatov Word in Excel pretvorili v zapis, ki je bolj primeren za hrambo kot tudi za nadaljnje pretvorbe v formate, namenjene posameznim orodjem, in sicer v XML shemo skladno s priporočili iniciative za kodiranje besedil TEI (TEI Consortium, 2020), bolj natančno, v shemo, ki jo raziskovalna infrastruktura CLARIN.SI priporoča za korpuse, deponirane v repozitoriju infrastrukture.[1]

Datoteke Word smo najprej pretvorili v osnovni TEI, in sicer z uporabo standardnih skript XSLT za pretvorbo v in iz dokumentov TEI,[2] medtem ko smo razpredelnico Excel z metapodatki o posameznih besedilih shranili kot datoteko TSV, ki je primerna za nadaljnje avtomatske obdelave. Nato smo z namensko skripto združili vsako besedilno datoteko TEI z njenimi

[1]https://github.com/clarinsi/TEI-schema
[2]https://github.com/TEIC/Stylesheets

metapodatki in jo formirali 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/11356/1226</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>
```

Celoten korpus je formiran kot en dokument XML s krovno datoteko (element <teiCorpus>), ki vsebuje kolofon korpusa in povezave (elementi XInclude) na besedilne datoteke korpusa. Kolofon korpusa poda korpusne metapodatke, tj. naslov, odgovorne osebe, velikost, založnika, distributerja, licenco, skupni opis vira in projekta ter seznam uporabljenih jezikov (slovenščina, za metapodatke tudi angleščina).

2.3 Posodabljanje besed

Korpus vsebuje dela, ki so nastala proti koncu devetnajstega in v začetku dvajsetega stoletja, ko so se marsikatere besede pisale drugače, kot pa je to v sodobni normi. Arhaične oblike besed otežijo iskanje po korpusu, odvisno od raziskave pa tudi njihovo nadaljnjo analizo. Poleg tega bi jezikoslovno označevanje zbirke izvornih besedil imelo dosti napak, saj orodja, ki so na voljo za tako označevanje, delujejo dobro le na sodobni standardni slovenščini.

Zaradi teh razlogov smo besede v korpusu najprej avtomatsko posodobili, za kar smo uporabili odprtokodno orodje za normalizacijo cSMTiser[3] (Scherrer in Ljubešić, 2016), ki temelji na principu statističnega strojnega prevajanja in orodju Moses (Koehn, 2010). cSMTiser smo naučili posodabljanja na gajičnem delu ročno posodobljenega korpusa slovenščine goo300k (Erjavec, 2016), podobno, kot smo že pred tem naredili za posodabljanje zbirke slovenskih romanov v okviru korpusa ELTeC (Schöch et al., 2021), za korpus starejše slovenske proze PriLit (Žejn in Erjavec, 2021) ali za zbirko slovenskih pregovorov (Babič in Erjavec, 2022). S tako izšolanim orodjem smo nato posodobili besede vseh besedil v korpusu. Pri tem velja opomba, da izvorne besedne oblike s tem niso izgubljene, pač pa so posodobljene oblike, kjer se razlikujejo od izvirnih, pripisane le-tem.

V postopku avtomatskega posodabljanja starejšega jezika so bile denimo ustrezno posodobljene besede, ki imajo polglasnik zapisan z »e« (npr. *miloserčnost* > *milosrčnost, serce* > *srce*, *smert* > *smrt*), končnica -vec > -lec (*bravec* > *bralec*), premena -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*), obliko *aposteljnov* > *apostolov*, in nekatere funkcijske besede: zaimek *gdo* > *kdo*, veznik *temuč* > *temveč*, veznik in členek *toraj*/*torej* > *torej*, členek *vže* > *že*, predlog *sè* > *s*, *ž* > *z*. Nekatere starejše oblike besed so ostale neposodobljene (npr. predlog *mej* > med, *blager* > blagor). Po drugi strani pa dela avtomatsko posodabljanje tudi napake. Določene besede so bile posodobljene v sicer obstoječo, a neustrezno besedo (npr. *čiger* > *čigra* nam. *čigar*), po nepotrebnem posodobljene v napačen zapis (npr. *vbogajme* > *ubogajme*) oz. nove skovanke (npr. *keterim* > *ketim* nam. *katerim*). Spodnja tabela 1 prikazuje nekatere tovrstne napake, ki smo jih identificirali v korpusu, seveda pa je takih napak po vsej verjetnosti še več.

[3]https://github.com/clarinsi/csmtiser

Tabela 1. Nekatere prezrte starejše oblike in napake, ki so nastale pri avtomatskem posodabljanju leksike.

Izvorna beseda	*Posodobljena beseda*	*Korigirana posodobljena beseda*
čiger	čiger (rod. mn. od 'čigra')	čigar
sobrat	zbrat	sobrat
keterim	ketim	katerim
marsiketerim	marsiketim	marsikaterim
neketerim	neketim	nekaterim
kesneje	kosno	kasneje
mariveč	mareveč	marveč
gvardijan	gvardian	gvardijan
vmerje	umerje	umrje
(v) tempeljnu	(v) tempelju	(v) templju
(pred) vratmi	(pred) vratami	(pred) vrati
vbogajme	ubogajme	vbogajme
tolikanj	tolikanj	toliko / tako
(s) tebo	(s) tebo	(s) tabo / teboj
obena	obena	nobena
češer	češer	česar
mej	mej	med
ke	ke	ko
jeli	jeli	jeli = začeli; je li = ali je
namestu	namestu	namesto
ničemernost	ničemernost	nečimrnost
apostelj	apostelj	apostol
kolikanj	kolikanj	koliko
odpuščenje	odpuščenje	odpuščanje
razodevenje	razodevenje	razodetje
vsegamogočni	vsegamogočni	vsemogočni
ražalil	ražalil	razžalil
marcija	marcija	marca
preskerbel	preskrbel	preskrbel/priskrbel/poskrbel (različen pomen)
verni	vrni	verni/vrni (različen pomen)

2.4 Jezikoslovno označevanje

Na osnovi avtomatsko posodobljenih besed smo nato korpus jezikoslovno označili. Tu smo uporabili odprtokodno orodje CLASSLA[4] (Ljubešić in Dobrovoljc, 2019), s katerim smo dodali naslednje jezikoslovne oznake v besedilo, npr. za besedno obliko “*bi*”:

- lemo oz. osnovno obliko besede, tu “biti”;
- oblikoskladenjsko oznako po priporočilih MULTEXT-East (Erjavec, 2012), »Va-c«, ki se razveže v oblikoskladenjske lastnosti “Verb Type=auxiliary VForm=conditional”, pri čemer obstaja tudi ekvivalentna oznaka v slovenščini, tu »Gp-g« in njena razvezava v “glagol vrsta=pomožni oblika=pogojnik”;
- oblikoskladenjske lastnosti po sistemu Universal Dependencies za slovenski jezik (Dobrovoljc et al., 2017), tu “UPosTag=AUX Mood=Cnd VerbForm=Fin”. Te oznake so sicer podobne oznakam MULTEXT-East, vendar z drugače definiranim naborom lastnosti in vrednosti, občasno se pa od njih tudi sistemsko razlikujejo, kot je tudi primer za “bi”;
- odvisnostna skladenjska razčlenitev povedi po sistemu Universal Dependencies za slovenski jezik.

Posodobljena in jezikoslovno označena različica korpusa je bila formirana kot svoj korpus; format označenega besedila je ilustriran v sliki 2.

[4]https://github.com/clarinsi/classla

Slika 2. Primer zapisa TEI označene povedi v korpusu.

```
<s xml:id="CVET-1887_7_2_43-46.1.s1">
   <w xml:id="CVET-1887_7_2_43-46.1.s1.t1" ana="mte:Ncmpn"
      msd="UPosTag=NOUN|Case=Nom|Gender=Masc|Number=Plur" lemma="vzrok">Vzroki</w>
   <w xml:id="CVET-1887_7_2_43-46.1.s1.t2" ana="mte:Cc" msd="UPosTag=CCONJ" lemma="in"
      >in</w>
   <w xml:id="CVET-1887_7_2_43-46.1.s1.t3" ana="mte:Ncfpn"
      msd="UPosTag=NOUN|Case=Nom|Gender=Fem|Number=Plur" lemma="korist">koristi</w>
   <w xml:id="CVET-1887_7_2_43-46.1.s1.t4" ana="mte:Ncnsg"
      msd="UPosTag=NOUN|Case=Gen|Gender=Neut|Number=Sing" norm="trpljenja"
      lemma="trpljenje">terpljenja</w>
   <w xml:id="CVET-1887_7_2_43-46.1.s1.t5" ana="mte:Agpfpg"
      msd="UPosTag=ADJ|Case=Gen|Degree=Pos|Gender=Fem|Number=Plur" lemma="pobožen"
      >pobožnih</w>
   <w join="right" xml:id="CVET-1887_7_2_43-46.1.s1.t6" ana="mte:Ncfpg"
      msd="UPosTag=NOUN|Case=Gen|Gender=Fem|Number=Plur" lemma="duša">duš</w>
   <pc xml:id="CVET-1887_7_2_43-46.1.s1.t7" ana="mte:Z" msd="UPosTag=PUNCT">.</pc>
   <linkGrp corresp="#CVET-1887_7_2_43-46.1.s1" targFunc="head argument" type="UD-SYN">
      <link ana="ud-syn:root"
         target="#CVET-1887_7_2_43-46.1.s1 #CVET-1887_7_2_43-46.1.s1.t1"/>
      <link ana="ud-syn:cc"
         target="#CVET-1887_7_2_43-46.1.s1.t3 #CVET-1887_7_2_43-46.1.s1.t2"/>
      <link ana="ud-syn:conj"
         target="#CVET-1887_7_2_43-46.1.s1.t1 #CVET-1887_7_2_43-46.1.s1.t3"/>
      <link ana="ud-syn:nmod"
         target="#CVET-1887_7_2_43-46.1.s1.t3 #CVET-1887_7_2_43-46.1.s1.t4"/>
```

Kot je razvidno iz primera, so povedi označene z elementom <s>, besede z <w> ter ločila s <pc>. Iztočnična oblika besede je podana kot vrednost atributa @lemma, oznake MULTEXT-East kot vrednosti atributa @ana, lastnosti Universal Dependencies kot @msd, posodobljena oblika besede pa kot vrednost @norm. Skladenjska analiza je podana v elementu <linkGrp>, in sicer za vsako skladenjsko odvisnost (element <link>) kot povezava med dvema pojavnicama (prek sklica na njuna identifikatorja v atributu @target), pri čemer je oznaka skladenjske odvisnosti podana kot vrednost atributa @ana.

V različici korpusa, ki vsebuje posodobljena in jezikovno označena besedila, je dopolnjen tudi kolofon s taksonomijo skladenjskih oznak Universal Dependencies in z opisom uporabljenih orodij.

2.5 Objava in velikost korpusa

Različico 1.0 izdelanega korpusa, poimenovanega "CVET", smo objavili v repozitoriju CLARIN.SI (Košir in Erjavec, 2024) pod odprto licenco Creative

Commons, priznanje avtorstva (CC BY).

Korpus je za prevzem na voljo v štirih stisnjenih datotekah. Vsaka vsebuje direktorij, v njem pa datoteke za eno od variant korpusa:

- Cvet.TEI: korpus v zapisu TEI, kot je predstavljen v razdelku 2.2. Poleg korenske XML datoteke korpusa vsebuje še 230 XML besedil z metapodatki in direktorij s shemo CLARIN.SI TEI;
- Cvet.TEI.ana: korpus z jezikoslovno označenimi besedili;
- Cvet.txt: korpus brez oznak (navadno besedilo), avtomatsko pretvorjen iz TEI.ana, in razpredelnica z metapodatki posameznih besedil; format je primeren za neposreden uvoz v razna orodja. Za namene raznovrstnih raziskav je ta varianta na voljo v kombinacijah sledečih različic:
    - izvorno besedilo (npr.: “*Ti pa nikaker ne poslušaj hudobe,*”); posodobljeno besedilo (“Ti pa nikakor ne poslušaj hudobe,”); lematizirano besedilo (“ti pa nikakor ne poslušati hudoba,”);
    - besedilo z izvorno kapitalizacijo (kot zgoraj);
    - besedilo v malih črkah (“ti pa nikaker ne poslušaj hudobe,”);
    - besedilo z izvorno stičnostjo in besedilo razdeljeno po pojavnicah (“Ti pa nikaker ne poslušaj hudobe ,”);
- korpus v t. i. vertikalnem formatu, ki je primeren za uvoz v spletne konkordačnike; dodana je tudi konfiguracijska datoteka korpusa, kot se uporablja v konkordančnikih CLARIN.SI.

Repozitorijski vnos je povezan s konkordančniki CLARIN.SI, in sicer noSketch Engine in KonText, kar omogoča poizvedbe in analize korpusa brez znanja programiranja in z različnimi vizualizacijami in možnostjo shranjevanja rezultatov poizvedb.

Korpus vsebuje 230 besedil oz. 3.228 odstavkov in 10.109 (avtomatsko določenih) povedi. Pojavnic je skupno 212.703, od tega 176.700 besed (če štejemo vse pojavnice, ki niso označene kot ločilo) oz. 175.907 (kot besede prešteje konkordančnik, ki npr. cifer ne šteje kot besede). Avtomatsko je bilo posodobljenih 13.033 oz. 7,4 % besed.

3 STILISTIČNA ANALIZA GRADIVA

Stilistika je opredeljena kot jezikoslovno preučevanje sloga oz. namenske specifične rabe jezika, ki nakazuje odnos med kreativnim dosežkom (učinkom) in jezikovno manifestacijo (Leech in Short, 2007, str. 11, 55), na kar lahko vplivajo nejezikovne spremenljivke, kot so žanr, avtor, zgodovinsko obdobje ipd. (Jeffries in McIntyre, 2010, str. 1). Korpusna stilistika je razumljena kot aplikacija teorij, modelov in metod stilistike v analizi korpusa, pri čemer je poudarek na razločevanju vzorcev v jezikovni rabi s preučevanjem velikih količin jezikovnih podatkov (McIntyre, 2015, str. 61).

Pri stilistični analizi opazujemo sestavine sloga (angl. *features of style*, Leech in Short, 2007), ki so jezikovni elementi, ki v danem korpusu izstopajo – glede na ozadje pričakovanj bralca pojavnost nekega jezikovnega elementa povzroči presenečenje (notranji odklon, angl. *internal deviation*), in pomembno prispevajo k oblikovanju sloga. Te jezikovne entitete imenujemo slogovni označevalci (angl. *style markers*), ki so v določenih kontekstih najbolj ali najmanj frekvenčni (Enkvist, 1964, str. 34 v Leech in Short, 2007, str. 59). Te sestavine je pred slogovno analizo treba izbrati, kriterije za kvantitativno analizo pa po Leechu in Shortu določa preplet treh konceptov: devianca (angl. *deviance*), prominenca (angl. *prominence*) in književna relevanca (angl. *literary relevance*). Devianca je statistični pojem, povezan s kvantifikacijo, ki kaže na odklon frekvence rabe določenega jezikovnega (ali slogovnega) elementa od normalne frekvence; prominenca je psihološki pojem, intuitivni ekvivalent devianci, in je odvisen od bralčevega odziva na besedilo, ta pa je pogojen z vrsto dejavnikov (občutek za slog, bralne izkušnje, razpoloženje, koncentracija itd.), zato se zdi prominenca bolj subjektivna, čeravno je tudi tu možno izmeriti frekvenco. Književna relevanca je opredeljena kot »umetniško motiviran odmik od relativne norme« (angl. *foregrounding*), ki je kvalitativen (kreativna raba jezika v nasprotju s konvencionalno rabo) ali kvantitativen (devianca vsled nepričakovane frekvenčnosti) (Onič, 2014, str. 184). Pri naši analizi bomo opazovali posamezne jezikovne prvine, ki jih Leech in Short (2007, str. 61–63) uvrščata v skupini leksikalnih in slovničnih kategorij.

V digitalni humanistiki se je kot kvantitativna in statistična analiza avtorjevega stila uveljavila stilometrija, primerjalna metoda, ki »meri« podobnosti in razlike med besedili na različnih jezikovnih ravninah (Žejn, 2020). Pri stilističnih in stilometričnih analizah gre najpogosteje za analize na leksikalni ravni, pri čemer so med najpogostejšimi preučevanimi kategorijami (spremenljivkami) naslednje: najpogostejših 100 besed, distribucija pogostosti besedišča, raznolikost (gostota) besedišča in *hapax legomena* oz. *hapax dislegomena* (izrazi, ki se v danem kontekstu pojavijo samo enkrat oz. dvakrat), povprečna dolžina besed oz. povedi, besedni in črkovni n-grami, pogostost najpogostejših besednih kolokacij (zaporedje dveh, treh, štirih itd. besed), najpogostejših funkcijskih besed (zaimki, pomožni glagoli, predlogi, vezniki, členki, medmeti), ki same navadno nimajo leksikalnega pomena, izražajo pa slovnični odnos z drugimi besedami v stavku oz. odnos ali razpoloženje govorca; ritmično zaporedje naglašenih in nenaglašenih zlogov, razporeditev ločil idr. (Mosteller and Wallace, 1964; Hoover, 2002, 2003; Grieve, 2007; Eder, 2011; Žejn, 2020 idr.).

Za našo kvalitativno in kvantitativno analizo smo izbrali tri kategorije (spremenljivke), povezane s frekvenčnostjo: najpogostejših 100 besed, najmanj pogoste besede in ključne besede (v vseh treh primerih bodo iskane leme, v oklepaju bodo z oznako izv. pripisane izvorne oblike besed, če se te razlikujejo od lem). Orodje noSketchEngine omogoča funkcijo iskanja ključnih besed (*keywords*), ki po frekvenci rabe odstopajo navzgor ali navzdol glede na referenčni korpus, seznam pa zajame polnopomenske in nepolnopomenske (funkcijske) besede. Primerjali bomo seznam ključnih besed glede na izbrana referenčna korpusa starejšega gradiva (IMP 1.1, Erjavec, 2014; PriLit 1.0, Žejn in Erjavec, 2021). Pogostost rabe poda vpogled v gostoto besedišča in vsebinsko označi kontekst(e) besedil v korpusni zbirki, ključne besede pa razodevajo avtorjevo pomensko polje in skozi opazovanje njihovih pomenov in kontekstov rabe razkrivajo piščev spoznavni in literarni svet (prim. Mikolič, 2020, 2022). Pri interpretaciji dobljenih kvantitativnih podatkov je treba upoštevati tudi subjektivni interpretativni okvir raziskovalca.

3.1 Analiza avtorjevega pripovednega sloga z uporabo konkordančnikov

Med najpogostejšimi 100 besedami (lemami) v korpusu CVET 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 pričakovani glede na vsakdanje sporazumevalne okoliščine, pogosto pojavijo glagoli rekanja (*reči*, *prositi*, *govoriti*, nadalje tudi *praviti*, *odgovoriti*, *povedati*) in modalni glagoli (prvo število so pojavitve, drugo pa % besedil v korpusu, kjer se lema pojavi): *moči* (597 = 75 %), *hoteti* (577 = 70,87 %) in *morati* (465 = 66,52 %). Očitna prisotnost naklonskih glagolov zasluži natančnejšo analizo kontekstov rabe. Primeri, kot so *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*, sodijo v žanrski okvir nabožnega vzgojnopoučnega članka. Analizo bi na tem mestu lahko nadgradili in poglobili z opazovanjem a) glagolskega naklona (razmerje med rabo povednega, velelnega in pogojnega naklona), kar bi denimo služilo opazovanju razmerja med prevladujočo sporočevalno oz. vplivanjsko vlogo besedil, in b) glagolske osebe in števila (1. os. mn., 2. os. ed. ali mn., 3. os. ed. ali mn.), pri čemer bi opazovali frekvenco posrednega in neposrednega nagovarjanja naslovnika (bralca).

Religiozni diskurz opredeljuje moralno-vrednostno razmerje dobro : slabo in med najpogostejšimi besedami 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 % vseh besedil, kar pomeni, da je njihova zastopanost razpršena skozi celoten korpus.

Po pogostosti izstopa raba zaimka *kateri*, ki se pojavi kar 1.822-krat in v 90,87 % besedil. Izrazita je vezniška raba, kjer uvaja podredje z oziralnim odvisnikom, v današnji knjižni slovenščini bi ga nadomestil zaimek *ki* (*Ali ti, moj Jezus, keteremu je vse odkrito, in keteri si vstvaril vse v meri in številu ...*; *On ne ve, da so britkosti božje šibe polne ljubezni, da so mile kazni, pripravljene edino mojim izvoljenim, ketere čistim v ognju britkosti, kaker se čisti v goreči peči zlato*). Tudi sicer lahko ob hitrem pregledu gradiva ugotovimo, da so v Repičevih tekstih pogoste dolge povedi s podredji, kar bi bilo vredno natančnejše statistične analize.

Med načinovnimi prislovi po frekvenčnosti izstopa *jako* (342-krat; ob 11 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 nam jih posreduje korpus, lahko ugotovimo, da se starejša oblika 'jako' pojavlja skozi celotno gradivo (v 57,83 % besedil), novejša oblika 'zelo' pa ob njej nekoliko bolj konsistentno po letu 1913. Ker se prislov *jako*, kot bomo videli v nadaljevanju, pojavi tudi med ključnimi besedami glede na referenčni korpus starejših besedil, ga lahko označimo kot značilno potezo patrovega pripovednega sloga.

Med najmanj pogostimi besedami (ena do dve pojavitvi, korpus takih lem navaja 4.656) 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 oz. nepremišljeno izreči, zmoliti); deležnik na -č: *kretajoč (se)*, *pazeč*, deležnik na -ši: *znebivši*, *zbudivši*; členek *znabiti*, besedna zveza *čudapoln* mir, *prvak* apostolov, *zvezdoznanka*, *zgoja* (ob *vzgoja*), *zgojitelj*, *razjasnjenje*, *život* (ob *telo*), *nesposobnost*, *neizmernost*, *gotovost*, *enakoličnost* itd. Nepričakovane oz. 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 oz. nenaklonjenost.

Ključne besede v fokusnem korpusu CVET so bile izluščene skozi primerjavo z izbranim korpusom starejše pripovedne proze PriLit (Žejn in Erjavec, 2021). Med 100 ključnimi besedami se za korpus CVET pojavijo naslednje leme, ki smo jih zaradi večje preglednosti razdelili na a) leksikalne in b) funkcijske besede:
a) samostalniki: *Frančišek*, *Monald* (izv. tudi *Monaljd*), zemljepisna imena *Asiz* (= Assisi), *Koper, Padova*; *redovnik, frančiškan, sobrat, tretjerednik, zavetnica, voditelj, častilec* (izv. *častivec*), *posvetnjak*, *vodilo*, *miloščina*, *milosrčnost* (izv. *miloserčnost*), *sočutje, blager* (= blagor; 3-krat 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*; pridevniki: *redoven, blažen, brezmadežen, blagoslovljen, presladek, izreden, nepopisljiv*, (Marija) *Porcijunkuljska* (= nanaša se na cerkev Marije Angelske pri Assisiju, t. i. Porcijunkula), *vsakovrsten* (izv. *vsakoversten*), *brezštevilen*; glagoli: *pridigati, oznanjevati, občevati* (= biti v stiku, sporazumevati se), *spolnjevati* (redko *izpolnjevati*), *spodbujati, rabiti, zanemarjati, zoperstavljati se, vničiti*; prislovi: *jako, kesneje** (= kasneje), *naposled*;

b) vezniki: *koliker* (= kolikor), *čiger** (= čigar),[5] *keteri** (= kateri), *marivеč** (= marveč); členki: *nikaker* (= nikakor), *seveda* (pisano tudi *se ve da*), *potemtakem, blager* (26-krat členkovna raba).

Ugotovimo lahko, da leksikalne besede nedvomno pripadajo religijskemu (krščanskemu) diskurzu, nekatere med njimi še podrobneje usmerijo na red sv. Frančiška Asiškega (npr. *Frančišek, Asiz, frančiškan, vodilo,*

[5]Z znakom (*) so označene starejše oblike besed, pri katerih je v postopku avtomatskega posodabljanja prišlo do napak. Te so bile pojasnjene v poglavju 2.3.

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 (liturgija, pastorala), 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 (Mikolič, 2020). V smer krepitve argumenta vodi tudi pogosta raba prislova *jako* (redko *zelo*) in pomensko nasprotna si členka *nikaker* in *seveda*.

Seznam prvih 100 ključnih besed z referenčnim korpusom IMP (Erjavec, 2014) poda zelo podobne rezultate. Različno se denimo pojavijo še leme bogo* (za *Bog*), *tretjerednik* in *tretjerednica*, gvardian* (za *gvardijan*), *habit*, *častivec*, *ljubivec*, veznik *ke* (tudi *kè*, v pomenu ‘ko’ oz. ‘če’ v pogojnih odvisnikih, tudi za izražanje želje).

4 SKLEP

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 oz. opombah). Prikazana korpusna analiza s konkordančniki CLARIN.SI je bila usmerjena v raziskovanje avtorjevega pripovednega stila. Uporabljene so bile funkcije, ki jih ponuja prosto dostopni konkordančnik noSketchEngine: seznam najpogostejših in najmanj pogostih besed (lem) ter seznam ključnih besed (lem) glede na izbrani referenčni korpus. Analiza frekvenčnosti besedja je podala nekatere pričakovane rezultate, kot so glagoli vezani na vsakdanje praktično-sporazumevalne okoliščine, pri tem pa so po pogostnosti izstopali modalni glagoli. Žanrsko besedila v glavnem sodijo med nabožne vzgojno-poučne članke, pri čemer bi bilo z vidika osebnega stila nadalje zanimivo raziskati, kakšno je razmerje med sporočanjsko in vplivanjsko vlogo besedil v odnosu do vsebine (vzgojno-moralne problematike), z drugimi besedami, pri katerih temah je diskurz pripovedovalca odločnejši, strožji, koliko odkrito apologizira, po drugi strani pa je v nekaterih pogledih v primerjavi z drugimi verskimi pripovedniki nemara tudi milejši (znano je, da so se frančiškani, tedaj najštevilčnejši red na Slovenskem, uprli nekaterim rigoroznim verskim praksam med kleriki (Čebulj, 1922)), ter kakšen učinek takšen diskurz doseže pri bralcu.

Analiza najpogostejših samostalnikov in pridevnikov je smiselna, saj skoznjo vstopimo v avtorjevo pomensko polje in spoznamo vsebinski okvir besedil, z nadaljnjo raziskavo kontekstov rabe posamezne leksike v korpusnem gradivu pa dostopamo do vseh pojavitev v gradivu in lažje izluščimo specifike: dobesedni oz. preneseni pomen, ustaljene oz. izvirne metafore, prispodobe idr.

Opazovali smo tudi najpogostejše prislove, zaimke in členke, ki so prav tako lahko stilni označevalci, ki zaznamujejo tipično skladnjo (dolge povedi, veliko podredij), ali z vidika prepričljivosti argumentacije krepijo oz. šibijo pomen sporočila (Mikolič, 2020). Z vidika jezikovne pragmatike bi bila zanimiva tudi analiza diskurznih označevalcev v navezavi na žanr besedil: nekatera po vsebinski zgradbi in zaradi neposrednega nagovarjanja naslovnika spominjajo na pridižna besedila, prisotnost homiletičnih pripovednih prvin pa bi lahko ugotavljali tudi z analizo diskurznih označevalcev, značilnejših za govorjeni diskurz.

Korpus CVET 1.0 je prva tovrstna 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). Korpusno gradivo, označeno z metapodatki o avtorstvu in objavi (leto, letnik, zvezek), zato lahko služi za komparativne sinhrono-diahrone raziskave normiranja slovenskega jezika na prelomu 19. in 20. stoletja v periodičnem tisku ter za raziskave udejanjanja knjižne norme, ki jo je urednik Škrabec podrobno predstavil skozi svoje jezikoslovne znanstvene razprave na platnicah, v sami vsebini *Cvetja*.

V naslednji verziji korpusa bomo odpravili neposodobljene in napačno posodobljene besede, in sicer s temeljitim pregledom napak, ki mu bodo sledile ročne korekture. Takšen korpus bo ne samo omogočil boljše analize, pač pa bi lahko služil tudi kot dodatna učna množica za cSMTiser, s čimer bi odpravili napake, ki se trenutno pojavljajo pri posodabljanju.

LITERATURA

Babič, S. in Erjavec, T. (2022). Izdelava in analiza digitalizirane zbirke paremioloških enot. *Conference on Language Technologies & Digital Humanities Ljubljana*, 2022. Pridobljeno 10. maja 2024, https://nl.ijs.si/jtdh22/pdf/JTDH2022_Babic_Erjavec_Izdelava-in-analiza-digitalizirane-zbirke-paremioloskih-enot.pdf

Biber, D. in Conrad, S. (1998). *Register, Genre, and Style*. Cambridge University Press.

Čebulj, R. (1922). *Janzenizem na Slovenskem in frančiškani: inavguralna disertacija*. Frančiškanska provincija Slovenije.

Dobrovoljc, K. et al. (2017). The Universal Dependencies Treebank for Slovenian. *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics, str. 33–38, doi:10.18653/v1/W17-1406.

Eder, M. (2011). Style-Markers in Authorship Attribution. A Cross-Language Study of the Authorial Fingerprint. *Studies in Polish Linguistics, 6*(1), 99–114. Pridobljeno 15. maja 2024, https://ejournals.eu/en/journal/studies-in-polish-linguistics/article/style-markers-in-authorship-attribution-a-cross-language-study-of-the-authorial-fingerprint

Erjavec, T. (2014). *Digital library and corpus of historical Slovene IMP 1.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1031

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing, 22*(3), 251–270. https://doi.org/10.1093/llc/fqm020

Hoover, D. L. (2002). Frequent Word Sequences and Statistical Stylistics. *Literary and Linguistic Computing 17*(2), 157–180. https://doi.org/10.1093/llc/17.2.157

Hoover, D. L. (2003). Multivariate Analysis and the Study of Style Variation. *Literary and Linguistic Computing, 18*(4), 341–360. https://doi.org/10.1093/llc/18.4.341

Jeffries, L. in McIntyre, D. (2010). *Stylistics*. Cambridge University Press.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Korošak, B. (2006). *Anton Hijacint Repič (1863-1918), frančiškan pri sveti Ani v Kopru*. Branko.

Košir, D. (2022). Čitalniško gibanje na zahodnem in vzhodnem robu slovenskega kulturnega prostora. V M. Jesenšek (Ur.), *Čitalništvo in bralno društvo pri Mali Nedelji* (str. 66–92). Univerza v Mariboru, Univerzitetna založba. https://doi.org/10.18690/um.ff.1.2022

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

Leech, G. N. in Short, M. (2007). *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. 2nd ed. Pearson Longman. Pridobljeno 10. maja 2024, https://sv-etc.nl/styleinfiction.pdf

Ljubešić, N. in Dobrovoljc, K. (2019). 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*. Association for Computational Linguistics, str. 29–34, doi:10.18653/v1/W19-3704.

McIntyre, D. (2015). Towards an integrated corpus stylistics. *Topics in Linguistics 16*(1), 59–68. https://doi.org/10.2478/topling-2015-0011

Mikolič, V. (2000). Povezanost narodne in jezikovne zavesti. *Jezik in slovstvo 45*(5), 173–186. Pridobljeno 15. maja 2024, http://www.dlib.si/?URN=URN:NBN:SI:DOC-5BD79SLS

Mikolič, V. (2020). *Izrazi moči slovenskega jezika*. Annales ZRS; Slovenska matica.

Mikolič, V. (2022). *Ali bereš Cankarja?*. Slovenska matica.

Mosteller, F. in Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley. Pridobljeno 20. maja 2024, https://archive.org/details/inferencedispute00most/page/n3/mode/2up

Nećak-Lük, A. et al. (1998). *Medetnični odnosi v slovenskem etničnem prostoru*. Izsledki projekta. Inštitut za narodnostna vprašanja.

Onič, T. (2014). Univerzalnost literarnega sloga: vpogled v grafični roman. *Primerjalna književnost (Ljubljana), 37*(3), 179–198. Pridobljeno 15. maja 2024, https://ojs-gr.zrc-sazu.si/primerjalna_knjizevnost/article/view/6296/5954

Perenič, U. (2012). Čitalništvo v perspektivi družbenogeografskih dejavnikov. *Slavistična revija, 60*(3), 365–382. Pridobljeno 15. maja 2024, https://srl.si/ojs/srl/article/view/COBISS_ID-50413154

Scherrer, Y. in Ljubešić, N. (2016). Automatic Normalisation of the Swiss German ArchiMob Corpus Using Character-Level Machine Translation. *Proceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016), str. 248–255.

Schöch, C., Erjavec, T., Patraș, R. in Santos, D. (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*. DOI: 10.3828/mlo.v0i0.364.

Ulčnik, N. (2010). Slomškove Drobtinice. *Studia Historica Slovenica, 10*(2-3), 683–703. Pridobljeno 10. maja 2024, https://shs.zgodovinsko-drustvo-kovacic.si/sites/default/files/shs2010_2-3.pdf

Žejn, A. (2020). Računalniško podprta stilometrična analiza pripovedne literature Janeza Ciglerja in Christopha Schmida v slovenščini. *Fluminensia: časopis za filološka istraživanja, 32*(2), 137–158. https://doi.org/10.31820/f.32.2.5

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

CORPUS CVET 1.0: CREATION, DESCRIPTION AND ANALYSIS OF A COLLECTION OF OLDER TEXTS IN RELIGIOUS PERIODICALS

The paper presents the process of creation and linguistic tagging of 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* in the period 1881–1916. The texts were obtained in PDF format from the dLib portal, edited in the Word editor and then converted to TEI. Older words were automatically updated using an open-source normalisation tool, which facilitates corpus search and further analysis of the material. The article points out some errors that occurred during normalisation, which will be corrected manually in the next version of the corpus (e.g. *keterim* > ketim* > katerim; *kesneje* > kosno* > kasneje; *sobrat* > zbrat* > sobrat). The updated texts were then automatically linguistically annotated, including morphosyntactic annotations as well as morphological and syntactic annotations according to the Universal Dependencies Formalism for Slovenian. We converted the TEI-encoded versions into various formats and published the collection under an open licence in the CLARIN.SI repository and concordancers suitable for linguistic analysis of the material. The second part of the paper presents an example of the analysis of the author's narrative style performed with noSketchEngine, based on the frequency variables of the most and least frequent words and keywords.

Keywords: historical Slovenian language, religious texts, TEI, normalisation, stylistic analysis, lexis

SLOVENIAN PARTICLE: NOT A SYNTACTIC CATEGORY

Jakob LENARDIČ
Institute of Contemporary History, Ljubljana

This paper claims that there is insufficient syntactic or semantic evidence to distinguish particles as a separate part of speech from adverbs, a distinction which is otherwise made in mainstream Slovenian descriptive linguistics and which is also used for the morphosyntactic annotations of Slovenian corpora. To this end, the paper presents a corpus investigation which probes into the clausal distribution of frequent words tagged as particles on the one hand and as adverbs on the other, showing that there are no appreciable differences there. The paper also proposes that the well-known *wh*-question test, which is otherwise used to determine sentence elements like subject, object, and adjunct, is a theoretically weak and empirically inconsistent criterion for determining category membership. Furthermore, the paper argues that common interpretative criteria used to distinguish particles from adverbs suffer from conceptual fallacies and are empirically unmotivated.

Keywords: particle, adverb, corpus linguistics, theoretical linguistics, syntax

1 INTRODUCTION

The aim of this paper is to challenge the traditional viewpoint of Slovenian descriptive linguistics according to which the words in (1) belong to a different syntactic category (or part of speech) from the prototypical adverbs in (2) – namely, to the syntactic category of the so-called particles. This viewpoint, which I will argue against, is advocated for by the Slovenian descriptive grammar *Slovenska slovnica* (Toporišič, 2000) and is maintained in the work of many other authors as well (e.g., Žele, 2014; Žele, 2015; Krvina & Žele, 2018; Jakop, 2001).

(1) Particles
  a. *morda* 'possibly'
  b. *tudi* 'also'

c. *ne* 'not'
d. *naj* 'should'

(2) Adverbs

a. *včeraj* 'yesterday'
b. *danes* 'today'
c. *zlahka* 'easily'
d. *tam* 'there'

The question of how to classify these words, and whether to make a division between them, is not only relevant for theoretical purposes, but also (and perhaps even more so) for language resource creation. The linguistic annotations of the major Slovenian corpora, such as the written reference corpus *Gigafida 2.0* (Krek et al., 2019), the *Corpus of academic Slovene KAS 2.0* (Žagar et al., 2022), and the computer-mediated-communication corpora of the *JANES* family (Fišer et al., 2020), follow the MULTEXT-East specifications (Erjavec, 2017),[1] which recognise particles as a category separate from adverbs. Similarly, the annotated corpora of the *ParlaMint* project (see Erjavec et al., 2023, for the latest, annotated, version), whose creation is spearheaded by the Slovenian CLARIN.SI consortium,[2] make use of the Universal Dependencies schema (De Marneffe et al., 2021), which likewise posits that particles constitute their own category (see also Dobrovoljc et al., 2023, for a description of the Slovenian Universal Dependencies guidelines).

In this paper, I will claim that there is insufficient linguistic evidence – either syntactic or semantic – that would warrant treating words like those in (1) as being appreciably different grammatically from the words in (2). In other words, there is no good reason not to assume simply that the words in both (1) and (2) are anything other than adverbs.

The paper is structured as follows. In Section 2, I first present a corpus investigation using the *Gigafida 2.0* corpus with which I try to determine if words tagged as particles substantially differ in their clausal distribution from words tagged as adverbs. In Section 3, I criticise the well-known *wh*-question test ac-

[1] See http://nl.ijs.si/ME/V6/msd/html/msd-sl.html for the current – i.e., sixth – version of the Slovenian specifications.
[2] https://www.clarin.si

cording to which particles unlike adverbs cannot be targeted for *wh*-question formation; I also argue that particles are not a semantically well-defined category either, and that many particles are in fact quite similar to adverbs interpretatively. In Section 4, I end the paper with a discussion which highlights the pitfalls of relying on mostly interpretative rather than syntactic criteria for determining parts of speech and argue for a syntax-first approach.

2 THE SYNTACTIC DISTRIBUTION OF PARTICLES

2.1 A tentative syntax for particles

In *Slovenska slovnica* 'The Slovenian Grammar', particles are defined as follows:

> Particles are an uninflected syntactic category: they are used to establish connections to the wider discourse; they express certain semantic flavours of individual words, parts of the clause or entire sentences; they are also used to derive syntactic modalities. Some particles are functionally similar to conjunctions, others to adverbs. Particles are not constituents of the clausal structure, in the framework of which they occur; rather, they are fragments that substitute elided clauses which can express additional contextual meanings. For instance, *Sosedovi imajo* ***samo*** *enega otroka* 'The neighbours have only one child → *Sosedovi imajo enega otroka, imeli pa bi jih lahko več* 'The neighbours have one child, but they could have more of them.' (Or: *Navadno je v družinah več kot en otrok* 'There is usually more than one child in a family'). About one quarter of the particles are homophonous with adverbs and conjunctons.'
>
> (Toporišič, 2000, p. 445; translation by JL)

Toporišič (2000) then goes on to propose various ways of classifying particles semantically, drawing up classes like *pozivni členek* 'particle of addressing' and *vrednotenjski členek* 'particle of evaluation'.[3] In general, this approach and the definition above is primarily semantic in nature, with the only ostensibly syntactic characteristic being the idea that particles are not constituents of clausal structure, to which point I will return in Section 3.1.

If particles indeed constitute a different part of speech than adverbs as per the definition above, they should differ from one another in some obvious syntagmatic way, perhaps in the way that they are distributed in the clause.

On the face of it, many canonical particles, like *samo* 'only' in the definition above, do indeed seem to show a different clausal distribution from typical adverbs like *hitro* 'fast', primarily in relation to the clause initial position. In this respect, Marušič and Žaucer (2010) observe that the particle *že* 'already' is a phonologically weak element that cannot by itself host the so-called Wackernagel clitic cluster (Anderson, 1993). The cluster occupies the second syntactic position in the clause and has to follow a phonologically non-weak word to ensure well-formedness.[4] To see this, compare the grammaticality of the constructed examples (3a) and (4a),[5] in which the clitic cluster is typeset in bold and is directly preceded by an adverb, with the ungrammaticality of the examples in (3b) and (4b), in which the clitic cluster is directly preceded by the particle *že* 'already'.

(3) a. Že velikokrat **ji** **je** stopil na prste.
already manytimes she.Dat is stepped on toes
'He has stepped on her toes many times.'

[3] When exemplifying the various classes, Toporišič (2000) identifies phrasal particles as well, which to my mind further underscores the idea that particles taxonomically are not to be treated as a part of speech, but rather as a special interpretation of individual words, which need not be just adverbs, or of phrases. When for instance Smolej (2004) talks about particles, she uses the term to refer to phrases such as *konec koncev*, which are idiomatically used for discourse-organisational purposes; in terms of parts of speech, however, this is simply a noun complex in which the first noun is in nominative (i.e., *konec* '(the) end') and the second (i.e., *koncev* 'of ends') in genitive.

[4] See also Marušič (2008) for further observations on the seeming second-position placement of the clitic cluster.

[5] I will try to make it as clear as possible when I am talking about constructed or about corpus examples. Additionally, all the examples are glossed using the abbreviations of the Leipzig Glossing Rules (https://www.eva.mpg.de/lingua/resources/glossing-rules.php) – so for instance Dat in the interlinear glosses of example (3a) stands for Dative, 1Sg in example (5) for first person singular, and so on.

b. *Že **ji je** velikokrat stopil na prste.

(4) a. Že od včeraj **mi je** Peter dolžen 5 Eur.
already since yesterday I.Dat is Peter indebted 5 Eur
'Peter has owed me 5 Eur already since yesterday.'

b. *Že **mi je** Peter dolžen 5 Eur.

(Marušič & Žaucer, 2010, examples (2)–(3))

Similarly, the negator *ne* is also typically analysed as a phonologically weak element that in finite declarative clauses forms a syntactic unit with the finite verbal element (Ilc, 2008), together with which it can play host to the clitic cluster:

(5) [Ne dam] **ti ga**.
Neg give.1Sg you.Dat him.Acc
'I won't give it to you.'

(Milojević Sheppard & Golden, 2000, example (11))

Unlike the particle *že* 'already' in examples (3)–(4), prototypical adverbs can be left-adjacent to the second-position clitic cluster, as shown by the following constructed example.

(6) Danes **ga je** videl.
today him.Acc is saw
'Today he saw him.'

On the basis of these examples, it seems that particles could be – by inductive reasoning – defined as in (7).

(7) **(Tentative) particle definition**
A phonologically weak adverb, which cannot host a second-position clitic.

If it turns out that Slovenian particles differ from adverbs systematically in their clausal distribution along these lines (i.e., particles never directly precede the clitic cluster), then a case can indeed be made for their constituting a distinct syntactic category. It is worth noting that in English, words identified as particles do have a unique distribution that is tied to their being phonologically weak or clitic-like. Such English particles are words like *up*, which differ from prototypical adverbs in that they are invariably constituents of the verb phrase: *John looked the information* ***up*** or *John looked* ***up*** *the information*, but not **Up* *John looked the information* (Den Dikken, 1995, p. 1).

2.2 Corpus investigation

In order to test (7), I have conducted a corpus investigation that checks how Slovenian particles pattern with the Wackernagel clitic cluster. According to the constructed examples (3b) and (4b), the prediction is that particles should not be able to directly precede the cluster, as there should necessarily be a phonologically strong element like an adverb that intervenes between the particle and the cluster. The corpus used for the investigation is the deduplicated version of *Gigafida 2.0* (Krek et al., 2019).

Using CLARIN.SI's noSketch Engine concordancer,[6] I have first created a frequency list of the top 30 lemmas tagged as particles in the deduplicated *Gigafida 2.0*. This frequency list is shown in Table 1.

Many of the most frequent particles in this Table are words like *tudi* 'also', *ne* 'not', *še* 'yet', and *že* 'already', which intuitively adhere to the clitic-adjacency restriction described in the previous section. There are, however, several words, like *naj* 'should' and *morda* 'possibly', which a priori seem to show no such restriction and can host the clitic cluster; compare the constructed example in (8) with (3b) and (4b) from the previous section.

(8) Morda **ga** **je** videl.
possibly him.Acc is saw
'He has possibly seen him.'

So it seems rather that only a subset of the words tagged as particles in *Gigafida 2.0* adhere to (7).

[6] https://www.clarin.si/ske/

Table 1: The 30 most frequent particles in the *Gigafida 2.0* corpus.

Lemma	**Frequency**	**Lemma**	**Frequency**
tudi ‘also’	8,478,793	*morda* ‘possibly’	468,348
ne ‘not’	6,734,994	*niti* ‘not even’	458,790
še ‘yet’	5,787,718	*sploh* ‘really/even’	419,675
že ‘already’	3,735,419	*šele* ‘let alone’	388,144
le ‘only’	2,165,776	*pač* ‘but’	294,972
naj ‘should’	1,598,002	*zgolj* ‘just’	254,436
prav ‘precisely’	1,208,290	*zlasti* ‘especially’	242,203
sicer ‘otherwise’	1,078,123	*ravno* ‘just’	247,650
samo ‘only’	940,204	*pravzaprav* ‘actually’	198,500
predvsem ‘mostly’	844,382	*vsekakor* ‘most certainly’	159,819
seveda ‘of course’	725,852	*no* ‘well’	142,860
celo ‘even’	646,664	*najbrž* ‘likely’	123,679
skoraj ‘almost’	528,454	*menda* ‘evidently’	132,347
več ‘more’	520,542	*koli* ‘ever’	107,068
vsaj ‘at least’	507,625	*češ* ‘as if’	87,039

To see if this is indeed the case, I have also queried each particle from Table 1 with a CQL search string like the one in (9), which is concretely defined for the word form *Tudi* ‘also’.

(9) [word = "Tudi" & tag = "L"][tag = "Z.*k" & word != "mi"]

This search string yields bigrams in which the first element corresponds to a capitalised word belonging to the particle class, which is specified by the MSD tag L. Capitalisation is used as a heuristic to ensure that the particle is sentence initial, in parallel to the sentence-initial particle in the examples in (3)–(4). The second element is any pronominal clitic (defined by the MSD tag Z.*k). The word *mi* is filtered out because many cases of the plural nominative pronoun *mi* ‘we’ are incorrectly tagged with the same MSD as the syncretic singular dative pronoun *mi* ‘to me’. For instance, the form *mi* in the *Gigafida 2.0* examples in (10) is tagged in both cases as a dative clitic, concretely with the tag Zop-ed--k or the English equivalent Pp1-sd--y (Pronoun Type=personal Person=first Number=singular Case=dative Clitic=yes).

Table 2: Phonologically weak particles followed by a pronominal clitic.

Particle	**Capitalised word + clitic F.**	**Relative F.**	**Lemma F.**
tudi 'also'	775	91.4	8,478,793
ne 'not'	939	139.4	6,734,994
še 'yet'	679	117.3	5,787,718
že 'already'	477	127.7	3,735,419
le 'only'	41	18.9	2,165,776
prav 'precisely'	695	575.2	1,208,290
samo 'only'	177	188.3	940,204
celo 'even'	16	24.7	646,664
več 'more'	0	0.0	520,542
vsaj 'at least'	12	23.6	507,625
niti 'not even'	431	939.4	458,790
šele 'let alone'	9	23.2	388,144
pač 'but'	66	223.8	294,972
zgolj 'just'	0	0.0	254,436
ravno 'just'	363	1,465.8	247,650
zlasti 'especially'	603	2,489.6	242,203
no 'well'	17	119.0	142,860
koli 'ever'	0	0.0	107,068
češ 'as if'	0	0.0	87,039
Σ	5,300	160.9	32,949,187

(10) a. Tudi **mi** jih bomo "morali" poslušati.
also we.Nom them will.1Pl must.Pl listen.Inf
'We too will have to listen to them.'

b. Tudi **mi** ne gre več.
also I.Dat Neg goes anymore.
'I am no longer able to do it.'

But in actuality, *mi* in (10a) is a nominative pronoun, which in contrast to the dative form in (10b) is crucially not a clitic and is therefore irrelevant for the present investigation.

The results are given in Tables 2 and 3, which are organised as follows. The second column, labelled as Capitalised word + clitic F., gives the absolute frequency of the bigrams retrieved with search strings like (9) (relevantly modified for each particle separately). The column Relative F. provides the per-million-token frequency of the bigram relative to the absolute frequency of each lemma, which is reported in the last column.

The division of the particles between Tables 2 and 3 intuitively corresponds to their phonological strength. All the particles in Table 2 are those that are expected to pattern with the constructed examples in (3)–(4) in disallowing direct adjacency to a second-position pronominal clitic, while the particles in Table 3 are those that are intuitively expected to pattern with regular adverbs in allowing strict adjacency with the clitic, as in (6) and (8).

Interestingly, most of the particles in Table 2 can indeed be followed by a pronominal second position clitic, in contrast to what is expected by the assumed ungrammaticality of examples (3b) and (4b). In this Table, there are 775 examples of capitalised (therefore sentence-initial) *tudi* followed by a second-position clitic; 939 examples of capitalised *ne* followed by such a clitic; 679 examples of *še*; and so on. The following is a corpus example of capitalised *že* followed by the dative feminine clitic *ji* 'to her'; recall from the reported judgement of the constructed example (3b) that this string is expected to be ungrammatical. Indeed, the attestastion of such examples seems to conform to the idea that the placement of clitics in Slovenian is not fixed in any true sense, so their occurring second-position is merely a tendency (Marušič et al., 2024), albeit a strong one.

(11) Že ga je v mislih videla, kako stopa izza
already him.Acc is in thoughts saw.F how steps from-behind
grma in ji hiti naproti.
bush.Gen and her.Dat hurries towards
'In her thoughts she already saw him step out of the bush and hurry towards her.'

There is an important difference, however, in the distribution of the particles: most of the particles in Table 3 are much more frequent in this sentence-initial position followed by a clitic in comparison to the particles of Table 2. For instance, the relative frequencies of 6 out of the 11 capitalised particle + clitic bigrams in Table 3 – that is, the bigrams containing capitalised *seveda*

Table 3: Phonologically non-weak particles followed by a pronominal clitic.

Particle	**Capitalised word + clitic F.**	**Relative F.**	**Lemma F.**
naj 'should'	12,076	7,556.9	1,598,002
sicer 'otherwise'	6,150	5,704.4	1,078,123
predvsem 'mostly'	4,406	5,218.0	844,382
seveda 'of course'	21,650	29,827.0	725,852
skoraj 'almost'	730	1,381.4	528,454
morda 'possibly'	18,218	38,898.4	468,348
sploh 'really/even'	2,414	5,752.1	419,675
pravzaprav 'actually'	6,320	31,838.8	198,500
vsekakor 'most certainly'	5,014	31,373.0	159,819
menda 'evidently'	3,353	25,334.9	132,347
najbrž 'likely'	3,812	30,821.7	123,679
Σ	84,143	13,404.6	6,277,181

'of course', *morda* 'maybe', *pravzaprav* 'actually', *vsekakor* 'definitely', *menda* 'maybe', and *najbrž* 'probably' – are above 10,000 tokens per million, while all of the bigrams in Table 2 – save for those with capitalised *ravno* 'just' and *zlasti* 'mostly' – have relative frequencies below 1,000 tokens, many of them even below 100 tokens per million (e.g., *tudi* 'also', *le* 'only', *celo* 'even'). Taken together, the bigrams of Table 2 have an overall relative frequency of 160.9 tokens per million while the bigrams of Table 3 have an overall relative frequency of 13,404.6 tokens per million, which is roughly 83 times as many.

Let us now see how such particles compare to some of the most frequent lexemes tagged as regular adverbs in *Gigafida 2.0*. Table 4 shows how frequently 10 such adverbs play host to a second-position pronominal clitic. The main thing to notice in this Table is that the relative frequencies practically mirror those of the "non-weak" bigrams in Table 3. Six out of 10 bigrams (those with the capitalised adverbs *lahko* 'possibly/easily', *danes* 'today', *potem* 'then', *letos* 'this year', *nato* 'then', and *včeraj* 'yesterday') show relative frequencies above 10,000 per million tokens, while the overall relative frequency – that is, 13,748.6 per million – is barely higher than that of Table 3, which is 13,404.6 tokens per million.

To test these overall differences statistically, I have used the *Calc: Corpus Calculator* tool (Cvrček, 2021), which calculates χ^2 values (among other such test statistics) for the differences in pairwise absolute frequencies across differently sized sets of words. Aside from this, *Calc* also provides for each difference a DIN value, which is an effect size metric (Fidler & Cvrček, 2015).

Let us first compare the difference between the overall frequency of the "weak" particle bigrams (5,300 tokens out of 32,949,187; relative frequency = 160.9) in Table 2 and the "strong" particle bigrams (84,142 tokens out of 6,277,181; relative frequency = 13,404.6) in Table 3 on the other. This difference has a χ^2 score of 40,6508.45 and a p value of < .00001, which is statistically significant at the .05 cut-off point.

Let us now turn to the markedly smaller difference between the overall frequency of the "strong" bigrams (84,142 tokens out of 6,277,181; relative frequency = 13,404.6) in Table 3 and the adverbs (173,823 out of 12,642,922 tokens; relative frequency = 13,748.6) in Table 4. This difference has a χ^2 score of 36.95 with a p value of < .00001, which is also statistically significant at the .05 cut-off point.

As pointed out by Fidler and Cvrček (2015), the issue here is that when comparing large sample sizes, even differences that intuitively seem very small and possibly irrelevant quickly turn out to be statistically significant, which is what seems to be happening with this second difference between the "strong" particle and adverb bigrams. This is where the DIN effect-size metric comes into play. The DIN score for the overall difference between the "weak" and "strong" particle bigrams is $\pm$97.62, while the DIN score for the difference between the "strong" particle bigrams and the adverb bigrams is $\pm$1.26.

DIN values are to be interpreted as follows.

(12) a. A value of 0: the word occurs equally often in both corpora.
b. A value of $\pm$100: the word is present only in one of the two corpora.

(Adapted from Fidler & Cvrček, 2015, p. 230)

According to (12), the first DIN score of $\pm$97.62 indicates that the "strong" particles really do show a marked preference for being adjacent to second position

Table 4: Adverbs followed by a pronominal clitic.

Adverb	**Capitalised word + clitic F.**	**Relative F.**	**Lemma F.**
lahko ‘easily/possibly’	24,988	7,163.8	3,488,074
tako ‘thus’	56,520	17,182.5	3,289,402
vedno ‘always’	7,722	7,182.4	1,075,127
dobro ‘well’	6,186	6,824.8	906,394
danes ‘today’	19,339	21,365.6	905,146
potem ‘then’	22,895	28,107.2	814,559
skupaj ‘together’	1,954	3,179.6	614,538
letos ‘this year’	10,259	16,907.8	606,763
nato ‘then’	18,328	31,121.5	588,918
včeraj ‘yesterday’	5,632	15,909.6	354,001
Σ	173,823	13,748.6	12,642,922

clitics in comparison to the “weak” particles (even though such weak particles are attested in such positions), while the DIN score of ± 1.26 indicates that the “strong” particles are just as likely to directly precede second position clitics as the regular adverbs.

To my mind, there are two ways to interpret this in light of the present paper’s main claim. One interpretation takes these differences in effect sizes seriously, whereby it is possible to claim something like the following: while certain Slovenian particles, namely those of Table 2, indeed differ from the adverbs distributionally, there exists a subset of particles, namely those of Table 3, that does not differ from them at all. According to this point of view, the particle category syntactically overlaps with the adverbs. But this of course raises the question why not consider this overlapping part as simply belonging to regular old adverbs – one answer is perhaps due to non-syntactic but interpretative factors, but see Section 3.2 for an argument against this as well.

The other way of looking at this is that the corpus data just show a tendency for the “weak” particles to avoid positions where they host second position clitics, but that this is not a hard and fast rule – if it were, then no such “weak” particles would be attested in this position, contrary to fact. According to this interpretation, then, particles are not distributionally different from the adverbs at all, given that both can technically precede the clitics, albeit at different frequencies.

No matter which interpretation ultimately turns out to be correct (the answer might boil down to one's attitude towards the interpretation of word frequency in corpora, and how grammaticality is related to attestation, a non-trivial issue), neither seems to serve as a strong argument for the treatment of particles as a syntactically coherent category that is appreciably different from the adverbs.

3 OTHER DIAGNOSTICS FOR CATEGORY MEMBERSHIP

3.1 Wh-question formation

One classic (and very informal) argument in favour of treating particles as belonging to a different category from adverbs is related to the formation of *wh*-questions (Toporišič, 2000). Notice that adverbs can freely undergo *wh*-question formation, as in (13), where *kdaj* 'when' in the *wh*-question (13b) "substitutes" the adverb *včeraj* 'yesterday' in (13a). Particles, by contrast, are unable to undergo *wh*-question formation; descriptive Slovenian grammarians also typically tie this seemingly syntactic constraint to the idea that particles differ from adverbs semantically, in which sense they are closer to functional rather than lexical words (Jakop, 2001) or perhaps they are interpretatively something else entirely (see Section 3.2).

(13) a. **Včeraj** smo šli domov.
yesterday Aux.1Pl went home
'We went home yesterday.'
b. **Kdaj** smo šli domov?
when Aux.1Pl went home
'When did we go home?'

The *wh*-question test is used to identity sentence elements. The idea goes that since particles cannot be targeted by *wh*-question formation, they do not constitute a sentence element in contrast to adverbs, where for instance *Včeraj* in (13) would be analysed as an adjunct. Recall from the beginning of Section 2.1 that *Slovenska slovnica* considers particles not to be "constituents of clausal structure" (Toporišič, 2000, p. 445). What seems to be meant by this is that particles are not sentence elements like adjuncts.

But relying on sentence elements is inappropriate for determining category membership in any case. Notice that verbs are also strictly speaking unable to undergo *wh*-question formation (yet they are traditionally considered to constitute their own syntactic category). What is possible with verbs is to form a *wh*-question that substitutes both the lexical verb (and possible verbal auxiliaries) and its possible object complement (14), but there is no operation of *wh*-question formation that would target just the lexical verb (to the exclusion of everything else), in parallel to the *wh*-question that substitutes just the adverb in (13). Notice also that with forming *wh*-questions out of verb phrases there is no 1-to-1 substitution, but the supporting verb *delati* has to be invoked aside from the *wh*-word *kaj* 'what', which is quite unlike the simple *wh*-substitution with the adverb in (13).

(14) a. Moj prijatelj **igra** **računalniško** **igro**.
my friend plays computer.Adj.Acc game.Acc
'My friend is playing a video game.'

b. **Kaj** **dela** moj prijatelj?
what.Acc does my friend
'What is my friend doing?'

In terms of sentence elements, example (14a) would be parsed as in (15).

(15) $\underbrace{\text{Moj prijatelj}}_{\text{Subject}}$ $\underbrace{\text{igra}}_{\text{Predicator}}$ $\underbrace{\text{računalniško igro}}_{\text{Direct object}}$.

But the impossibility to form a *wh*-question out of just the verb in (14) shows that this test is not sensitive to this kind of division;[7] rather, *wh*-question formation is sensitive to the phrase structure in (16), as it can only target the entire VP

[7] Determining what is and is not a sentence element is also quite arbitrary and depends on the particular grammatical tradition of language analysis – for instance, in English, words like *possibly* (a close equivalent for Slovenian *morda*) are in the descriptive *Comprehensive Grammar of the English Language* (Quirk et al., 1985) indeed treated as sentence elements, namely as subjuncts, which are a subtype of adverbials along with adjuncts and disjuncts. Furthermore, sentence elements are simply metalinguistic descriptors for common syntactic patterns of argument structure

igra računalniško igro or the NP *računalniško igro* 'computer game' contained within it, but not the V *igra* 'plays' itself, which is a syntactic head rather than a phrase.[8]

(16) $[_{vP}$ $[_{NP}$ Moj prijatelj$]$ $[_{VP}$ igra $[_{NP}$ računalniško igro$]]]$

Furthermore, it is not strictly speaking true that it is impossible to form questions that output particles. Such questions just are not *wh*-ones, as shown by the pair in (17).

(17) a. Ali je on to naredil?
Q is he this.Acc made
'Has he done this?

b. Morda.
possibly
'Possibly.'

Consequently, the fact that particles cannot undergo *wh*-question formation is thus not a syntactic constraint, but is rather the result of a simple lexical gap according to which no *wh*-word exists in the Slovenian lexicon (or perhaps universally) that is interpretatively akin to words identified as the particles. Relying on such lexical gaps is not, by itself, a sufficient criterion for determining membership in an ultimately *syntactic* category; indeed, apart from verbs, other categories such as conjunctions and interjections also do not have *wh*-word equivalents, yet they are typically considered to be members of their own category all the same (precisely because of their syntactic/distributional characteristics, which distinguish them from one another). But, as was shown in Section 2.2, syntactically particles can occupy the same clausal positions as adverbs.

3.2 The semantics of particles and adverbs

Krvina and Žele (2018) distinguish particles from adverbs on semantic grounds, claiming that particles are neither lexical nor functional words. This idea of

realisation rather than real linguistic features, so they are only indirectly related to syntactic categories.

[8] In (16), *v*P is the phrase in which the external argument – that is, the NP *moj prijatelj* – is introduced; see (Larson, 2014).

particles supposedly going beyond the lexical–functional divide is tied to the claim by Žele (2014, pp. 9–10) that particles are not part of (the logical content of) the proposition but that they merely modify it.[9]

Let us return to the particles in Table 1. I would here like to propose that this Table actually contains 2 groups of words from the perspective of semantics. In fact, it seems that the semantic division between the two groups that I will propose shortly more or less aligns with the division between Tables 2 (the phonologically "weak" particles) and 3 (the "strong" particles). Note that a similar division is proposed by Žele (2014), who claims that the two main groups of particles in Slovenian correspond to discourse particles (Table 2) and modal particles (Table 3). Contra Žele (2014), however, I will now claim that these two groups are substantially different from each other semantically, and cannot be subsumed under her umbrella category of "propositional modifiers" (*ibid.*, p. 10)

In Table 2, there are words like *tudi* 'also', *že* 'already', and *samo* 'only'. The main semantic characteristic of these words is that they preserve the truth conditions of the proposition they modify – according to Grosz (2020), such particles "simply constrain the context in which an utterance is felicitous by acting as truth-conditionally vacuous presupposition triggers" (see also Kaufmann, 2010). For instance, the presence of *tudi* in example (18) (taken again from *Gigafida 2.0*) does not change the truth conditional meaning of the prejacent proposition – that is, regardless whether the particle is present or not, the example conveys that what one will be able to see is a set of many colour slides; the particle merely lexicalises the presupposition that other things aside from the slides will be on display as well. These words, then, are presuppositional rather than propositional modifiers.

(18)	Na	ogled	bodo	**tudi**	številni	barvni	diapozitivi.
	on	*display*	*will-be.3Pl*	*also*	*many*	*colour.Adj.Nom.Pl*	*slides.Nom*

'One will also be able to see many colour slides.'

[9] This claim is made on functionalist rather than formalist grounds, and crucially for Žele (2014), only sentence elements make up the propositional content of an utterance. But the classification of particles as not sentence elements is made on shaky ground in any case; see the criticism of the *wh*-question test in the previous Section as well as Footnote 7.

In Table 3, there are words like *morda* 'maybe', *seveda*, *naj* 'should', and *najbrž* 'possibly'. These words are semantically completely different from the presuppositional modifiers, as they do not preserve the truth conditions of the prejacent proposition. They correspond to modal expressions and as such they shift the interpretation of the prejacent proposition into the realm of possibility and necessity, so that the sentence modified this way no longer speaks of what holds in the actual world (von Fintel, 2006). Unlike the presuppositional modifiers, these expressions have an evaluative component to them – a word like *morda* 'possibly' is an epistemic modal (Gomboc Čeh, 2021, pp. 27–28), which means that the possibility expressed is tied to the speaker's world knowledge. What is then conveyed by *morda* 'possibly' is not only that the proposition corresponds to a possibility, but that the speaker is uncertain about it as well.

For the particles in this second group, it is quite unclear to me why they should be distinguished from prototypical adverbs at all, save of course from possible pretheoretically understood differences in degrees of abstraction. One could speculate, following Jakop (2001), that the particles correspond to functional words whereas adverbs are lexical words. But it is not the case that all adverbs semantically correspond to lexical words, which means that a lexeme being a lexical word cannot be a necessary condition for its membership in the adverb category either.

Adverbs like *takrat* 'at that time', *potem* 'then' are not referring expressions that pick out and crucially provide encyclopedic labels for events or individuals, which is how lexical words are usually conceptualised with regard to their semantics (Cann, 1999), but play a wholly functional role; *takrat* 'at that time', for instance, is a definiteness operator that picks out a temporal interval or point before the time of utterance, so it is similar to the article *the* in English, a functional word that picks out a unique individual in the context. Similarly, an adverb of frequency like *občasno* 'sometimes' semantically corresponds to a quantifier over temporal intervals/points (Lepore & Ludwig, 2007), just like *najbrž* 'probably' corresponds to a quantifier over possibilities (or possible worlds). It seems, then, that a word like *morda* 'possibly' is considered to be a particle rather than an adverb just because modality is conceptually more abstract than temporality. Apart from that, however, a particle like *morda* 'possibly' and an adverb like *občasno* 'sometimes' operate on the proposition in a completely parallel way.

In sum, only a subset of the lexemes understood as particles in contemporary Slovenian grammar substantially differ from adverbs in interpretation, this being the presuppositional modifiers of Table 2. But not only is the second group – that of Table 3 – similar to adverbs in its clausal distribution (see Section 2.2), we have just shown that it is also similar to them semantically in its altering of the truth conditions of the prejacent proposition. The particle category thus does not seem to be robustly defined on semantic grounds either.

4 DISCUSSION AND CONCLUSION

The main problem that goes beyond the putative particle–adverb distinction (but of which this distinction is a symptom) lies in what mainstream Slovenian grammarians like Toporišič (2000) consider as defining criteria for determining category membership. Rooted in general functionalist approaches to linguistic description rather than formal ones, such grammarians use mainly interpretative criteria for distinguishing syntactic categories, so for instance nouns are defined as those expressions that typically refer to persons, animals, and inanimate things,[10] adjectives are those linguistic expressions that head phrases which semantically characterise nouns, and so on.

There are two massive drawbacks of relying on interpretative criteria for defining category membership, however, especially if interpretation is the only criterion that is really used, as is the case of particles and adverbs in *Slovenska slovnica* (Toporišič, 2000). The first is that lexemes that are syntactically wholly different from one another quickly end up getting lumped together under the same category because of similar semantics. Let us for instance consider the modals *lahko* 'can' and *treba* 'must', which denote logical possibility and necessity, respectively (see Lenardič & Fišer, 2021, for a discussion). In the *Gigafida*

[10] The Universal Dependencies provides precisely such a semantic definition of nouns:

> Nouns are a part of speech typically denoting a person, place, thing, animal or idea.
>
> (https://universaldependencies.org/u/pos/NOUN.html)

2.0 corpus, whose tagset follows the *MULTEXT-East* specifications (Erjavec, 2017), both words are consistently tagged as adverbs. Similarly, in the Slovenian part of the annotated *ParlaMint 4.0* corpus (Erjavec et al., 2023), whose morphosyntactic tags follow a different tagset – that of the Universal Dependencies framework (Dobrovoljc et al., 2023) –, both words are also in all cases defined as adverbs.[11]

However, *treba* in no way syntactically behaves like *lahko*; while the latter is syntactically indeed an adverb, the former behaves like a predicative adjective that, on the one hand, complements a copular verb (Rothstein, 1999), and on the other, combines in standard Slovenian with an infinitival verbal phrase (and rarely with a finite clause according to Uhlik, 2016, p. 55). (Note that non-finite complementation of adjectives is cross-linguistically common; see Rickman & Rudanko, 2018, for English examples. Additionally, such adjective+non-finite-verb combinations typically have a modal, evaluative, component to them, just like *treba*.) Since it is *treba*, a predicative adjective, that syntactically governs the type of the verbal phrase it combines with (i.e., the infinitival *povedati* in example (19a)), its omission from the sentence results in ungrammaticality, while *lahko*, being an adverb, is always syntactically omissible (though of course dropping it changes the meaning of the sentence). This is shown by the pair of examples in (19) from the Slovenian subset of the *ParlaMint 4.0* corpus, where the non-optionality of *treba* is indicated by *(.

(19) a. Hkrati pa je *(treba) povedati, da se študentje
simultaneously but is must tell.Inf that Refl students
vedno vključujejo ...
always join-in.3Pl
'But at the same time it *(must) be said that students always join in...'

b. Tako da je (lahko) študent tudi neodvisen.
thus that is possibly student also independent
'Thus the student is also (possibly) independent.'

[11] The Slovenian Universal Dependencies do recognize the particle class, but consider it to be a legacy category inherited from the *JOS* schema (Erjavec et al., 2010); see here https://universaldependencies.org/sl/pos/PART.html.

The second problem is that using semantics as a criterion for determining syntactic category membership can quickly result in fuzzily demarcated groups of words which contain members that are so to speak neither fish nor fowl interpretatively. As previously mentioned, Žele (2014) in her *Dictionary of Slovenian Particles* (see also Žele, 2015) defines two groups of particles in terms of interpretation – the modal particles, which modify the proposition and crucially for Žele (2014) have an evaluative component to them (e.g., *morda* 'possibly' conveys the lack of speaker's truth commitment; *skoraj* 'almost' is an approximator and thereby hedges the statement insofar as it denies a possible universal reading of the thing it modifies); and the discourse particles, which are used to organise the discourse but do not modify the prejacent presupposition and lack the evaluative component of the modals or approximators.

However, there is (at least) one word belonging to the group of phonologically weak adverbs (Table 2) which has interpretative properties of both groups – this is the negator *ne* 'not'. Like the discourse particles and unlike the modals, it has no inherent evaluative component to it (its use simply entails that a proposition is not the case); like the modals but unlike the discourse particles, it is a propositional rather than a presuppositional modifier (compare what the use of *ne* 'not' does to the truth conditions with what *tudi* 'also' does to them; i.e., only the latter preserves them). Thus, divvying up groups of words on the basis of informally defined semantic categories immediately runs into the problem of categorical fuzziness.

In light of this, I end this paper with a very brief proposal that is aimed both at descriptive Slovenian grammarians and at the developers of Slovenian corpora, who have to rely on some kind of theoretical basis when annotating the corpus data for parts of speech or even syntactic dependencies. The proposal is simply that syntactic category membership should first and foremost be defined on syntactic rather on semantic grounds (and never exclusively on the basis of semantics).

Adverbs could in this sense be defined as follows:

(20) **Adverb – a syntactic definition**
A word that heads a phrase which does not syntactically interact (e.g.,

by undergoing syntactic agreement or syntagmatically determining categorical selection) with the rest of the clausal structure in any way.

Virtually all other categories, save from possibly interjections, do interact with the rest of the clausal structure, by which I mean that they take part in observable and/or inferable syntactic dependencies. For instance, nouns head phrases which occupy clausal positions in which they receive case and, if in nominative, agree with the verb phrase. Lexical verbs determine transitivity, show Voice contrasts (active vs. passive vs. perhaps middle), agree with the noun phrase in nominative, etc. Auxiliary verbs need to be complemented by a participle. Attributive adjectives undergo syntactic agreement with the headword of the noun phrase in which they are embedded. Predicative adjectives like *treba* in (19a) determine the type of clause that syntactically complements them. Prepositions require nominal complementation and also determine the case value of the complement. And so on.

Of course, the way adverbs are distributed in a clause is not completely unconstrained, but the limited distribution invariably seems to be semantic rather than syntactic in nature (Ernst, 2007), stemming either from the semantics of the adverbs themselves (e.g., manner adverbs modify the lower parts of the clause in comparison to temporal adverbs or modals, so they tend to occur lower in the clause as well) or from the semantics of the main clause predicate (e.g., *to #deliberately contain something*). Syntax-wise adverbs can of course be constituents of different phrases, like the verbal or adjectival phrase (and possibly many others, see Cinque, 2004), but they never seem to establish any kind of *relational*(="syntagmatic") dependency with any other phrase in the clause.

A reviewer points out that *preveč* 'too much' in a structure like (21a) is involved in an observable dependency – namely, it governs the fact that the nominal *denarja* 'money' appears in genitive case. Traditionally (and in Slovenian corpora), *preveč* in such structures is indeed analysed as an adverb. According to (20), however, it is not, which is correct in my opinion. Syntactically, *preveč* 'too much' completely mirrors the syntactic behaviour of numerals assigning the so-called genitive of quantification (see also Stegovec, 2022), which a prototypical adverb does not do. For another structural parallel with the numeral, notice that when *preveč* occurs in subject position, as in (21a), clausal agreement fails to take place, resulting in the verbal phrase appearing with default agreement features (third-person singular auxiliary and neuter on the main verb in periphrastic tense constructions). This to me implies that such uses of words like *preveč* should be analysed – together with numerals – as their own separate part of speech, possibly as quantifiers/determiners, rather than as belonging to a category like adverbs from whose prototypical members they differ so substantially in terms of grammar.

(21) a. Preveč denarja je bilo v banki.
too-much money.Gen Aux.3Sg was.N in bank
'too much money'

b. Pet ljudi je šlo domov.
five people.Gen Aux.3sg went.N home
'Five people went home.'

The fact that there is no reference to semantics in (20) contrasts with the definition of adverbs in, for instance, the Universal Dependencies (De Marneffe et al., 2021) framework. Like *Slovenska slovnica* (Toporišič, 2000), the Universal Dependencies guidelines rely mainly on semantic criteria for defining syntactic categories – notice the reference to semantic notions such as 'place', 'direction', and 'manner' in the following definition:

> Adverbs are words that typically modify verbs for such categories as time, place, direction or manner. They may also modify adjectives and other adverbs, as in *very briefly* or *arguably wrong*.
>
> (https://universaldependencies.org/u/pos/ADV.html)

Note also the use of *typically* in this definition. I believe that syntactic categories should be defined more robustly than this, and (undefined) exceptions should be avoided.

If a definition like (20) is adopted, the particle class could be retained to refer simply to the subset of those adverbs whose clausal distribution is limited due to phonological considerations, but not as its own category independent of adverbs – this would then parallel the way in which nominal clitics are already treated as subsets of pronouns in the MULTEXT-East specifications (Erjavec, 2017). In my mind, relying on such syntactic criteria rather than interpretative ones would also simplify any potential work on manual annotation, since annotators would thereby be less likely to run into the aforementioned neither-fish-nor-fowl problem, which quickly arises whenever different parts of speech are stipulated (due to apparently different semantics) for words that in actuality are syntactically indistinguishable, as is the case of the current understanding of the division between particles and adverbs.

ACKNOWLEDGMENTS

Work described in this paper has been carried out in the context of the research project *Z6-4616: Slovenian Verbal Valency: Syntax, Semantics, and Use*, as well as under the auspices of the national research programme *P6-0436: Digital Humanities: Resources, Tools and Methods* and *DARIAH-SI: Slovenian Digital Research Infrastructure for the Arts and Humanities*.

REFERENCES

Anderson, S. R. (1993). Wackernagel's revenge: Clitics, morphology, and the syntax of second position. *Language*, *69*(1), 68–98. https://doi.org/10.2307/416416

Cann, R. (1999). Functional versus lexical: A cognitive dichotomy. In *The nature and function of syntactic categories* (Vol. 32, pp. 37–78). Brill.

Cinque, G. (2004). Issues in adverbial syntax: Taking up the gauntlet-adverbs across frameworks. *Lingua*, *114*(6), 683-710. https://doi.org/10.1016/S0024-3841(03)00048-2

Cvrček, V. (2021). *Calc 1.03: Corpus calculator*. Czech National Corpus. https://www.korpus.cz/calc/

De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Den Dikken, M. (1995). *Particles: On the syntax of verb-particle, triadic, and causative constructions*. Oxford University Press.

Dobrovoljc, K., Terčon, L., & Ljubešić, N. (2023). Universal dependencies za slovenščino: nove smernice, ročno označeni podatki in razčlenjevalni model. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, *11*(1), 218–246. https://doi.org/10.4312/slo2.0.2023.1.218-246

Erjavec, T. (2017). MULTEXT-East. *Handbook of linguistic annotation*, 441–462. https://doi.org/10.1007/978-94-024-0881-2_17

Erjavec, T., Fiser, D., Krek, S., & Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. In *LREC 2010 proceedings.* http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agerri, R., Agirrezabal, M., … Fišer, D. (2023). *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1860

Ernst, T. (2007). On the role of semantics in a theory of adverb syntax. *Lingua*, *117*(6), 1008-1033. https://doi.org/10.1016/j.lingua.2005.03.015

Fidler, M., & Cvrček, V. (2015). A data-driven analysis of reader viewpoints: Reconstructing the historical reader using keyword analysis. *Journal of Slavic linguistics*, 197–239. https://www.jstor.org/stable/24602151

Fišer, D., Ljubešić, N., & Erjavec, T. (2020). The janes project: language resources and tools for Slovene user generated content. *Language resources and evaluation*, *54*(1), 223–246. https://doi.org/10.1007/s10579-018-9425-z

Gomboc Čeh, K. (2021). Izražanje naklonskih pomenov z glagolom imeti. In D. Zuljan Kumar & H. Dobrovoljc (Eds.), *Škrabčevi dnevi 12. zbornik prispevkov s simpozija 2021* (pp. 23–33). https://www.dlib.si/details/URN:NBN:SI:DOC-SFGF9P4O

Grosz, P. G. (2020). Discourse particles. In *The wiley blackwell companion to semantics.* John Wiley & Sons. https://doi.org/10.1002/9781118788516.sem047

Ilc, G. (2008). O zanikanju in nikalnici v slovenščini. *Jezik in slovstvo*, *53*(2), 65–79. https://www.dlib.si/details/URN:NBN:SI:doc-CVP39DNK

Jakop, N. (2001). Funkcijska delitev členkov: značilnosti naklonskih členkov. *Jezik in slovstvo*, *46*(7–8), 305–316. http://www.dlib.si/details/URN:NBN:SI:DOC-SI4C339V

Kaufmann, S. (2010). *Strong and weak presupposition: German 'ja' under quantifiers.* (Manuscript)

Krek, S., Erjavec, T., Repar, A., Čibej, J., Arhar Holdt, Š., Gantar, P., … Logar, N. (2019). *Corpus of written standard slovene gigafida 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1320

Krvina, D., & Žele, A. (2018). O členkih, zlasti o njihovih razločevalnih lastnostih: poudarjen slovarski vidik. *Jezik in slovstvo*, *63*(1), 39–64. https://journals.uni-lj.si/jezikinslovstvo/article/view/17767

Larson, R. K. (2014). *On shell structure*. Routledge.

Lenardič, J., & Fišer, D. (2021). Hedging modal adverbs in slovenian academic discourse. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, *9*(1), 145–

180. https://doi.org/10.4312/slo2.0.2021.1.145-180

Lepore, E., & Ludwig, K. (2007). Temporal adverbials and quantifiers. In *Donald Davidson's truth-theoretic semantics.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199290932.003.0010

Marušič, F. (2008). Slovenian clitics have no unique syntactic position. In *Formal approaches to Slavic linguistics* (Vol. 16, pp. 266–281). Retrieved 20 May 2024, from https://www2.ung.si/~fmarusic/pub/marusic_2007_clitics_fasl.pdf

Marušič, F., Mišmaš, P., & Žaucer, R. (2024). Placement and ordering of the (en)clitics. In D. Šipka & W. Browne (Eds.), *The Cambridge handbook of Slavic linguistics* (p. 365–384). Cambridge University Press. https://doi.org/10.1017/9781108973021.018

Marušič, F., & Žaucer, R. (2010). *An argument against syntactic positioning of Slovenian 2nd position clitics.* https://www2.ung.si/~fmarusic/pub/marusic&zaucer_2010_clitics_third_position.pdf

Milojević Sheppard, M., & Golden, M. (2000). Imperatives, negation and clitics in slovene. In *Razprave XVII* (pp. 93–109). Ljubljana: Slovenska akademija znanosti in umetnosti, Razred za filološke in literarne vede.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Rickman, P., & Rudanko, J. (2018). *Corpus-based studies on non-finite complements in recent English*. Springer.

Rothstein, S. (1999). Fine-grained structure in the eventuality domain: The semantics of predicative adjective phrases and be. *Natural language semantics*, *7*(4), 347–420. https://doi.org/10.1023/A:1008397810024

Smolej, M. (2004). Členki kot besedilni povezovalci. *Jezik in slovstvo*, *49*(5), 45–57. https://www.dlib.si/details/URN:NBN:SI:DOC-25RXY8HV

Stegovec, A. (2022). Number in Slovenian. In P. Acquaviva & M. Daniel (Eds.), *Number in the world's languages – a comparative handbook* (pp. 187–260). De Gruyter Mouton. https://doi.org/10.1515/9783110622713-008

Toporišič, J. (2000). *Slovenska slovnica*. Obzorja.

Uhlik, M. (2016). Nekatere značilnosti izražanja nujnosti oz. obveznosti v slovenščini in ruščini. *Jezikoslovni zapiski*, *22*(2), 45–59. https://doi.org/10.3986/JZ.22.2.6973

von Fintel, K. (2006). Modality and language. In D. Borchert (Ed.), *Encyclopedia of philosophy* (pp. 20–27). Macmillan Reference.

Žagar, A., Kavaš, M., Robnik-Šikonja, M., Erjavec, T., Fišer, D., Ljubešić, N., ... Hrovat, G. (2022). *Corpus of academic slovene KAS 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1448

Žele, A. (2014). *Slovar slovenskih členkov*. Založba ZRC.

Žele, A. (2015). *Dictionary of Slovenian particles*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1128

SLOVENSKI ČLENEK NI SAMOSTOJNA BESEDNA VRSTA

V prispevku predložimo, da se členki premalo razlikujejo v skladenjskem in semantičnem smislu od prislovov, da bi jih v nasprotju z obstoječimi pristopi v slovenskem opisnem jezikoslovju lahko razčlenjevali kot samostojno besedno vrsto. V prvem delu prispevka predstavimo korpusno raziskavo, s katero pokažemo, da slovenski členki niso robustno opredeljena skladenjska kategorija oz. da so skladenjsko prekrivni z običajnimi prislovi. V članku tudi predložimo, da tvorba vprašalnih stavkov, ki jo sicer uporabljamo za identifikacijo stavčnih členov, konceptualno in empirično ne služi kot ustrezen preizkus za določanje besednih vrst. Nadalje trdimo, da so splošno sprejeti pomenski kriterij tudi neustrezni za razločevanje med členki in prislovi, saj jih pesti vrsta konceptualnih problemov, razloček pa niti ni ustrezno empirično motiviran.

Keywords: členek, prislov, korpusno jezikoslovje, teoretsko jezikoslovje, skladnja

MIĆI PRINC - A LITTLE BOY TEACHING SPEECH TECHNOLOGIES THE CHAKAVIAN DIALECT

Nikola LJUBEŠIĆ,[1,2,3] Peter RUPNIK,[1] Tea PERINČIĆ[4]
[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2]University of Ljubljana, Slovenia
[3]Institute for Contemporary History, Ljubljana, Slovenia
[4]Maritime and History Museum of Croatian Littoral, Rijeka, Croatia

This paper documents our efforts in releasing the printed and audio book of the translation of the famous novel *The Little Prince* into the Chakavian dialect, as a computer-readable , AI-ready dataset, with the textual and the audio components of the two releases now aligned on the level of each written and spoken word. Our motivation for working on this release is multiple. The first one is our wish to preserve the highly valuable and specific content beyond the small editions of the printed and the audio book. With the dataset published in the CLARIN.SI repository, this content is from now on at the fingertips of any interested individual. The second motivation is to make the data available for various artificial-intelligence-related usage scenarios, such as the one we follow upon inside this paper already - adapting the `Whisper-large-v3` open automatic speech recognition model, with decent performance on standard Croatian, to Chakavian dialectal speech. We can happily report that with adapting the model, the word error rate on the selected test data has being reduced to a half, while we managed to remove up to two thirds of the error on character level. We envision many more usages of this dataset beyond the set of experiments we have already performed, both on tasks of artificial intelligence research and application, as well as dialectal research. The third motivation for this release is our hope that this, now highly structured dataset, will be transformed into a digital online edition of this work, allowing individuals beyond the research and technology communities to enjoy the beauty of the message of the little boy in the desert, told through the spectacular prism of the Chakavian dialect.

Keywords: The Little Prince, Chakavian dialect, text and speech dataset, automatic speech recognition

1 INTRODUCTION

We have recently witnessed staggering improvements in processing language in both textual (Zhao et al., 2023) and speech modality (Radford et al., 2022). Regardless of these drastic improvements in performance, they are mostly directed at well-resourced languages in their standardised form, disregarding the dialectal variation (Kantharuban et al., 2023) present in both the textual, but especially the spoken modality of language.

Our language in focus in this paper, the Croatian language, a member of the western group of South Slavic languages, has recently obtained its first open, large, searchable spoken dataset, namely the ParlaSpeech-HR corpus (Ljubešić et al., 2022), based on parliamentary proceedings recordings and transcripts, currently consisting of 3,061 hours of spoken material and linguistically processed transcripts (Ljubešić, Koržinek, & Rupnik, 2024).[1] The two only earlier examples of open spoken datasets of Croatian language that have to be mentioned here, especially important for their pioneering efforts, are the Croatian Adult Spoken Language Corpus (HrAL) (Kuvač Kraljević & Hržica, 2016), 250,000 tokens in size, and the CCCL Croatian corpus of child language (Kovačević, 2002), consisting of recordings and detailed transcriptions of speech of three children.

The only open dialectal dataset of Croatian we are aware of is the recent textual translation of the COPA commonsense reasoning benchmark into the Chakavian dialect of Žminj, part of the DIALECT-COPA benchmark set (Ljubešić et al., 2024). There have, however, not been any open spoken dialectal datasets of Croatian. Here we are changing this, by releasing a small and aesthetically pleasing dataset, the aligned audio and printed book of the translation of *The Little Prince* into various Chakavian micro-dialects - *Mići Princ*. The contributions of this work are the following:

1. We are constructing and releasing via a FAIR (findable, accessible, interoperable, and reusable) repository the first open dataset of dialectal speech of the Croatian language, with speech aligned to its transcripts.

[1] The corpus is searchable through the CLARIN.SI concordancers at https://tinyurl.com/parlaspeech.

2. We are releasing the dataset both in a rich, verbatim format, but also adapted for automatic speeech recognition (ASR) experiments, with instances of up to 30 seconds long, ready to be used for adapting or evaluating various ASR systems.

3. We are showcasing the usefulness even of such a small dataset for modern speech technologies by successfully adapting the Whisper ASR model to the Chakavian dialect.

4. We are releasing the first ASR system capable of processing the Chakavian dialect, lowering the relative word error rate on unseen speakers for around 40%.

5. We are paving the road to a digital online release of the underlying work, which will make the beauty of the Chakavian dialect significantly more accessible to the wider audiences.

This paper is structured as follows: we present the Mići Princ book in Section 2, in Section 3 we describe how we compiled the word-aligned dataset, and discuss steps needed to transform it into an ASR-specific dataset. In Section 4 the encoding of both datasets is explained in detail. In Section 5 our ASR model fine-tuning is presented and the results obtained are commented. We discuss some limitations of our approach in section 6 and wrap up with conclusions (Section 7).

2 ORIGIN OF THE DATA – THE MIĆI PRINC BOOK

Mići Princ (Saint-Exupéry, 2021) is a translation of *Le Petit Prince* into various Chakavian micro-dialects, released both as a printed book and an audio book. Its special distinction is that almost every character in the book uses a different micro-dialect, which was achieved by using numerous translators and voice actors. In total, 16 translators and 16 voice actors were involved in the process, representing the 16 different characters in the audio and the text book.

The audio book spans 113 minutes, which also includes the music that is sometimes used to start or end a chapter. The duration of voiced segments only is 79 minutes. The text portion (after removal of bullet points and newlines and

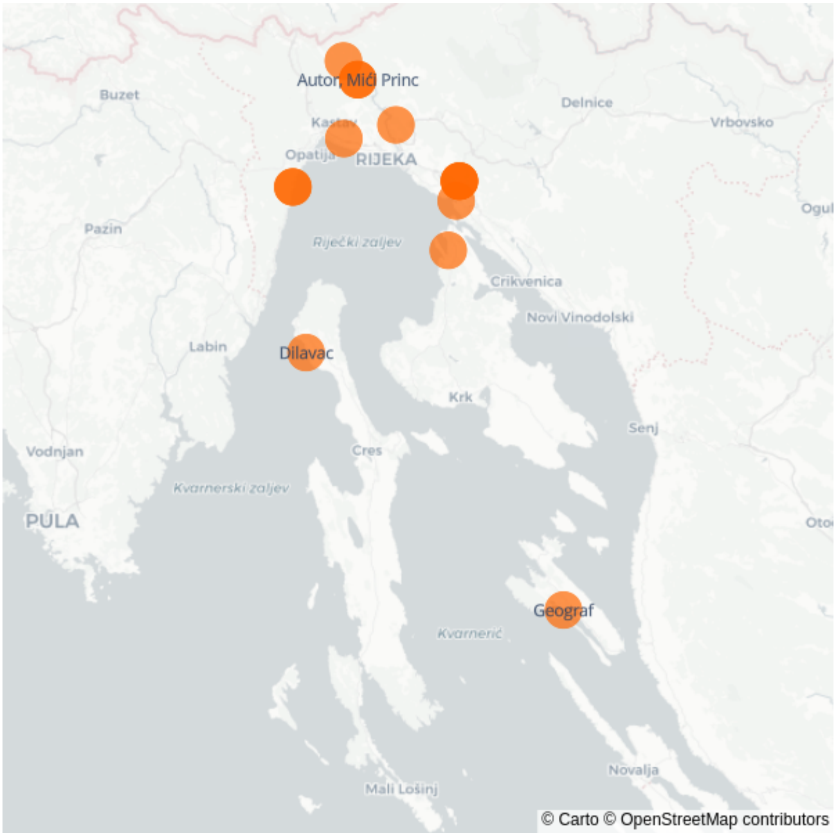


Figure 1: Origins of the voice actors, plotted on a map of northern Kvarner Gulf (Croatia). Speakers with labelled markers were used for model evaluation (see Section 5).

with numerals transcribed to words) is 60129 characters long, which equates to 11591 words and 547 turns.

3 DESCRIPTION OF THE DATA PROCESSING PIPELINE

In this section we are describing the process of transforming the data obtained by the first author of the original book, and the last author of this paper, to obtain the final word-aligned computer-readable dataset, useful both for developing, adapting and evaluating speech technologies, as well as releasing the work in a digital form.

Measure	Quantity
Number of characters	60,129
Number of words	11,591
Number of speaker turns	547
Audio book duration	113 min
ASR dataset duration (speech only)	79 min

The processing involved the following steps:

Chapter-based Segmentation Audio and text were manually segmented into chapters, where chapter 00 denotes the preface and chapters 01 to 28 contain the body section of the book.

Voice Activity Detection Every chapter was analyzed with a voice activity detection model (Bredin et al., 2020; Bredin & Laurent, 2021)[2] which automatically gives spans that contain human speech. Some chapters begin or end with music which would hinder downstream processing. By detecting the parts of the audio where only speech is present, the relevant data can be successfully processed in the downstream.

Trimming For every chapter the audio was trimmed so that only the parts containing speech are preserved.

Diarisation Intervals, where specific speakers are speaking, were identified with a diarisation model[3] (Plaquet & Bredin, 2023; Bredin, 2023).

Exporting For manual corrections and inspection, the trimmed audio and diarised data were exported in EXB format.

Manual interventions EXB files were inspected in Exmaralda Partitur Editor[4]. Any misdiarised turns were manually labeled and automatically corrected afterwards. The speakers identified during diarisation, or added during the manual inspection, were labelled with their characters' names (e.g. Mići Princ, Pisac, Ružica, ...).

[2] https://huggingface.co/pyannote/segmentation
[3] https://huggingface.co/pyannote/speaker-diarization-3.1
[4] https://exmaralda.org/en/partitur-editor-en/

Alignment Texts were normalised (special dialect characters î, ï, ä, â, and ë substituted with analogs from standard Croatian, punctuation characters were removed, numerals were changed to words). The normalised texts were aligned with the audio using Kaldi (Povey et al., 2011), similar to the process recently used to word-align the Slovenian Gos corpus (Verdonik et al., 2024). In rare cases, additional manual interventions were performed on texts to assure successful alignment (e.g. `Exupéry` was changed to `Eksuperi` for alignment and then reverted after successful processing, some transcript errors were also identified during that process and rectified). The resulting aligned data, each word from the transcript having the start and end timestamp in the recording, were encoded in a json and the Exmaralda EXB format.

Data transformation for ASR With the entire Mići Princ diarised, aligned, and manually inspected, the construction of an ASR flavour of the dataset was possible. Since most modern ASR models require relatively short segments, the dataset was re-segmented so that the instances' duration is shorter than 30 seconds. Instances from chapters 13 and 15 were kept aside for constructing the testing subset. They feature two speakers that are very common in the book (Autor and Mići Princ), as well as two additional speakers, Geograf and Dilavac, who do not occur in the rest of the data at all, which allows for examining performance differences on new versus known speakers.

4 FINAL ENCODING OF THE RESULTING DATASET

The final encoding of the constructed dataset was uploaded to the CLARIN.SI FAIR repository [5] (Ljubešić, Rupnik, & Perinčić, 2024). The encoding consists of the following files:

MP.wav.tgz audio files in wav format, one file per chapter

MP.mp3.tgz audio files in mp3 format, one file per instance in the ASR dataset

MP.json.tgz verbatim dataset in JSON format, as described below in Subsection 4.1

[5] http://hdl.handle.net/11356/1765

MP.asr.json.tgz ASR-specific dataset in JSON format, as described below in Subsection 4.2

MP.exb.tgz dataset in EXB format, suitable for viewing in the Exmaralda Partitur Editor

speakers.json a file with speaker metadata information, describing who translated and read parts for a specific character and the provenience of the speaker (name and wikidata link)

4.1 MP.json encoding

In the JSON encoding of the verbatim dataset, containing all the available information, each JSON file covers one chapter. Each entry covers one speaker turn and contain the following attributes:

speaker Character who is speaking in the current turn

text Original text, as it appears in the book, with no alterations except 1. numerals being written with words and 2. parts not pronounced in the audio book omitted.

char_s Character offset start, denoting how many characters from the start of the chapter in textual format the turn starts

char_e Character offset end, i.e., how many characters from the start of the chapter does the turn end in the text version of the chapter

time_s Temporal offset start, i.e., how many seconds after the start of the chapter recording the turn start

time_e Temporal offset end, i.e., how many seconds after the start of the chapter the turn ends in the recording

words A list of key:value pairs for attributes `char_s, char_e, time_s, time_e` for individual words, i.e., information for each word where it is located in the textual version and the audio version of the dataset.

Example entry

In the example entry below, we see a short turn of the *Mići Princ* saying *Prosin vas, narišite mi ovcu*. Futhermore, we know that in the textual version of the chapter we can find this turn between characters at indices 595 and 623. We also know that the spoken form of this turn can be found in the recording between seconds 102.87 and 104.92. Finally, for each of the words, we have similar offset information, for the first word, *Prosin*, the text version being available between character indices 595 and 601, and its pronunciation between seconds 102.87 and 103.34.


```
{   ”char_e”: 623,
    ”char_s”: 595,
    ”speaker”: ”Mići Princ”,
    ”text”: ”Prosin vas, narišite mi ovcu!”,
    ”time_e”: 104.92,
    ”time_s”: 102.87,
    ”words”: [{”char_e”: 601,”char_s”: 595,
             ”time_e”: 103.34,”time_s”: 102.87},
              {”char_e”: 605,”char_s”: 602,
              ”time_e”: 103.59,”time_s”: 103.34}, ...]
}
```


4.2 MP.asr.json encoding

In this section the encoding of the ASR flavour of the dataset is described. It is much simpler than the verbatim encoding described in the previous chapter. Each json covers one chapter. Each entry covers speech in the length of up to 30 seconds. In case of chapters 13 and 15, the testing chapters, it is ensured that each turn is spoken by just one speaker. Each instance contains the following attributes:

audio Name of the audio file in MP.mp3.tgz

text Text of the instance

normalized_text Text without bullet points and newlines, with special characters substituted

speaker Character speaking the instance. This attribute is only present in the testing chapters 13 and 15.

The biggest changes in this ASR-flavoured version of the data are that 1. recording snippets are available in mp3 format for each instance, up to 30 seconds long, 2. there is no alignment information available, 3. text normalization was performed, with bullet points and newlines removed, and accented characters that do not appear in standard Croatian being substituted. With these three changes it is easy to produce instances of short speech and corresponding text snippets, as preferred by the ASR community.

To further boost the visibility of the dataset overall, and especially its application in ASR, this ASR version of the Mići Princ dataset was also published to HuggingFace dataset hub[6], which enables adapting or evaluating speech technologies on this dataset in a few lines of code.

Example entry

In the below example we can observe that each instance has an mp3 file attached (the file name encoding the chapter, as well as time offsets), and the text having a normalised version consisting only of standard Croatian characters.

```
{
    ”audio”: ”MP_13_260.92-261.63.mp3”,
    ”text”: ”I to je sȅ?”,
    ”normalized_text”: ”I to je se?”
    ”speaker”: ”Mići Princ”,
}
```

5 ASR EXPERIMENTS

In this section we are describing our preliminary, but very successful ASR experiments on the dataset described in the previous sections. In the first subsection we are describing the setup of the ASR model fine-tuning procedure, the second subsection describes the overall evaluation of the resulting model, while a

[6] https://huggingface.co/datasets/classla/Mici_Princ

more detailed, speaker-specific analysis of the output is provided in the final subsection.

5.1 Finetuning setup

For the ASR technology we want to adapt to the Chakavian dataset we chose the `Whisper-large-v3`[7] (Radford et al., 2022) model due to its reasonable[8] performance on standard Croatian language. In preliminary experiments, a brief hyper-parameter optimization was performed, in which 80 epochs and learning rate of 1e-5 were chosen as optimal, with effective batch size set to 16.

During fine-tuning with the chosen hyper-parameters, the model was serialized (i.e. saved to disk) every 4 epochs, so that various evaluations could be calculated post-festum as our fine-tuning was progressing.

5.2 Evaluation

The metrics used for evaluation are the two most standard metrics for ASR evaluation: word error rate (WER), which calculates the percentage of mistranscribed words, and character error rate (CER), the percentage of characters that were mistranscribed. Given that the metrics calculate the percentage of errors, lower values show better ASR performance. We use the implementation of the two metrics in the `evaluate`[9] package. Before calculating those metrics, both the model outputs and reference text is lower-cased and stripped of punctuation.

In Figure 2 we present the development of both metrics, overall and by speaker, as the fine-tuning progresses through the 80 selected epochs (each instance in the dataset is used 80 times in fine-tuning). Both metrics exhibit the same profile during the fine-tuning process. The left-most datapoints, at epoch 0, on both plots show performance of 'vanilla' Whisper, before any fine-tuning took place. It is evident that on both metrics and all speakers fine-tuning improved

[7] https://huggingface.co/openai/Whisper-large-v3

[8] In (Samardzic et al., 2024) Whisper-large-v3 is evaluated on a new Croatian ASR dataset, especially adapted to challenges in ASR (e.g. numbers being transcribed as numerals instead of words). In this setting Whisper-large-v3 outperformed other models and achieved character error rate as low as 6.68% and word error rate of 16.18%.

[9] https://pypi.org/project/evaluate/

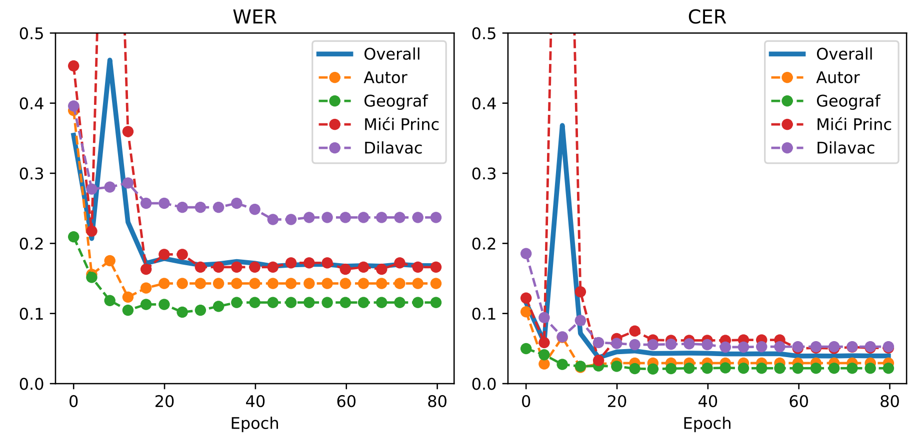


Figure 2: Metrics, achieved by the model during fine-tuning.

results visibly, which is especially important given that two out of four speakers, namely *Dilavac* and *Geograf*, were not seen by the model during fine-tuning.

The 'Overall' curve on Figure 2 shows a very much expected profile. In the first part of the training both metrics get worse, but after a while the model picks up on the specificities of the dialect and the metrics drop. An inspection of the outputs after just a few epochs shows for the drastic deterioriation of the output is due to hallucinations (many repetitions of predicted character sequences) the Whisper model is known for while being adapted with additional data.

Comparing the output of the model with the reference text on the test split showed the model to be well adapted to the dialects present in the dataset, with differences mostly due to:

- incorrect segmentation (e.g. `zvizdami` being transcribed as `zvizda␣mi`)
- regressing to standard language (e.g. `š njimi`→`s njimi`)
- minor differences, where audio could be transcribed in two ways, and in a dialect without orthography both versions could be considered valid (e.g. `diljivaš`↔`dilivaš`)

In one case our model correctly identified a mismatch between the printed and audio versions of the book that we missed to rectify in our processing pipeline. Since the disagreement was indisputably evident from listening to the audio (`priti` versus `arivat`), the ASR dataset was corrected to better reflect the task at hand. This identified mistranscription, performed better by the model than by us, humans, shows: 1. that today's ASR has become very good and 2. that we might have additional minor issues in the data that we will have to improve for the second release of the data. What is important is that in the whole test set, this was the only mistranscription identified, which shows these mistranscriptions to be very infrequent, and thereby the dataset of high quality.

5.3 Speaker-based error analysis

In this subsection we look at the performance of the model as it is being fine-tuned on the basis of each speaker in our testing data, namely two speakers very well represented in the fine-tuning data, *Autor* and *Mići princ*, and two speakers not present in our fine-tuning data, namely *Dilavac* and *Geograf*.

At the beginning of the fine-tuning process, Autor, Mići Princ, and Dilavac perform worse than Geograf on both metrics. After the first serialization at 4 epochs into the training the metrics drop for all speakers, after which performance of some speakers keeps improving, while for other both error rates explode. This phenomenon was investigated by examining the outputs of the models, and wild hallucinations (mostly repetitions a single word at the end of the output) were found to be the root cause for the significantly increased error rates. For some insofar yet unexplained reason, speaker reading lines for Mići Princ seems to be the most affected by this phenomenon. After some additional fine-tuning, these hallucinations become much less frequent, yielding better and more accurate results.

One hypothesis about this speaker-dependent difference is that Mići Princ is the speaker most different to standard pronunciation on the word level (on epoch 0 it has the highest word error rate), and that its need for adaptation gets affected with hallucinations. This speaker is also the most frequent speaker in the test data, which makes the overall metric explode as well.

Not all speakers followed the same metrics profile during training. Speakers present in the training data, Mići Princ and Autor, suffer from the aforementioned performance drop in the beginning of the testing, while new speakers seem not to. Geograf quickly achieves optimal performance, whereas the metrics for Dilavac drop much later in the training. Another possible hypothesis for differences in behaviour is not (just) the initial performance, but also that speakers present in the fine-tuning data are especially prone to hallucinations (over-generation) until the model gets properly fine-tuned.

We have stated two hypotheses on the difference in per-user performance as fine-tuning progresses, testing any of these going outside the scope of this paper, and will therefore have to be inspected more thoroughly in future research.

Comparing the per-speaker metrics with the map from Figure 1 shows no significant geographic trend. Dilavac and Geograf both live very far from the weighed average of training data, which lies just south of Mići Princ and Autor, yet they achieve the worst and the best metrics respectively on the majority of finetuned models. To properly address the search for the existance of geographic trends a much bigger dataset would be needed, where content-based differences would average out.

In Table 1 we present the initial, epoch 0 evaluation (*vanilla*) of the model on both metrics and per each speaker and overall. We compare this Whisper-v3-large-before-adaptation performance with the final performance of the model at epoch 80 (*fine-tuned*). We also report the relative error reduction, which encodes the percentage of errors that were successfully removed from the output of the system with our model adaptation through model fine-tuning.

After 80 epochs the model reaches CER of 3.95% and WER 16.83%, which are very good numbers for the complexity of the underlying problem. What is most important, similar numbers can be observed on the previously unseen speakers, which shows the generality of our adaptation. However, one still has to bear in mind that these are studio-recorded spoken utterances and that transcribing speech in less controlled environments would quite likely be much more challenging.

As expected, the relative error reduction of word error rate on the speakers seen during fine-tuning is higher (63.32% and 63.33%) than for the two unseen speakers (44.75% and 40.15%). This trend, however, does not hold for character error rate, where the overall largest improvement is measure with the *Dilavac* character, which is speaking in a heavy dialect, with a very high character error

speaker	vanilla	finetuned	relative error reduction
all	35.43%	16.83%	52.50%
Autor	38.96%	14.29%	63.32%
Geograf	20.94%	11.57%	44.75%
Mići Princ	45.32%	16.62%	63.33%
Dilavac	39.60%	23.70%	40.15%

(a) Word error rate (WER)

speaker	vanilla	finetuned	relative error reduction
all	11.54%	3.95%	65.77%
Autor	10.24%	2.93%	71.39%
Geograf	4.99%	2.19%	56.11%
Mići Princ	12.21%	5.09%	58.31%
Dilavac	18.55%	5.27%	71.59%

(b) Character error rate (CER)

Table 1: Breakdown of metrics achieved with vanilla (`Whisper-large-v3`) and the finetuned model.

rate of the vanilla model of 18.55%, shrinking with the adaptation to 5.27%, thereby 71.59% of the error on character level being removed.

We can overall report very good results due to adaptation, with 52.5% of error being removed on the level of words, and 65.77% being removed on the level of characters.

The final model was published on HuggingFace model hub[10], hoping to increase visibility of the dataset it has been fine-tuned on, but also to motivate future data-driven projects on this and other dialects.

[10]https://huggingface.co/classla/Whisper-large-v3-mici-princ

6 LIMITATIONS

There is a series of limitations that we want to put forward.

The frequency of special characters (e.g. ë, ï, ä) is low, the most common occurs 29 times out of the total 60,129 characters in the dataset, the least common only appears once. With this in mind we omitted modelling them with our ASR model, which is a limitation of this approach, but with so few occurences of so many special characters we feel the introduction of them would only render the model less reliable.

As with the aforementioned `priti↔arivat` example, it is possible that there are other discrepancies between the audio and the printed version of the book. However, we expect for such potential discrepancies to be very infrequent, given that we were able to find just one in all of the test data.

Our hyper-parameter search was by no means exhaustive, and it is possible that a better fine-tuning setup could exist. Additionally, in our hyper-parameter search we used the same data for training and evaluating as we did for fine-tuning itself, which is not the best practice.

Finally, while our error reductions, as well as overall measured performance is very reassuring, we must stress that this evaluation was performed on acted speech, recorded in a studio setting. Any dialectal speech production out in the wild will surely be much more challenging.

7 CONCLUSIONS

In this paper we have presented our efforts in ensuring the usefulness of two traditional releases, a printed book, and an audio book, both being a translation of *The Little Prince* into Chakavian micro-dialects, beyond these two traditional means of publication.

The first use case for the new dataset, one we have already followed in this paper, is the adaptation of an automatic speech recognition system to the Chakavian dialect. Similar usage can be expected in the future as well, with the dataset becoming both a fine-tuning and an evaluation dataset for future models.

Another use case is the application of data in dialectal research, although the data are acted, so caution is needed for such data usage. However, given the absolute lack of open dialectal data for research, we consider this dataset to improve the data landscape on this front as well.

The third use case that we very much hope for is the preparation of a digital online edition of the translation, where audio and text content could be followed in parallel. Our own experience with the content is that neither the textual nor the audio content is informative enough to delve deep in the rich and aesthetically pleasing content available in the two separate traditional releases.

With the first use case we have illustrated the feasibility of adapting existing tools and frameworks for standard languages to either dialects or other related languages with little resources. In the process two datasets were compiled and published, one following closely the structure of the Mići Princ printed book and audiobook, and the second, compiled with specific ASR applications in mind.

During the ASR system fine-tuning process we noticed interesting disadvantageous transient phenomena, mostly overgeneration of the final text, but after a long enough fine-tuning process, the output is stable with little bias towards new speakers.

We hope that this project will be motivation for further similar endeavours where content right holders will be open for technology-savvy language and speech preservation enthusiasts to encode their work under an open license for the benefit of all involved parties, as well as society as a whole.

ACKNOWLEDGMENTS

This work was partially funded by the programme P6-0411 "Language Resources and Technologies for Slovene", the CLARIN.SI infrastructure, and the project J7-4642 "MEZZANINE - Development of Spoken Language Resources and Speech Technologies for the Slovenian Language", all financed by the Slovenian Research and Innovation Agency (ARIS).

We would very much like to thank the following organisations and individuals: the PEEK&POKE museum for allowing, together with the final author, Tea Perinčić, for these two traditional releases to be made available in a new, AI-friendly format, under a permissive license; John Scott, Marko Simonović and Keith Langston for making the first author aware of texual and audio releases in the Chakavian dialect.

REFERENCES

Bredin, H. (2023). pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. interspeech 2023.*

Bredin, H., & Laurent, A. (2021, August). End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. interspeech 2021.* Brno, Czech Republic.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., ... Gill, M.-P. (2020, May). pyannote.audio: neural building blocks for speaker diarization. In *Icassp 2020, ieee international conference on acoustics, speech, and signal processing.* Barcelona, Spain.

Kantharuban, A., Vulić, I., & Korhonen, A. (2023, December). Quantifying the dialect gap and its correlates across languages. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 7226–7245). Singapore: Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.481 doi: 10.18653/v1/2023.findings-emnlp.481

Kovačević, M. (2002). *Croatian corpus, childes.*

Kuvač Kraljević, J., & Hržica, G. (2016). Croatian adult spoken language corpus (hral). *FLUMINENSIA: časopis za filološka istraživanja*, *28*(2), 87–102.

Ljubešić, N., Koržinek, D., & Rupnik, P. (2024). *Parliamentary spoken corpus of croatian ParlaSpeech-HR 2.0.* http://hdl.handle.net/11356/1914 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., Koržinek, D., Rupnik, P., & Jazbec, I.-P. (2022, June). ParlaSpeech-HR - a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. In D. Fišer, M. Eskevich, J. Lenardič, & F. de Jong (Eds.), *Proceedings of the workshop parlaclarin iii within the 13th language resources and evaluation conference* (pp. 111–116). Marseille, France: European Language Resources Association. https://aclanthology.org/2022.parlaclarin-1.16

Ljubešić, N., Rupnik, P., & Perinčić, T. (2024). *The "mići princ" text and speech dataset of chakavian micro-dialects.* http://hdl.handle.net/11356/1765 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., Galant, N., Benčina, S., Čibej, J., Milosavljević, S., Rupnik, P., & Kuzman, T. (2024). DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to south slavic dialects. In Y. Scherrer, T. Jauhiainen, N. Ljubešić, P. Nakov, J. Tiedemann, & M. Zampieri (Eds.), *Eleventh workshop on nlp for similar languages, varieties and dialects (vardial 2024).* Mexico

City, Mexico: Association for Computational Linguistics.

Plaquet, A., & Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. interspeech 2023.*

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011, December). The kaldi speech recognition toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding.* IEEE Signal Processing Society. (IEEE Catalog No.: CFP11SRW-USB)

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision.* arXiv. https://arxiv.org/abs/2212.04356 doi: 10.48550/ARXIV.2212.04356

Saint-Exupéry, A. d. (2021). *Mići princ* (T. Perinčić, Ed.). Rijeka: Udruga Calculus, Muzej informatike "Peek&Poke". (Prijevod djela: Le petit prince.)

Samardzic, T., Rupnik, P., Starović, M., & Ljubešić, N. (2024, September 19–20). Mak na konac: A multi-reference speech-to-text benchmark for croatian and serbian. In *Proceedings of the language technologies and digital humanities 2024 conference (jt-dh 2024).* Ljubljana, Slovenia.

Verdonik, D., Dobrovoljc, K., Erjavec, T., & Ljubešić, N. (2024). 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)* (pp. 7825–7830).

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... Wen, J.-R. (2023). *A survey of large language models.*

MIĆI PRINC - KAKO JE FANTIČEK V PUŠČAVI RAZPOZNAVALNIK NAUČIL ČAKAVŠČINO

V delu opišemo svoj prispevek pri izdaji tiskane in zvočne oblike prevoda slavnega *Malega Princa* v čakavskem dialektu kot računalniško berljivo podatkovno množico, primerno za uporabo na področju umetne inteligence, z besedilno in zvočno poravnavo vsake izgovorjene in zapisane besede. Pri delu nas je vodilo več vzgibov, prvenstveno želimo te dragocene in zelo specifične vsebine ohraniti, kar smo zagotovili z objavo na FAIR repozitoriju CLARIN.SI, s čimer je Mići Princ odslej vedno na voljo zainterestiranim. Naš drugi cilj je bila priprava podatkov v obliki, primerni za uporabo v različnih aplikacijah umetne intelegence, kot je denimo aplikacija, opisana v tem prispevku: prilagodili smo razpoznavalnik `Whisper-large-v3`, ki na knjižni hrvaščini dosega dobre rezultate, za razpoznavo čakavskega dialekta. Z veseljem lahko poročamo, da smo z učenjem razpoznavalnika na podatkovni množici Mići Princ prvotno stopnjo besednih napak prepolovili, stopnjo napak na znakih pa zmanjšali kar za dve tretjini. Predvidevamo, da bo podobnih in drugih primerov uporab v bodočnosti še več, tako s strani raziskovalcev na področju jezikovnih tehnologij kot v dialektologiji. Kot zadnje pa upamo, da bo ta dobro strukturirana podatkovna množica kmalu tudi transformirana v hibridno digitalno obliko, ki bo sleherniku omogočala vpogled in poslušanje čakavske različice očarljive zgodbe malega dečka v puščavi.

Keywords: Mali princ, Mići Princ, čakavski dialekt, avtomatska razpoznava govora

CLASSLA-STANZA: THE NEXT STEP FOR LINGUISTIC PROCESSING OF SOUTH SLAVIC LANGUAGES

Nikola LJUBEŠIĆ,[1] Luka TERČON,[2] Kaja DOBROVOLJC[1,2]
[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2]Faculty of Arts, University of Ljubljana

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 2.1 release of the pipeline. We also report performance scores produced by the pipeline for different languages and varieties. CLASSLA-Stanza exhibits consistently high performance across all the supported languages and outperforms or expands its parent pipeline Stanza at all the supported tasks. We also present the pipeline's new functionality enabling efficient processing of web data and the reasons that led to its implementation.

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

1 INTRODUCTION

The South Slavic languages make up one of the three major branches of the Slavic language family. Despite their widespread use, many members of this group remain relatively low-resourced and under-represented in the field of natural language processing. Goldhahn et al. (2016) 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 if South Slavic languages are ever to become capable of competing with linguistic giants such as English, steps have already 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[1]), was established

[1] https://www.clarin.si/info/k-centre/

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 (Ljubešić et al., 2022). The efforts of the knowledge centre gave rise to the CLASSLA-Stanza[2] pipeline for linguistic processing, which arose as a fork of the Stanza neural pipeline (Qi et al., 2020). CLASSLA-Stanza was created with the aim of providing state-of-the-art automatic linguistic processing for South Slavic languages (Ljubešić & Dobrovoljc, 2019), and currently supports Slovenian, Croatian, Serbian, Macedonian, and Bulgarian . Additionally, Slovenian, Croatian, and Serbian have support for both standard and non-standard, Internet varieties. In comparison to its Stanza parent pipeline, CLASSLA-Stanza expands to cover the standard Macedonian language, as well as the non-standard, Internet varieties of Slovenian, Croatian and Serbian. Beside the expanded coverage of languages and varieties, CLASSLA-Stanza shows improvements in performance on 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 2.1 release. 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 2. Afterwards, section 3 introduces the datasets used for training the models. Section 4 then gives a general description of the model training process, which is followed by an analysis of the results produced by the newly-trained models in section 5.

At present, the CLASSLA-Stanza annotation tool supports a total of six tasks: tokenization, morphosyntactic annotation, lemmatization, dependency parsing, semantic role labeling, and named-entity recognition. Tokenization is handled by one of two external rule-based tokenizers included in CLASSLA-Stanza, either the Obeliks tokenizer[3] for standard Slovenian (Grčar et al., 2012) or the ReLDI

[2] https://github.com/clarinsi/classla
[3] https://github.com/clarinsi/obeliks

tokenizer[4] for non-standard Slovenian and all other languages (Samardžić et al., 2015). While the basic tasks of tokenization, morphosyntactic annotation, lemmatization and dependency parsing are covered at least for some languages in the parent Stanza pipeline, semantic role labeling and named entity recognition for South Slavic languages are available only in CLASSLA-Stanza.

The current version of the models was trained on data that are 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 (de Marneffe et al., 2021) and will henceforth be referred to as UPOS and UFeats—and the MULTEXT-East V6 specifications for morphosyntactic annotation (Erjavec, 2012), which are implemented as the language-specific XPOS tags in the CoNLL-U file format[5], 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 (Erjavec, Fišer, et al., 2010). Additionally, the annotation schema described in Krek et al. (2016) was used for semantic role label annotation.

The outline of the model training process given in this paper describes all six tasks supported by CLASSLA-Stanza, however it must be noted that not all tasks are available for every supported language and variety. Dependency parsing has dedicated models for the standard variety of every language except Macedonian. Named entity recognition is also not supported for Macedonian. Processing of the non-standard variety is available only for Slovenian, Croatian and Serbian, while it is not available for Macedonian and Bulgarian. Semantic role labeling 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 (Ljubešić & Samardžić, 2023). Table 1 provides an overview of every language variety and the tasks it supports.

[4] https://github.com/clarinsi/reldi-tokeniser
[5] https://universaldependencies.org/format.html

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 labeling.

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

2 DIFFERENCES BETWEEN CLASSLA-STANZA AND STANZA

The Stanza neural pipeline is centered around a bidirectional long short-term memory (Bi-LSTM) network architecture (Qi et al., 2020). 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 in the end 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 learned 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 learned by the model during the training process. Due to this drawback, 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 (Grčar et al., 2012) and the ReLDI tokenizer for non-standard Slovenian and all other languages (Samardžić et al., 2015).[6]

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 (Ljubešić & Dobrovoljc, 2019). 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 an external lexicon usage 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 additionally exploits 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 that contain the full morphosynctactic information, while Stanza uses 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 available than for dependency parsing, which is exploited by CLASSLA-Stanza. Thus, for example, instead of using only the 210

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

thousand tokens of data that are used for training the dependency parser, the latest set of standard Croatian models in CLASSLA-Stanza includes mophosyntactic tagging and lemmatization models which were trained on additional 290 thousand tokens, which were manually annotated only on these two levels of annotation.

CLASSLA-Stanza also has a special way of handling "closed-class" words. Closed-class control is a feature of the tokenizers and ensures that punctuation and symbols are assigned appropriate morphosyntactic tags and lemmas. It also restricts the set of possible tokens that can be assigned morphosyntactic tags and lemmas for punctuation and symbols to only those tokens that are defined as such in the tokenizer. In addition to punctuation and symbols, the Slovenian package also includes closed-class control for pronouns, determiners, adpositions, 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 for any token to be labeled with a closed class label if this token was not defined as such in the lexicon.[7]

The Stanza pipeline expects pretrained word embeddings as input. While it uses embedding collections based on Wikipedia data, CLASSLA-Stanza does the extra mile by using the CLARIN.SI embeddings (Terčon et al., 2023; Terčon & Ljubešić, 2023b, 2023b, 2023d, 2023c, 2023a), which are skipgram-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 multiple times larger text collections than Wikipedia, obtained through web crawling (Bañón et al., 2022), which ensures drastically more diverse word embeddings and thereby also better unseen word handling.

When working with Slovenian, Croatian, or Serbian, the pipeline can be set to any of three predetermined settings, which are used for processing different varieties of the same language. These settings are called *types* and can be either *standard*, *nonstandard*, or *web*. The processing types determine which model is used on every level of annotation (either standard or nonstandard) and are all associated with their respective language varieties: the *standard* type is used

[7] 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.

for processing standard language, the *nonstandard* type is used for processing nonstandard Internet language, and the *web* type is used for processing texts obtained from the web. The reasons for introducing a separate processing type for web texts are described in section 5.2. Below is an overview showing which model is used on every layer for every type:

Table 2: Overview of processing types in CLASSLA-Stanza.

Processing type	Tokenizer	Morphosyntactic tagger	Lemmatizer	dependency parser
standard	standard	standard	standard	standard
nonstandard	nonstandard	nonstandard	nonstandard	standard
web	standard	nonstandard	nonstandard	standard

The reason why the nonstandard and the web processing type use the standard dependency parsing model is primarily the lack of training data for training a model beyond standard text. The lack of motivation for building a dataset for parsing non-standard text lies in the fact that the parsing model has upstream lemma and morphosyntactic information at its disposal, 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 on the SloBENCH evaluation dataset. SloBENCH[8] (Žitnik & Dragar, 2021) is a platform for benchmarking various natural language processing tasks for Slovenian, which includes also a dataset for evaluating the tasks supported by CLASSLA-Stanza. The performance scores are presented in the form of micro-F1 scores, while the relative error reduction between the scores of the pipelines is presented in percentages.

3 DATASETS

The latest models included in the 2.1 release of CLASSLA-Stanza were trained on a variety of datasets in five different languages: Slovenian, Croatian, Serbian, Macedonian, and Bulgarian. Slovenian, Croatian, and Serbian were all associated with two training datasets—one consisting of standard-language texts and one consisting of non-standard texts, while Bulgarian and Macedonian only had a standard-language training dataset available.

[8] https://slobench.cjvt.si/

Table 3: Comparison of performance on the SloBENCH evaluation dataset by both pipelines. Metrics are micro-F1 scores. Downstream tasks use upstream predictions, not gold labels.

Task	Stanza	CLASSLA-Stanza	Rel. 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%

Slovenian standard language models were trained using the SUK training corpus (Arhar Holdt et al., 2022). 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 labeling annotations. The corpus is a continuation of the ssj500k Slovene training corpus (Krek et al., 2021). Non-standard models were trained on a combination of the standard training corpus and the non-standard Janes-Tag training corpus (Lenardič et al., 2022), 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.

Croatian standard language models were trained on the hr500k training corpus (Ljubešić & Samardžić, 2023), 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 labeling annotations. Croatian non-standard models were trained on a combination of the standard training corpus and the non-standard ReLDI-NormTagNER-hr training corpus (Ljubešić et al., 2023a). The ReLDI-NormTagNER-hr corpus contains about 90 thousand tokens of non-standard 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 (Batanović et al., 2023), 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 non-standard models were trained, similar to the previously introduced languages, on a combination of the standard dataset and the non-standard ReLDI-NormTagNER-sr training corpus (Ljubešić et al., 2023b). 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 (Erjavec, Barbu, et al., 2010) 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 (Ljubešić & Stojanovska, 2023). At the time of writing this paper, only the SETimes.MK corpus has been made publicly available, while the "1984" corpus is still awaiting to being published by its authors. Both corpora are manually annotated on the levels of tokenization, sentence segmentation, morphosyntactic tagging, and lemmatization. The combining of the corpus 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 only the training data split 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 (Osenova & Simov, 2015), 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 labeling.

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
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

4 MODEL TRAINING PROCESS

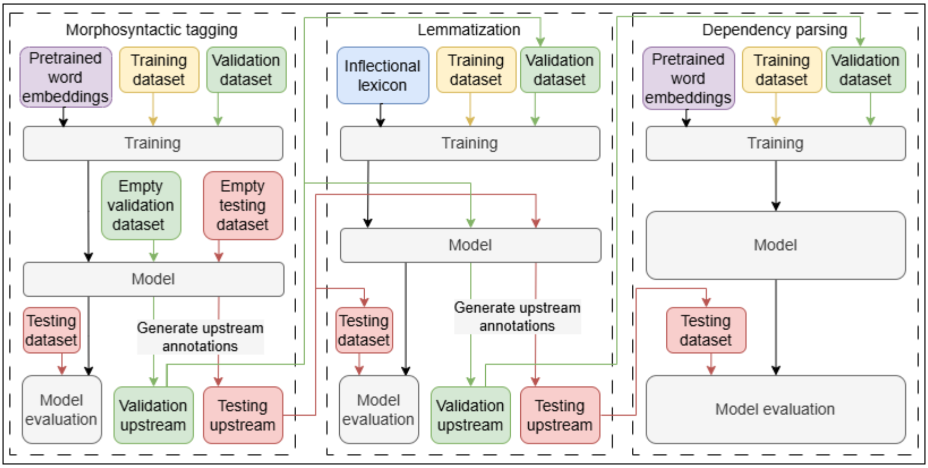
In this section, the model training process is described in detail. Only a descriptive account of the process is provided here. For a list of the specific commands and oversampling scripts used, refer to the GitHub repository of the training procedure.[9]

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 refer to the CLASSLA-Stanza technical report, a longer and older version of this paper available on arXiv (Terčon & Ljubešić, 2023). The language-specific steps were necessary due to some features and levels of annotation (semantic role labeling, oversampling of the training data, etc.) being unique to only certain languages, while all languages at the same time share the common steps described below.

An 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.

[9] https://github.com/clarinsi/classla-training

Figure 1: Diagram of the basic model training process for standard morphosyntactic tagging, lemmatization and dependency parsing models.



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 pretokenized 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 of cross-validation method such as jackknifing (splitting 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 training, 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, so as 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 one of the inputs. 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 layer model.

The dependency parser module 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.[10] The parser was run in training mode using the training and validation datasets[11] as inputs along with the pretrained word embeddings. After training, the parser was run in predict mode to obtain evaluation results.

For this latest release of CLASSLA-Stanza, no new models for named entity recognition were trained. However, the process for training models for named-entity recognition is quite similar to the other tasks. The tagger trainer for this task accepts pretrained word embeddings and training and validation datasets

[10] In comparison to UD, 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 specific applications.

[11] For most languages, only a portion of the original datasets contained dependency parsing annotations. In these cases, a separate set of training, validation, and test datasets consisting of only this portion of the original data had to be extracted.

as inputs. After training, the named entity recognition tagger can be run in predict mode to obtain evaluation results.

The non-standard models were trained using the same process as the standard models, with a few exceptions. Firstly, no syntactic dependency annotations are present in the non-standard datasets. As a result, no non-standard dependency parsing models were trained.

Before training the non-standard models, approximately 20% of diacritics were removed from the training datasets in order to ensure that the models will learn to effectively handle dediacritized forms, which occur prominently in online communication.

It is important to note that the non-standard models were regularly trained on a mixture of standard and non-standard data for best possible performance, while still informing models of non-standard linguistic features. For that reason, non-standard training data were regularly oversampled so that their combination with the standard data would not make non-standard data much less represented, which would hinder learning the non-standard linguistic features.

5 MODEL PERFORMANCE ANALYSIS

We know, as noted in section 2, that CLASSLA-Stanza significantly outperforms Stanza on the Slovenian benchmark, with 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 perform in this section a series of additional performance analyses. In Section 5.1 we give a detailed rundown of the performance of the models for each UPOS and each UD label for each language. In Section 5.2 we continue with a more qualitative investigation of the performance of the models on web-specific data.

5.1 Model Performance on UPOS and UD Labels

To obtain a sense of which specific categories a model struggles with and which ones it handles with particular ease, 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. 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 either "st" for *standard* or "nonst" for *non-standard*.

UPOS tag	Accuracy								
	sl-st	sl-nonst	hr-st	hr-nonst	sr-st	sr-nonst	mk-st	bg-st	Average
ADJ	99.31	90.71	97.93	92.27	99.27	94.58	97.74	98.28	96.26
ADP	99.90	98.54	99.96	99.82	100.00	99.84	99.75	99.92	99.72
ADV	95.98	91.89	95.35	91.59	95.42	87.93	95.14	97.60	93.86
AUX	98.62	96.31	99.60	99.59	100.00	98.81	99.50	92.75	98.15
CCONJ	98.01	97.03	96.53	97.21	98.95	97.21	97.94	97.87	97.59
DET	99.29	93.29	95.68	94.08	98.88	96.74	100.00	87.79	95.72
INTJ	80.00	75.82	71.43	90.22	n/a	87.65	71.43	47.58	74.88
NOUN	98.88	93.75	98.33	93.98	99.23	97.66	99.55	98.53	97.49
NUM	99.74	98.41	98.87	100.00	98.71	100.00	100.00	98.17	99.24
PART	99.46	95.12	85.16	90.64	94.12	89.39	90.16	79.94	90.50
PRON	99.47	97.25	98.68	98.19	97.64	98.47	98.84	99.15	98.46
PROPN	98.71	78.23	93.65	77.81	97.31	83.68	97.97	98.14	90.69
PUNCT	100.00	99.79	100.00	99.73	100.00	99.82	100.00	100.00	99.92
SCONJ	99.78	97.99	95.72	94.79	99.52	98.25	94.70	99.61	97.55
SYM	100.00	99.85	90.91	99.10	100.00	99.38	n/a	n/a	98.21
VERB	97.05	94.12	99.30	97.84	99.18	98.76	99.74	96.79	97.85
X	59.13	75.67	77.15	80.10	43.33	62.86	n/a	0.00	56.89

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 2. 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 abbreviations, 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 all UD relations.

UD relation	Accuracy				
	sl	hr	sr	bg	Average
punct	100.00	100.00	100.00	99.91	99.98
amod	98.61	95.97	97.38	98.66	97.66
case	99.63	99.32	99.21	99.86	99.51
nmod	92.74	91.22	90.99	91.49	91.61
nsubj	90.49	93.39	94.30	91.10	92.32
obl	91.99	85.31	87.24	77.17	85.43
conj	92.51	90.92	93.06	93.95	92.61
root	93.14	94.98	95.77	95.97	94.97
obj	93.33	82.84	91.39	90.18	89.44
aux	99.48	97.88	97.57	90.46	96.35
cc	97.83	97.63	97.96	99.14	98.14
advmod	96.74	93.58	91.82	97.91	95.01

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 (Dobrovoljc et al., 2022). 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 also a likely reason for the performance drops of the **obl** and **obj** relations in other languages as well.

5.2 Model Performance on Web Data

The model evaluations described in the previous subsection provide a good summary of how well the CLASSLA-Stanza pipeline performs on both purely standard and purely non-standard data. However, modern corpus construction techniques—especially for low-resource languages—often rely on crawling data from online conversations, articles, blogs, etc. (Goldhahn et al., 2016), which typically consists 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 (Bañón et al., 2023b, 2023a). In preparation for the annotation process, a short test was conducted with the goal of determining which of the two sets of models—the standard or the non-standard—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 non-standard models. The two outputs were then compared and a qualitative analysis of the differences was conducted.

Quite a few of the analyzed differences in the model outputs were connected to the processes of sentence segmentation and tokenization. In the CLASSLA-Stanza annotation pipeline, both of these processes are controlled by the tokenizer. As stated in section 2, the pipeline uses two different tokenizers depending on the language and the annotation type used. [12] The analysis showed that sentence segmentation was performed much more accurately by Obeliks and the standard mode of the ReLDI tokenizer. The non-standard 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 non-standard tokenizer very consistently produces a new sentence after periods, question marks, exclamation marks, and other punctuation, even when

[12] The ReLDI tokenizer can be used in two different settings: standard and non-standard. The Obeliks tokenizer, on the other hand, only supports tokenization of standard text.

these characters do not signify the end of a segment. 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 “
```

Besides sentence segmentation issues, the standard models also performed better than the non-standard models when assigning certain types of grammatical features, such as with disambiguating between the UD part-of-speech labels AUX and VERB for the verb *biti* (Eng. “to be”). However, the difference between the two model outputs for these grammatical features was not as noticeable as on the levels of tokenization and sentence segmentation.

The non-standard models, on the other hand, handled non-standard 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 lemmatizer output for the word form "hoce" (which corresponds to *hoče* in standard Slovene (Eng. "he/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
```

Non-standard 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 non-standard models for all subsequent levels of annotation. In light of this, a new *web* type was implemented for the CLASSLA-Stanza pipeline. This new type combines the standard tokenizer and non-standard models for the other layers in a single package and is intended specifically for the annotation of texts originating on the Internet.

6 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 for a more detailed description of the training process for each language the technical documentation (Terčon & Ljubešić, 2023) should be consulted. We also presented per-label performance scores for UPOS labels from standard and non-standard models, and most frequent UD labels from standard models.

CLASSLA-Stanza gives consistent results across all supported languages and outperforms the Stanza pipeline on all supported NLP tasks, as illustrated in sections 2 and 4. However, overall low accuracies are still seen for infrequent labels and pairs of labels that are not so 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. 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 type implemented within CLASSLA-Stanza.

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, such as, for instance, semantic role labeling, which is currently only available for Slovenian, 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 (Nguyen et al., 2021), which was shown to outperform Stanza over a large number of languages and datasets.

ACKNOWLEDGMENTS

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), financed by the Slovenian Research Agency, the MEZZANINE project (Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language, project ID: J7-4642) and the SPOT project (A Treebank-Driven Approach to the Study of Spoken Slovenian, Z6-4617), financed by the Slovenian Research Agency, and the CLARIN.SI research infrastructure.

REFERENCES

Arhar Holdt, Š., Krek, S., Dobrovoljc, K., Erjavec, T., Gantar, P., Čibej, J., ... Zajc, A. (2022). *Training corpus SUK 1.0.* http://hdl.handle.net/11356/1747 (Slovenian language resource repository CLARIN.SI)

Bañón, M., Chichirau, M., Esplà-Gomis, M., Forcada, M. L., Galiano-Jiménez, A., García-Romero, C., ... Zaragoza-Bernabeu, J. (2023a). *Croatian web corpus MaCoCu-hr 2.0.* http://hdl.handle.net/11356/1806 (Slovenian language resource repository CLARIN.SI)

Bañón, M., Chichirau, M., Esplà-Gomis, M., Forcada, M. L., Galiano-Jiménez, A., García-Romero, C., ... Zaragoza-Bernabeu, J. (2023b). *Slovene web corpus MaCoCu-sl 2.0.* http://hdl.handle.net/11356/1795 (Slovenian language resource repository CLARIN.SI)

Bañón, M., Esplà-Gomis, M., Forcada, M., García-Romero, C., Kuzman, T., Ljubešić, N., ... Zaragoza, J. (2022). MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In L. Macken et al. (Eds.), *Eamt 2022 - proceedings of the 23rd annual conference of the european association for machine translation* (pp. 303–304). European Association for Machine Translation.

Batanović, V., Ljubešić, N., Samardžić, T., & Erjavec, T. (2023). *Serbian linguistic training corpus SETimes.SR 2.0.* http://hdl.handle.net/11356/1843 (Slovenian language resource repository CLARIN.SI)

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021, July). Universal Dependencies. *Computational Linguistics*, *47*(2), 255-308. https://doi.org/10

.1162/coli_a_00402 doi:10.1162/coli_a_00402

Dobrovoljc, K., Terčon, L., & Ljubešić, N. (2022). Universal Dependencies za slovenščino: nadgradnja smernic, učnih podatkov in razčlenjevalnega modela [Universal Dependencies for Slovenian: An Upgrade to the Guidelines, Annotated Data and Parsing Model]. In D. Fišer & T. Erjavec (Eds.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference [language technology and digital humanities: Conference proceedings]* (p. 30–39). Inštitut za novejšo zgodovino [Institute of Contemporary History]. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf

Erjavec, T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, *46*(1), 131–142. Retrieved 2023-06-21, from http://www.jstor.org/stable/41486069

Erjavec, T., Barbu, A.-M., Derzhanski, I., Dimitrova, L., Garabík, R., Ide, N., ... Zdravkova, K. (2010). *MULTEXT-East "1984" annotated corpus 4.0.* http://hdl.handle.net/11356/1043 (Slovenian language resource repository CLARIN.SI)

Erjavec, T., Fišer, D., Krek, S., & Ledinek, N. (2010, May). The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf

Goldhahn, D., Janicki, M., & Quasthoff, U. (2016). Corpus collection for under-resourced languages with more than one million speakers. In *Workshop on Collaboration and Computing for Under-Resourced Languages (CCURL), LREC.*

Grčar, M., Krek, S., & Dobrovoljc, K. (2012, October). Obeliks: statisticni oblikoskladenjski oznacevalnik in lematizator za slovenski jezik [Obeliks: Statistical Morphosyntactic Tagger and Lemmatizer for Slovene]. In J. PBI G. T. Erjavec (Ed.), *Proceedings of the eighth language technologies conference.* Ljubljana: Jožef Stefan Institute.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., ... Zajc, A. (2021). *Training corpus ssj500k 2.3.* http://hdl.handle.net/11356/1434 (Slovenian language resource repository CLARIN.SI)

Krek, S., Gantar, P., Dobrovoljc, K., & Škrjanec, I. (2016). Označevanje udeleženskih vlog v učnem korpusu za slovenščino [Semantic Role Labeling in the Training Corpus for Slovene]. In (p. 106–110). Znanstvena založba Filozofske fakultete [Ljubljana University Press, Faculty of Arts]. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Krek-et-al_Oznacevanje-udelezenskih-vlog-v-ucnem-korpusu-za-slovenscino.pdf

Lenardič, J., Čibej, J., Arhar Holdt, Š., Erjavec, T., Fišer, D., Ljubešić, N., ... Dobrovoljc, K. (2022). *CMC training corpus Janes-Tag 3.0.* http://hdl.handle.net/11356/1732 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., & Dobrovoljc, K. (2019, August). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th workshop on balto-slavic natural language processing* (pp. 29–34). Florence, Italy: Association for Computational Linguistics. https://aclanthology.org/W19-3704 doi: doi:10.18653/v1/W19-3704

Ljubešić, N., Erjavec, T., Batanović, V., Miličević, M., & Samardžić, T. (2023a). *Croatian Twitter training corpus ReLDI-NormTagNER-hr 3.0.* http://hdl.handle.net/11356/1793 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., Erjavec, T., Batanović, V., Miličević, M., & Samardžić, T. (2023b). *Serbian Twitter training corpus ReLDI-NormTagNER-sr 3.0.* http://hdl.handle.net/11356/1794 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., & Samardžić, T. (2023). *Croatian linguistic training corpus hr500k 2.0.* http://hdl.handle.net/11356/1792 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., & Stojanovska, B. (2023). *Macedonian linguistic training corpus SETimes.MK 0.1.* http://hdl.handle.net/11356/1886 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., Erjavec, T., Petrović, M. M., & Samardžić, T. (2022). Together We Are Stronger: Bootstrapping Language Technology Infrastructure for South Slavic Languages with CLARIN.SI. In D. Fišer & A. Witt (Eds.), *Clarin. the infrastructure for language resources.* (pp. 429–456). Berlin, Boston: De Gruyter. Retrieved 2023-06-21, from https://doi.org/10.1515/9783110767377-017 doi: doi:10.1515/9783110767377-017

Nguyen, M. V., Lai, V., Veyseh, A. P. B., & Nguyen, T. H. (2021). 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.*

Osenova, P., & Simov, K. (2015, September). Universalizing BulTreeBank: a Linguistic Tale about Glocalization. In *The 5th workshop on Balto-Slavic natural language processing* (pp. 81–89). Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA. https://aclanthology.org/W15-5313

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations.* https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Samardžić, T., Ljubešić, N., & Miličević, M. (2015, September). Regional Linguistic Data Initiative (ReLDI). In *The 5th workshop on Balto-Slavic natural language processing* (pp. 40–42). Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA.

https://aclanthology.org/W15-5306

Terčon, L., & Ljubešić, N. (2023a). *Word embeddings CLARIN.SI-embed.bg 1.0.* http://hdl.handle.net/11356/1796 (Slovenian language resource repository CLARIN.SI)

Terčon, L., & Ljubešić, N. (2023b). *Word embeddings CLARIN.SI-embed.hr 2.0.* http://hdl.handle.net/11356/1790 (Slovenian language resource repository CLARIN.SI)

Terčon, L., & Ljubešić, N. (2023c). *Word embeddings CLARIN.SI-embed.mk 2.0.* http://hdl.handle.net/11356/1788 (Slovenian language resource repository CLARIN.SI)

Terčon, L., & Ljubešić, N. (2023d). *Word embeddings CLARIN.SI-embed.sr 2.0.* http://hdl.handle.net/11356/1789 (Slovenian language resource repository CLARIN.SI)

Terčon, L., Ljubešić, N., & Erjavec, T. (2023). *Word embeddings CLARIN.SI-embed.sl 2.0.* http://hdl.handle.net/11356/1791 (Slovenian language resource repository CLARIN.SI)

Terčon, L., & Ljubešić, N. (2023). *CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.* https://doi.org/10.48550/arXiv.2308.04255

Žitnik, S., & Dragar, F. (2021). *SloBENCH evaluation framework.* http://hdl.handle.net/11356/1469 (Slovenian language resource repository CLARIN.SI)

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. Opisujemo vse glavne izboljšave, ki jih prinaša CLASSLA-Stanza v primerjavi s Stanzo in podamo podroben opis postopka učenja modelov v različici 2.1, 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 nalogah. Predstavimo tudi novo funkcijo cevovoda, ki omogoča učinkovito procesiranje spletnih besedil, in razloge za njeno implementacijo.

Keywords: južnoslovanski jeziki, avtomatsko procesiranje jezika, označevalni cevovod, jezikovno označevanje

COMMUNICATIVE INTENT DIVERGENCE OF DISCOURSE MARKERS IN SIMULTANEOUSLY INTERPRETED SPEECH

Simona MAJHENIČ
Faculty of Electrical Engineering and Computer Science, University of Maribor

The present paper examines the divergence in the communicative intent of discourse markers in simultaneous interpretations from English into Slovene in the European Parliament. Specifically, it explores the factors of the original speaker's delivery rate and emotive intensity, and the effects of these factors on the share of equivalent translations, i.e., the translations of discourse markers with the same communicative intent. Moreover, the paper presents a comprehensive translation type classification tailored specifically to the use of English and Slovene discourse markers in simultaneous interpretation. Our classification is grounded in empirical data containing approximately 47,000 tokens, and covers the encountered discourse marker translation types. The results, based on the corpus TolAnSi, indicate that neither the delivery rate nor the emotive intensity level of the speaker affects the share of equivalent translations significantly.

Keywords: discourse marker, simultaneous interpreting, delivery rate, emotive intensity, translation type classification

1 INTRODUCTION

In communication, we do not rely only on propositional content to convey our message. Non-propositional content, or metadiscourse (Hyland, 2005), plays an equally significant role in communication (Hyland, 2005; Mauranen, 2023), and helps us reveal "our intended meanings and communicative intentions" (Mauranen, 2023, p. 1). As a subgroup of metadiscourse (Aijmer, 1996; Maschler, 2009), discourse markers can structure discourse (Schiffrin, 1987; Redeker, 2006), oil the wheels of communication (Müller, 2005), serve as indicators on how to interpret a message (Schiffrin, 1987; Redeker, 2006), and express our attitude towards others (Andersen, 1998; Maschler, 2009).

One would, therefore, expect discourse markers (DMs) to be indispensable in settings where convincing others of one's argument is pivotal. An example of such a setting is heated debates in the European Parliament, where each Member of the European Parliament is permitted to speak in their mother tongue due to the European Union's multilingualism policy.

To make use of this privilege, simultaneous interpreters are needed to provide live translations. These translations must be as close to the original as possible, so that the listeners relying on the simultaneous interpretation are able to understand the message, and, if need be, respond accordingly. Nevertheless, simultaneous interpreting is a cognitively demanding task, as the interpreter must provide an accurate translation within a very limited timeframe. With this in mind, one might assume interpreters devote more attention to the propositional content and less to the metadiscourse, including DMs.

Consider the following excerpt of a response in the European Parliament within a debate on the Brexit referendum: *"thank you | so | we have to protect ourselves from these evil Russians who are controlling the internet do we"*[1] (Etheridge, 2019) and its rendition into Slovene: *"hvala lepa | seveda | zaščititi zavarovati se moramo pred temi zlobnimi Rusi"* (anonymous interpreter) [thank you | of course | we have to safeguard and protect ourselves from these evil Russians]. The DM "so" loses its concluding function, which reflects the speaker's summarisation of previous responses, but gains an affirmative or agreeing function in the translation ("of course"), which is not present in the source language (SL). Conversely, the DM "do we" which serves as a provocation or negative emphasis, is omitted completely in the interpretation. The communicative intent of building on the message of the previous responses, as well as provoking, is consequently lost in the translation.

The present paper examines such divergences in the communicative intent of DMs, and strives to hone in on the factors contributing to them. In particular, we look into how the SL delivery rate and the speaker's emotive intensity affect the simultaneous interpretation of English DMs into Slovene. To guide our investigation, the following research questions have been formulated:

[1] The transcription is divided into segments which are separated by the pipe character. Translations for the Slovene interpretations are provided in square brackets.

- RQ1: Does the SL delivery rate affect the share of the DMs' equivalent translation types?
- RQ2: Does the level of emotive intensity in the SL affect the share of the DMs' equivalent translation types?

2 RELATED WORKS

2.1 Discourse markers

Despite being a popular topic of research, to date, there is no widely accepted definition of DMs (Maschler & Schiffrin, 2015; Fedriani & Sansò, 2017). Instead, consensus is found mostly by listing their properties, which include non-truth conditionality, syntactical detachability (Schourup, 1999; Redeker, 2006), multifunctionality (Schiffrin, 2001), as well as having a metalingual interpretation (Maschler, 2009; Maschler & Schiffrin, 2015). They are one- or multi-word units (Schourup, 1999; Redeker, 2006) encompassing different parts of speech (Schourup, 1999; Fischer, 2006). Similar to Schiffrin's (1987, p. 31) canonical definition of DMs being "sequentially dependent elements which bracket units of talk", Maschler (2009, p. 6) describes them as "metalingual utterances occurring at conversational action boundaries, or frame shifts", whereby frame shifts can be major, such as at the beginning of a new story, or subtle, such as when introducing a side comment.

DMs are generally divided into those pertaining to the ideational or the textual domain, the rhetorical, the interpersonal or the interactional, as well as the structural or the sequential domain (cf. Crible, 2017; Crible & Degand, 2019). These domains are then specified further into thirty functions (Crible, 2017), with which the communicative intent of a DM can be specified further. Maschler (2009), however, includes an additional domain, or "realm", as she terms it, namely, the domain pertaining to the cognitive processes of the participants, which include the functions of realising the need to rephrase one's utterance, information processing, or realising new information. Consequently, she devised a classification of DMs that distinguishes the domain of textual, interpersonal and cognitive DMs. Especially for settings where the cognitive load is very high, such as simultaneous interpreting, the inclusion of the cognitive domain is crucial, as it can hint at possible explanations of translation choices.

The translation of DMs is, per their definition, a highly complex task (Aijmer & Simon-Vandenbergen, 2003), as DMs are characteristically multifunctional. Consequently, they are often omitted in written translations (Aijmer, 2007). This raises the question of how simultaneous interpreters deal with DMs, where the task of providing an accurate translation is even more demanding.

2.2 Simultaneous interpreting of DMs

Simultaneous interpreting is a cognitively taxing task, as the interpreter has to listen to the original speech, translate, utter the translation, and, at the same time, listen to the speaker's next utterance while monitoring one's own output. It is, therefore, a continuous process of listening, comprehending, translating and speaking. At the same time, the interpreter is committed to providing a translation that is as close as possible to the original message. Omitting DMs or changing their meaning would, therefore, breach this principle.

Nevertheless, previous research on the interpretation of DMs has shown that, in legal settings, interpreters systematically omit DMs (Hale, 2004; Blakemore & Gallai, 2014). A study on the simultaneous interpretation of the DM "indeed" into Polish in the European Parliament revealed that the DM is omitted in one-third of the cases, and, when translated, in two-thirds of the cases, the translations are non-equivalent (e.g., generalisations, particularisations, or functional diversions) (Rozumko, 2021). The analysis of the simultaneous interpretation of causal and concessive connective items[2] in the European Parliament has shown that functionally equivalent translations represent only 40% to 45% of the provided translations (Defrancq et al., 2015).

The above findings provoked the exploration of the causes leading interpreters to omit or change the communicative intent of the DMs functionally.

[2] In the present study, connective items or connectives (e.g., because, but, so, however) are considered a part of DMs provided they fulfill the requirements of discourse markers as described in Section 3.4.

2.3 Delivery rate

As noted by past research (Defrancq et al., 2015; Rozumko, 2021), the speed at which the speaker delivers the original message (delivery rate) might affect the translation of DMs. Magnifico and Defrancq (2020) found a positive correlation between the factors delivery rate and omissions of connective items, albeit only in the causal and concessive functions. The authors categorised the delivery rate into three groups (below 120 words per minute, 120-159 words per minute, 159 or more words per minute). They found that, beyond the delivery rate of 120 words per minute, the share of the translated connective items starts to drop.

Still, according to general simultaneous interpreting research, a delivery rate of up to 120 words per minute is considered comfortable (Gerver, 1969/2002; Setton & Dawrant, 2016), while, in the past, rates beyond 170 words per minute were considered not possible for simultaneous interpreting (Lederer, 1981). Although simultaneous interpreters have to interpret speeches with a far greater delivery rate than 170 words per minute, current research (Setton and Dawrant, 2016) also notes that, at a delivery rate surpassing 160 words per minute, even the most skilled interpreters are forced to resort to omissions as a coping strategy.

2.4 Emotive intensity

To the knowledge of the author, the link between the emotive intensity of the original speaker, which we define as a combination of non-verbal communication (gestures and facial expressions), paralinguistic phenomena (prosody), verbal behaviour (emotionally charged language), as well as the level of extemporaneity, and translation accuracy of DMs, have not been explored so far.

Nevertheless, the link between emotionally charged language and translation accuracy is documented, and studies show a tendency of simultaneous interpreters in the European Parliament to mitigate their translations (see Kučiš & Majhenič, 2018; Beaton-Thome, 2020; Bartlomiejczyk, 2022).

Even though interpreters tend to water down emotionally laden utterances, research into the physiological response of interpreters found that interpreters are affected by the speaker's emotions (Korpal & Jasielska, 2018). Moreover, interpreters tend to converge emotionally with the original speaker and mimic their physiological arousal, which, as Korpal and Jasielska (2018) presumed, might help to comprehend the communicative intent of the original speaker.

While the two findings of mitigating emotive intensity and emotionally converging with the speaker seem contradictory, they are not irreconcilable. Simultaneous interpreters might well converge emotionally with the speaker and grasp the original communicative intent, yet, for some reason, still opt for a translation divergence.

3 METHODOLOGY

3.1 Dataset

The dataset in this study is part of the corpus TolAnSi and contains 98 speeches by 45 speakers that were simultaneously interpreted by 27 interpreters. The speeches were held in the European Parliament between November 2018 and January 2020 during the Brexit negotiations. They had to be non-read, performed by native English speakers from the United Kingdom and the Republic of Ireland, and translated by native Slovene accredited interpreters. The leveraged part of the corpus contains 46,999 tokens and is not publicly accessible yet.

3.2 Annotation procedure

The speeches were transcribed using the transcription tool Transcriber 1.5.1, while the alignment of the source language and the interpretation, as well as the annotation of DMs, was performed using the audio analysis tool Praat 6.3.08. To reduce functional ambiguity during the annotation process, the functional classification of DMs was carried out multimodally, i.e., by viewing the video recordings of the SL speeches in the annotation tool Elan 5.9. All annotations were performed by a linguist.

The transcriptions are separated into segments, corresponding predominately to prosodic units. A trained interpreter aligned the source language segments manually with the corresponding simultaneously interpreted segments.

3.3 Segmentation

As Setton (2011) cautioned, a syntactic-based segmentation is not optimal due to language differences, and could lead to very lengthy segments which are more difficult to align. A prosody-based segmentation, on the other hand, generally results in shorter segments (Setton, 2011). In this light, the Chinese interpreting corpora CECIC (Hu, 2016/2011) and CEIPPC (Wang & Tang, 2020) were segmented according to prosodic units. While pauses are a commonly used delimitator of prosodic units, they are not as useful in simultaneously interpreted speech, as the interpreter often pauses, for example, to ensure that their rendition is accurate (Setton, 2011).

Therefore, in our dataset, a prosodic unit is delimited by a combination of the most commonly used parameters, i.e., pitch reset (Degand & Simon, 2009; Cabedo, 2014; Degand et al., 2014; Zwitter Vitez, 2018; Beňuš, 2021), deacceleration or acceleration (Degand & Simon, 2009; Cabedo, 2014; Degand et al., 2014; Zwitter Vitez, 2018; Beňuš, 2021), pauses (Cabedo, 2014; Degand et al., 2014; Zwitter Vitez, 2018; Beňuš, 2021), breathing patterns (inhales) (Beňuš, 2021), and the presence of DMs (Degand & Simon, 2009; Cabedo, 2014; Degand et al., 2014; Zwitter Vitez, 2018; Beňuš, 2021).

3.4 DM annotation and classification

To identify potential DMs, the candidate items had to have a metaliguistic meaning and be syntactically and semantically detachable. As an optional second-level requirement, the item's multifunctionality and number of potential translation equivalents were considered.

The classification of DMs implemented in this research consolidates Crible's (2017) and Crible and Degand's (2019) with Maschler's (2009) topology. Maschler's (2009) three-pronged topology is used as an overarching categorisation of DMs by categorising DMs as textual, interpersonal, or cognitive. The domains of ideational, sequential, and rhetorical DMs are nested within the textual domain, as they denote semantic relations pertaining to the extralingual world and contain a lower degree of subjectivity. The functions used in the present research are adopted from Crible's (2017) functions, with the addition of cognitive functions and the interpersonal function of contact maintaining (corresponding to Crible's (2017) monitoring function), which are adopted from Maschler (2009). Due to the specificity of simultaneous interpreting, we added the functions of speech production or planning, as well as hesitation, which Crible and Degand (2019) nested within the sequential domain. Moreover, we followed Crible and Degand's (2019) work in allowing the functions of DMs to cross domains.

Despite this functionally comprehensive and annotationally liberal classification method, as DMs are multifunctional, per definition, a clear-cut functional classification seems inadquate. Therefore, we allowed for multiple functional annotations of DMs in both the SL and the simultaneous interpretations. Doing this, the various functional aspects of a given DM can be captured better, which, in turn, allows for an improved functional translation comparison.

3.5 DM translation classification

To analyse the share of equivalent and non-equivalent translations, translations of DMs were labelled provisionally as equivalents, shifts and omissions. However, the data from the corpus required additional tags, such as when interpreters produce a DM but omit the surrounding segment (tag Missing segment with equivalent), or when the whole segment containing the DM is omitted (tag Missing segment). Additionally, due to the syntactical differences between the languages, we had to account for the criterium of syntactical detachability when deciding on the type of translation provided by the simultaneous interpreter. The translation tags starting with »Non-DM« denote that the DM was translated, but that it does not fulfill the criteria of DMs (as presented in Section 3.4) in the other language.

As a result, translations were divided into two groups – the equivalent translation types:

- Equivalent – a translation with a DM that is functionally equivalent in the target language (see example in Section 4.2.1)
- Non-DM translation – an equivalent translation, but not with a DM, or with an expression that does not qualify as a DM in the target language (see example in Section 4.2.2)

- Close equivalent – a functionally very close translation, but not completely equivalent (see example in Section 4.1.1)
- Non-DM close equivalent – a functionally very close translation, but not completely equivalent, whereby the expression does not qualify as a DM in the target language (see example in Section 4.1.2)

and the non-equivalent translation types:

- Shift – a translation divergence with a DM that does not match the function of the DM in the source language
- Non-DM shift – a functionally different translation with an expression that is not a DM in the target language (see example in Section 1)
- Transformed segment – a non-equivalent translation where the segment is changed completely or a translation with a DM is not possible
- Omission – the DM is omitted in the target language (see example in Section 4.1.3)
- Missing segment – the segment containing the DM in the source language is omitted
- Missing segment with equivalent – the DM is translated, but the pertaining segment is omitted in the target language (see example in Section 4.2.3)

3.6 SL delivery rate classification

The delivery rate of the SL was divided into four rates: low, which designates speeds up to 130 tokens per minute (TPM), moderate, which is up to 160 TMP, high, which is up to 180 TPM, and very high, which is reserved for delivery rates of more than 180 TPM (cf. Monti et al., 2005; Magnifico & Defrancq, 2020).

3.7 SL emotive intensity classification

The following parameters were examined, to determine the level of emotive intensity:

- non-verbal communication, which can be perceived through facial expressions, gestures, posture and physical appearance (e.g., flushed face). The study will look at whether the speaker makes use of them or not, and, if so, to what degree and how frequently.

- paralinguistic phenomena, i.e., speech prosody – the use of pauses, whether the speaker uses a narrower or a wider pitch/intensity range (which can be perceived as a lively speech), etc.

- verbal communication, i.e., emotionally charged language – the use of provocations, sarcasm, insults, or offensive, vulgar, or politically incorrect expressions.

- the use of speech notes. The study will examine if the speaker uses notes, and, if so, to what degree they rely on them.

The frequency and the degree to which these parameters are deployed by the speakers correspond to the level of emotive intensity, whereby level 1 represents the lowest level of emotive intensity and level 5 the highest level of emotive intensity. Thus, the following levels were implemented:

1 – completely detached, i.e., almost no perceivable non-verbal communication, insignificant speech prosody, relying strongly on notes

2 – calm, but not completely detached (i.e., some perceivable/subtle facial expressions and/or gestures, more varied speech prosody), relying partially on notes

3 – some engagement, i.e., partially lively speech (perceivable through non-verbal communication and paralinguistic phenomena), may contain a mild form of provocation, but without prosodic prominence (raising one's voice or crying), or the use of sarcasm, not relying on notes

4 – lively speech with visible emotions perceivable through paralinguistic phenomena, non-verbal and verbal communication, may include provocation and/or sarcasm, not relying on notes

5 – very lively speech with strong emotions perceivable through paralinguistic phenomena, non-verbal and verbal communication, may include provocation, sarcasm, insults, offensive, vulgar, and/or politically incorrect expressions, not relying on notes.

An example of emotive intensity level 5 is given in Figure 1 (Davies, 2019) which exhibits a negative facial expression, prominent prosody (raised voice), a »pin-pointing« hand gesture while uttering a vulgar expression (»bollocks to Brexit«).

Figure 1: Emotive intensity 5

3.8 Hypotheses

Building upon the outlined methodology, we posit the following hypotheses:

- H1: Speeches with a delivery rate classified as high and very high have a smaller share of equivalent DM translation types than speeches with a low and moderate delivery rate.
- H2: Speeches with a higher emotive intensity, classified as levels 4 and 5, have a smaller share of equivalent DM translation types than speeches with level 2 or 3 emotive intensity.

4 RESULTS

Overall, 879 DMs were identified in the SL. These 879 DMs correspond to 394 DMs in the target language, irrespective of the translation type. The remaining 485 expressions cannot be classified as DMs, either because they were omitted, or were other expressions that did not fulfill the requirements of discourse markerhood in Section 3.4, in the target language.

The average SL delivery rate was 161.36 TPM and the median was 159.40 TPM, whereby the lowest was 116.81 TPM, and the highest delivery rate was

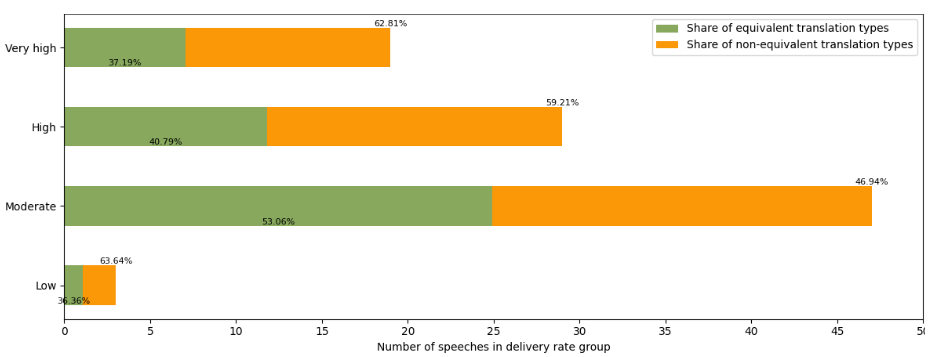
211.92 TPM. Regarding the categorisation of the speeches into delivery rate categories, only three speeches (3.06%) were categorised as having a low delivery rate, while the majority was represented by the moderate (47.96%) and high (29.59%) categories, leaving approximately a fifth (19.39%) to the very high delivery rate category.

The average value of the emotive intensity was 3.12 and the median was 3.0, whereby no speech was classified with the value 1, and six speeches (6.12%) had the top emotive intensity value 5. The lion's share (42.86%) of speeches had the emotive intensity level 3, and a quarter (25.51%) of the speeches are represented by both the emotive intensity levels 2 and 4.

4.1 SL delivery rate and equivalent translation types

The link between the SL delivery rate and the share of equivalent translation types of DMs is represented in Figure 2, where the bars on the vertical axis represent the SL delivery rates, and the horizontal axis denotes the number of the speeches in each delivery rate group, while the percentages in the stacked bars denote the share of each translation type.

Figure 2: SL delivery rate and translation types



In the following three subsections, examples of the individual translation types will be presented. It must be noted, that the translation type examples chosen for each delivery rate do not indicate that the given translation type is more or less likely to occur in the given delivery rate group. The examples merely serve translation type illustration purposes.

4.1.1 MODERATE DELIVERY RATE

The results show that the moderate delivery rate of 130 to 160 TPM had the greatest share of equivalent translation types, as more than half (53.06%) of the DMs preserved the original communicative intent. The equivalent translation types include the categories Equivalent, Non-DM translation, Non-DM close equivalent, and Close equivalent, which can be illustrated with the example below:

> *and I hope the next Parliament will address this | <u>secondly</u> sustainable finance should become the norm* (Gill, 2019)
>
> *in upam da bo naslednji Parlament to rešil | <u>potem</u> trajnostno financiranje naj bo norma (anonymous interpreter)*
>
> *[and I hope that the next Parliament will solve this | <u>then</u> sustainable financing should be the norm]*

The enumerating sequential DM "secondly" is translated with the Slovene DM "potem" [then], a topic resuming or continuing but also enumerating sequential DM. The functions of both DMs are indeed very close, yet an equivalent translation is also possible ("drugič" [secondly]). Therefore, the equivalent translation type cannot be assigned. Such instances were labelled as translation type Close equivalent, which designates a close and completely acceptable, yet not »perfect« equivalent.

4.1.2 HIGH DELIVERY RATE

The second-best results were obtained for the high delivery rate, where 40.79% of the DMs in the category were translated as equivalent types. An illustration of a close translation type with an expression other than a DM is the following case:

> *a many different organisations and stakeholders have made contact with us throughout our work on this | <u>and</u> I also thank them for their engagement and interest* (Palmer, 2019)
>
> *imamo veliko organizacij in deležnikov ki so stopili z nami v stik pri delu na tem dosjeju | <u>tako da</u> se tud njim zahvaljujem za angažiranost in interes* (anonymous interpreter)
>
> *[we have many organisations and stakeholders who contacted us regarding the*

> *work on this dossier | so I also thank them for their engagement and interest]*

At a first glance, "and" seems a mere adding sequential marker, but it can also be seen as a concluding rhetorical marker with respect to the previous utterance. Such instances warranted two functional annotations in the SL. In Slovene, the expression "tako da" [so] conveys primarily the communicative intent of conclusion, not as much the sequential addition function. As the translated expression corresponds to at least one of the source DM's functions, an equivalent translation type must be assigned. As the Slovene expression is, in this case, syntactically integrated, it cannot be considered a DM. Consequently, the only fitting translation type is Non-DM close equivalent, designating a functionally close translation, yet with an expression other than a DM.

4.1.3 VERY HIGH AND LOW DELIVERY RATES

The delivery rate categories very high and low had similar results, with the category very high containing 37.18% of equivalent translation types, and the category low containing 36.36% of equivalent translation types. The example below illustrates the translation type Omission that occurred in a speech with a very high SL delivery rate:

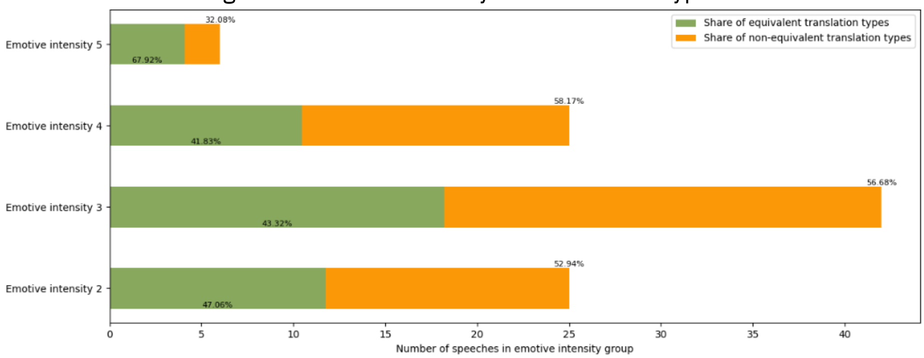
> *well thank you mister President | well I mean there's a big file but ninety percent of the interventions have been on one issue (Dalton, 2019)*
>
> *hvala | to je velik dosje ampak devetdeset odstotkov komentarjev je bilo v zvezi z enim vprašanjem (anonymous interpreter)*
>
> *[thank you | this is a big file but ninety percent of the comments pertained to one question]*

The three DMs in the SL were all omitted, even though the Slovene equivalents are relatively straightforward ("no" or "ja" [well], and "mislim" [I mean]). The effects of the high delivery can be connected with the interpreter's time lag, since the speaker had already begun to speak while the interpreter was still translating the Chairperson's message.

4.2 Emotive intensity and equivalent translation types

The interplay of the emotive intensity level and the share of equivalent translation types is represented in Figure 3, where the four bars represent the SL emotive intensity levels, and the horizontal axis the number of speeches in the emotive intensity group, while the percentages in the stacked bars denote the share of each translation type.

Figure 3: Emotive intensity and translation types



In the following three subsections, examples of the individual translation types will be given. Similar to Section 4.1, the examples chosen for each emotive intensity group do not represent the individual group.

4.2.1 HIGH EMOTIVE INTENSITY – LEVEL 5

As per Graph 2, the emotionally most intense speeches (group 5) also exhibited the greatest share, i.e., 67.97%, of equivalent translation types. Such translation types can be illustrated by the next example:

> *<u>well</u> this wasn't the plan <u>was it</u> | instead of sunlit uplands the best we are offered is survival (Howarth, 2019)*
>
> *<u>ja</u> | to ni bil načrt | <u>ali ne</u> | namesto vseh velikih načrtov nam ponujajo preživetje (anonymous interpreter)*
>
> *[<u>well</u> | this wasn't the plan | <u>was it</u> | instead of great plans we are offered survival]*

Both the SL and the interpretation contain the same DMs with the same communicative intent, as "well" and "ja" [right/well] can both signal a negative emphasis at the beginning of the speech. They were classified as negative emphasis interpersonal markers. Likewise, "was it" and "ali ne" [was it] serve the same provocative purpose, and were both classified as monitoring interpersonal DMs. As the functions in both languages overlap, the equivalent translation type is justified.

4.2.2 LOW EMOTIVE INTENSITY – LEVEL 2

The emotionally least intense speeches (group 2) contained almost half (47.06%) of all equivalent translation types, some of which were suitably translated with expressions other than DMs, as in the following example:

> *we've now received the latest reports from the platforms | on what they're doing | covering the month of March | a- <u>and</u> we'll publish our assessment of those reports in the coming days (King, 2019)*
>
> *zdaj beremo najnedavnejša poročila platform za mesec marec | <u>in</u> bomo našo oceno teh poročil objavili v prihodnjih dneh (anonymous interpreter)*
>
> *[we're now reading the latest reports from the platforms for the month of March | <u>and</u> we'll publish our assessment of these reports in the coming days]*

While the expression "in" [and] conveys the same communicative intent, i.e., that of sequential addition, in Slovene, the expression is syntactically integrated. Therefore, it does not meet the criteria for DMs. Consequently, only translation type Non-DM translation applies.

4.2.3 MIDDLE EMOTIVE INTENSITY – LEVELS 3 AND 4

A slightly greater share of non-equivalent translation types can be observed in the groups with the middle emotive intensity values (3 and 4), where equivalent translation types represented 43.32% and 41.83%. The example in the Introduction illustrates a case where the communicative intent diverged, and the corresponding expression was not a DM (translation type Non-DM shift). The case below, on the other hand, demonstrates a translation type where the DM preserves its original communicative intent, but the segment in which it is embedded is omitted (translation type Missing segment with equivalent):

> *<u>but</u> the problem is this | even though human rights are universal (Daly, 2019)*
>
> *<u>ampak</u> čeprav so človekove pravice univerzalne (anonymous interpreter)*
>
> *[<u>but</u> even though human rights are universal]*

In the example above, the interpreter merged both segments into one, whereby most of the first segment was omitted. As our analysis pertains to the translation of DMs, such cases were considered equivalent translation types, since the communicative intent of the DMs did not diverge.

5 DISCUSSION

According to the overall results, the average delivery rate of 161 TPM in the corpus TolAnSi is far from the ideal 100-120 words per minute and is considered very high, as it is at the threshold of what is currently considered still manageable, yet with potential strategic omissions by the interpreter (Setton & Dawrant, 2016). This suggests that, even though the results in Figure 2 seem less than ideal, as non-equivalent translations are predominant, they are indeed understandable and acceptable, considering the unfavourably high delivery rate. Nevertheless, since the greatest share of equivalent translation types was present in the moderate delivery rate group, this suggests that neither a very low nor a very high delivery rate is beneficial to the translation of DMs. Considering these results, the translation of DMs seems unaffected by the delivery rate, and other relevant factors need to be explored.

The analysis of the average emotive intensity of 3.12 on a scale from 1 to 5 in the corpus indicates that the speeches were rather emotionally laden than neutral. The results in Figure 3 are particularly interesting, as, without the speeches with the highest emotive intensity, the second hypothesis would have held true. Nevertheless, it must be highlighted that the group with the highest emotive intensity represents only 6.12% (and not, as it would be ideally, 20%) of the speeches. It remains, therefore, unclear, whether a higher emotive intensity affects the translation of DMs negatively, as the six speeches with the highest emotive intensity may coincidently, or, for a different reason, have a greater share of equivalent translation types.

Irrespective of the presented research findings, it is essential to address a critical limitation. Per definition, DMs are multifunctional, meaning that they convey different intents simultaneously. The question that arises is whether their multifunctionality contradicts the process of annotating functionally (non-)equivalent pairs across languages. In our research, we strived to accommodate this limitation by allowing multiple functional annotations per

DM in both languages, and using a comprehensive functional classification allowing cross-domain mobility. Moreover, the presented DM translation type classification includes several subtypes of (close) equivalent and non-equivalent translation types, and covers the encountered translation types in the corpus. Therefore, by combining a detailed functional and translation type classification, we minimised the risk of failing to identify (non-)equivalent translation pairs.

6 CONCLUSIONS AND LIMITATIONS

As the results in Figure 2 indicate, the share of equivalent DM translation types was not reduced by the rising SL delivery rate, but, conversely, was the smallest in the category with the lowest SL delivery rate and the highest in the moderate delivery rate category. Therefore, H1 was rejected.

Likewise, the emotive intensity did not seem to influence the share of equivalent translation types significantly, as the group with the highest emotive intensity (group 5) contained the most DMs with a preserved communicative intent, and it was followed by the group with the lowest emotive intensity (group 2). Consequently, H2 was also rejected.

The results gathered in the present research imply that neither the SL delivery rate nor the SL emotive intensity level influences to what degree the interpreters preserve the communicative intent of DMs. However, as emphasised in the previous section, more data are needed to provide more generalisable results. Despite both hypotheses being rejected, this research hints that other, not yet considered variables (perhaps the interpreter's experience level and the complexity of the source speech), need to be examined for their effect on the translation of DMs.

Moreover, as the DM annotation and emotive intensity assessment was provided by one annotator, it is prone to subjectivity, despite including some more objective and measurable parameters, such as speech prosody (pitch, intensity), detection of emotionally charged language, and the use of speech notes. In our future research, we therefore aim to include more human annotators, as well as automatise the process of emotive intensity assessment. Moreover, including additional languages would allow for more generalizable findings and enable cross-cultural comparisons.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency project MEZZANINE (Grant number J7-4642).

I sincerely thank the anonymous reviewers for their invaluable feedback and constructive comments.

REFERENCES

Aijmer, K., & Simon-Vandenbergen, A. M. (2003). The discourse particle well and its equivalents in Swedish and Dutch. *Linguistics, 41,* 1123–1161.

Aijmer, K. (1996). *Conversational Routines in English. Convention and Creativity.* London: Longman.

Aijmer, K. (2007). The meaning and functions of the Swedish discourse marker alltså – evidence from translation corpora. *Catalan Journal of Linguistics, 6:* 31–59.

Andersen, G. (1998). The pragmatic marker like from a relevance-theoretic perspective. In A. H. Jucker & Y. Ziv (Eds.), *Discourse Markers: Description and Theory* (pp. 147–170). Amsterdam, John Benjamins.

Bartłomiejczyk, M. (2022). Interpreting nonmainstream ideology (Euroscepticism) in the European Parliament. *Perspectives, 30*(4), 678–694. https://doi.org/10.1080/0907676X.2021.1939740

Beaton-Thome, M. (2020). Flagging the homeland: Interpreting Brexit à la Nigel Farage in the European Union. In K. Strani (ed), *Multilingualism and politics. Revisiting multilingual citizenship* (pp. 105–127). Palgrave Macmillan.

Beňuš, Š. (2021). *Investigating spoken English: A practical guide to phonetics and phonology using Praat*. Palgrave Macmillan.

Blakemore, D., & Gallai, F. (2014). Discourse markers in free indirect style and interpreting. *Journal of Pragmatics, 60,* 106–120.

Cabedo, A. (2014). On the delimitation of discursive units in colloquial Spanish: Val. Es. Co application model. In *Discourse Segmentation in Romance Languages* (pp. 157–183). John Benjamins.

Crible, L., & Degand, L. (2019). Domains and functions: A two-dimensional account of discourse markers. Discours. *Revue de linguistique, psycholinguistique et* *informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24), 1–35.

Crible, L. (2017). Towards an operational category of discourse markers: A definition and its model. In A. Sanso & C. Fedriani (Eds.), *Discourse markers, Pragmatics Markers and Modal Particles: New Perspectives* (pp. 99–124). John Benjamins.

Dalton, D. (2019). Speech at the European Parliament. Retrieved May 30, 2024, from: https://www.europarl.europa.eu/streaming/?event=20190416-0900-PLENARY (timestamp: 21:46:49)

Daly, C. (2019). Speech at the European Parliament. Retrieved May 30, 2024, from: https://www.europarl.europa.eu/streaming/?event=20191218-0900-PLENARY (timestamp: 21:00:54)

Davies, C. (2019). Speech at the European Parliament. Retrieved May 30, 2024, from: https://www.europarl.europa.eu/streaming/?event=20190918-0900-PLENARY (timestamp: 12:34:13)

Defrancq, B., Plevoets, K., & Magnifico, C. (2015). Connective Items in Interpreting and Translation: Where Do They Come From? In J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics, (3)* (pp. 195–222).

Degand, L., & Simon, A. C. (2009). On identifying basic discourse units in speech: theoretical and empirical issues. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics, (4)*.

Degand, L., Simon, A. C., Tanguy, N., & Van Damme, T. (2014). Initiating a discourse unit in spoken French. *Discourse Segmentation Romance languages* (250), 243–273.

Etheridge, B. (2019). Speech at the European Parliament. Retrieved May 30, 2024, from: https://www.europarl.europa.eu/streaming/?event=20190416-0900-PLENARY (timestamp: 17:57:08)

Fedriani, C., & Sansò, A. (2017). Pragmatic Markers, Discourse Markers and Modal Particles: What do we know and where do we go from here? In A. Sanso and C. Fedriani (Eds), *Discourse markers, Pragmatics Markers and Modal Particles: New Perspectives* (pp. 1–33). John Benjamins.

Fischer, K. (2006). Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume. In K. Fischer (Ed), *Approaches to Discourse Particles* (pp. 1–20).

Gerver, D. (1969/2002). The Effects of Source Language Presentation Rate on the Performance of Simultaneous Conference Interpreters. In F. Pöchhacker & M. Shlesinger (Eds.), *The Interpreting Studies Reader* (pp. 52–66). London/New York: Routledge.

Gill, N. (2019). Speech at the European Parliament. Retrieved May 30, 2024, from https://www.europarl.europa.eu/streaming/?event=20190313-0900-PLENARY (timestamp: 17:57:50)

Hale, S. B. (2004). *The Discourse of Court Interpreting: Discourse Practices of the Law, the Witness, and the Interpreter.* Amsterdam/Philadelphia: John Benjamins.

Howarth, J. (2019). Speech at the European Parliament. Retrieved May 30, 2024, from: https://www.europarl.europa.eu/streaming/?event=20190918-0900-PLENARY (timestamp: 12:24:08)

Hu, K. (2016/2011). *Introducing Corpus-based Translation Studies.* New York: Springer.

Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing.* London in New York: Continuum.

King, J. (2019). Speech at the European Parliament. Retrieved May 30, 2024, from: https://www.europarl.europa.eu/streaming/?event=20190416-0900-PLENARY (timestamp: 17:36:49)

Korpal, P., & Jasielska, A. (2018). Investigating interpreters' empathy. *Target,* 31(1), 2–24.

Kučiš, V., & Majhenič, S. (2018). Cultural and stress-related manifestations of political controversial language in the European Parliament from the view of interpreters. *Babel*, 64(1), 33–62.

Lederer, M. (1981). *La Traduction Simultanée Expérience et Théorie.* Paris: Minard – Lettres Modernes.

Magnifico, C., & Defrancq, B. (2020). Norms and gender in simultaneous interpreting: a study of connective markers. *The International Journal of Translation and Interpreting Research,* 12(1), 1–17.

Maschler, Y., & Schiffrin, D. (2015). Discourse markers: Language, meaning, and context. In D. Tannen, H. E. Hamilton & D. Schiffrin (Eds.), *The Handbook of Discourse Analysis* (pp. 189–221), Hoboken, ZDA: John Wiley and Sons.

Maschler, Y. (2009). *Metalanguage in Interaction. Hebrew discourse markers.*

Amsterdam/Philadelphia: John Benjamins Publishing Company.

Mauranen, A. (2023). *Reflexively Speaking: Metadiscourse in English as a Lingua Franca.* Berlin, Boston: De Gruyter Mouton.

Monti, C., Bendazzoli, C., Sandrelli, A., & Russo, M. (2005). Studying directionality in simultaneous interpreting through an Electronic Corpus: EPIC (European Parliament Interpreting Corpus). *Meta,* 50(4), np.

Müller, S. (2005). *Discourse markers in native and non-native English discourse. John Benjamins.* Amsterdam.

Palmer, R. (2019). Speech at the European Parliament. Retrieved May 30, 2024, from: https://www.europarl.europa.eu/streaming/?event=20190327-0900-PLENARY (timestamp: 19:23:30)

Redeker, G. (2006). Discourse markers as attentional cues at discourse transitions. In K. Fischer (Ed.), *Approaches to Discourse Particles* (pp. 339–358).

Rozumko, A. (2021). Underspecification in the translation of discourse markers: A parallel corpus study of the treatment of connective functions of indeed in Polish translations. *Journal of Pragmatics,* 177, 122–134.

Schiffrin, D. (1987). *Discourse Markers.* Cambridge: Cambridge University Press.

Schiffrin, D. (2001). Discourse markers: language, meaning, and context. In D. Schiffrin, D. Tannen, and E. H. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 54–75).

Schourup, L. (1999). Discourse markers. *Lingua, 107(3),* 227–265.

Setton, R. (2011). Corpus-based interpreting studies (CIS): Overview and prospects. In A. Kruger, K. Wallmach, & J. Munday (Eds.) *Corpus-based translation studies. Research and applications* (pp. 33–75). London: Continuum.

Setton, R., & Dawrant, A. (2016). *Conference Interpreting: A Trainer's Guide.* Amsterdam, Netherlands: John Benjamins.

Wang, B., & Tang, F. (2020). Corpus-based Interpreting Studies in China: Overview and Prospects. In K. Hu & K. H. Kim (Eds.) *Corpus-based Translation and Interpreting Studies in Chinese Contexts; Present and Future* (pp. 61–87). Palgrave Macmillan.

Zwitter Vitez, A. (2018). Enota analize spontanega govora: interakcija prozodije, pragmatike in skladnje. *Jezik in slovstvo, 63(2–3)*, 157–175.

RAZHAJANJE KOMUNIKATIVNEGA NAMENA DISKURZNIH OZNAČEVALCEV V SIMULTANO TOLMAČENEM GOVORU

Pričujoči članek preučuje razhajanje komunikativnega namena diskurznih označevalcev pri simultanem tolmačenju iz angleščine v slovenščino v Evropskem parlamentu, in sicer raziskuje vpliva hitrosti govora izhodiščnega govornika in njegove čustvene jakosti na delež enakovredno prevedenih diskurznih označevalcev, tj. takšnih, ki ohranijo enak komunikativni namen. Članek prav tako predstavlja klasifikacijo prevodov diskurznih označevalcev, namenjeno simultanemu tolmačenju angleških in slovenskih diskurznih označevalcev. Predstavljena klasifikacija izhaja iz empiričnega gradiva, ki zajema skoraj 47.000 pojavnic in upošteva vse zaznane prevodne tipe diskurznih označevalcev. Rezultati, ki temeljijo na korpusu TolAnSi, kažejo, da tako hitrost govora izhodiščnega govorca kot tudi stopnja čustvenosti ne vplivata bistveno na delež enakovredno prevedenih diskurznih označevalcev.

Keywords: diskurzni označevalci, simultano tolmačenje, hitrost govora, stopnja čustvenosti, klasifikacija tipa prevoda

SISTEM ZA ZAZNAVANJE SPREMEMB V RABI BESED IN NJEGOVA UPORABA ZA SOCIOLINGVISTIČNO ANALIZO

Matej MARTINC,[1] Veronika BAJT,[2] Špela ROT,[3] Senja POLLAK[1]
[1]Institut Jožef Stefan, Ljubljana
[2]Mirovni Institut, Ljubljana
[3]Filozofska fakulteta Univerze v Ljubljani, Ljubljana

V prispevku predstavimo prvi spletni, javno dostopni sistem za zaznavanje sprememb v rabi besed v slovenščini, ki omogoča rangiranje besed glede na spremembo v rabi v različnih časovnih obdobjih, ter interpretacijo teh sprememb s pomočjo spletnega uporabniškega vmesnika. Najprej predstavimo specifikacije sistema, grafični uporabniški vmesnik in metodologijo za odkrivanje sprememb. Nadaljnje demonstriramo, kako lahko sistem uporabimo za iskanje in analizo besed, ki so spremenile rabo v različnih časovnih obdobjih. Pokažemo, da sistem omogoča zaznavo in analizo kratkoročnih (tj. letnih) in dolgoročnih (tj. več kot dvajsetletnih) sprememb v rabi besed, ter da je mogoče meriti spremembe v rabi besed tekom več zaporednih obdobij. Nazadnje sistem uporabimo na primeru reprezentacije migracij v različnih časovnih obdobjih z ročno določenimi ločnicami, ki so signifikantno vplivale na odnos do migracije in migrantov v Sloveniji. Sistem je javno dostopen na http://kt-nlp-demo.ijs.si:8080.

Ključne besede: Zaznavanje sprememb v rabi besed, sociolingvistika, migracije

1 UVOD

Jezik je dinamičen sistem, ki odraža kulturni in tehnološki razvoj družbe (Aitchison, 2001). To pomeni, da se pomen in raba besed nenehno spreminjata in razvijata z uporabo v družbenih interakcijah, s spremembami kulturnih praks ter razvojem tehnologije. Preučevanje semantičnih premikov in sprememb v rabi besed je zaposlovalo učenjake že dolgo pred pojavom sodobnega jezikoslovja v poznem 19. in zgodnjem 20. stoletju (Tahmasebi in sod., 2021). Večinoma evolucija pomena besede poteka skozi več večpomenskih stopenj, zaradi česar je prepoznavanje in razumevanje semantičnih sprememb težka naloga (Hopper in sod., 1991). Kljub temu pa je delo na tej temi pomembno za jezikoslovne raziskave in družbeno analizo, saj semantični premiki v jeziku odražajo spremembe v družbi in jih je mogoče uporabiti kot približek za odkrivanje kulturnih in jezikovnih trendov (Gillani in Levy, 2019; Gantar in sod., 2018).

Študije jezikovne evolucije se bodisi osredotočajo na dolgoročne spremembe pomena besed ali pa na precej pogoste kratkoročne evolucijske pojave, kot je npr. pojavitev besede v novem kontekstu, kjer pomen besede ostaja nespremenjen v leksikografskem smislu (Montariol in sod., 2021). Ko v tej študiji govorimo o spremembah v rabi besed, se nanašamo na vse vrste jezikovne evolucije – kratkoročne ali dolgoročne, z ali brez pomenskih sprememb. Gre torej za široko kategorijo, ki poleg semantičnih premikov vključuje tudi premike v kontekstu, v katerem se beseda pojavlja.

Področje avtomatskega odkrivanja 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 (Juola, 2003; Hilpert in Gries, 2008), so raziskave na to temo dobile zagon z idejo o uporabi besednih vektorskih vložitev (tj. predstavitev besed s pomočjo vektorjev) za konstrukcijo časovnih reprezentacij (Tahmasebi in sod., 2018). Najsodobnejši sistemi za zaznavanje sprememb tako uporabljajo različne vrste besednih vložitev, v namen sistematične primerjave različnih metod pa je bilo nedavno organiziranih tudi več tekmovanj in delavnic (Schlechtweg in sod., 2020; Basile in sod., 2020; Kutuzov in sod., 2021; Zamora-Reina in sod., 2022).

Te delavnice so večinoma zajemale odkrivanje sprememb v rabi besed v jezikih z veliko viri in veliko govorci, kot so angleščina, ruščina, nemščina, italijanščina in španščina. Po drugi strani je bila večina jezikov z manj viri (med njimi tudi slovenščina) zanemarjena. Z namenom zmanjšanja te identificirane vrzeli, smo v tej študiji naredili in objavili prvi javno dostopen sistem za zaznavanje sprememb v rabi posameznih besed za slovenščino in ga uporabili za analizo več časovnih korpusov. Bolj specifično, glavni doprinosi študije so sledeči:

- Prilagoditev metode za zaznavanje sprememb v rabi besed za slovenščino. Metoda temelji na nenadzorovanem gručenju (tj. gručenju s pomočjo algoritma, ki za delovanje ne potrebuje učenja na ročno označenih podatkih) besednih vložitev in se je izkazala za zelo uspešno na več različnih jezikih (Montariol in sod., 2021).

- Implementacija spletnega uporabniškega vmesnika za prikaz in interpretacijo sprememb v rabi besed. Uporabniški vmesnik je javno dostopen na spletnem naslovu http://kt-nlp-demo.ijs.si:8080/.
- Demonstracija uporabnosti sistema za iskanje in analizo besed, ki so najbolj spremenile rabo v daljšem časovnem obdobju (tj. okoli 20 let) ter v krajših (letnih) intervalih. Na primeru migracije pokažemo, da je sistem uporaben tudi za sociolingvistične raziskave.

2 POVEZANE RAZISKAVE

Prvi sistemi za samodejno zaznavanje sprememb v rabi besed so bili razviti pred več kot desetletjem. Zanašali so se na metode, ki temeljijo na frekvenci (Juola, 2003; Hilpert in Gries, 2008) in se danes redko uporabljajo, saj je uporaba besednih vložitev (Mikolov in sod., 2013) privedla do razvoja učinkovitejših metod za to nalogo. Podroben pregled teh nekoliko starejših metod, ki temeljijo na frekvenci, lahko najdete v Tahmasebi in sod. (2018) in Tang (2018). V zadnjem času področje avtomatskega zaznavanja sprememb v rabi besed postaja vse pomembnejše, saj je uporabno ne le v jezikoslovju, temveč tudi pri različnih socioloških in kulturoloških raziskavah. Sem sodi spremljanje besedilnih tokov (streams) z namenom zaznavanje dogodkov (Kutuzov in sod., 2017), analiza gledišča (Azarbonyad in sod., 2017; Martinc in sod., 2021), analiza toka novic (Montariol in sod., 2021) ter zaznavanje sprememb v diskurzu, ki jih povzročijo krizni dogodki (Stewart in sod., 2017).

Najnovejši sistemi za zaznavanje sprememb v rabi besed uporabljajo dve vrsti nenadzorovanih metod. Temeljijo na konstrukciji časovnih reprezentacij z uporabo statičnih ali kontekstualnih besednih vložitev. Metode, ki uporabljajo statične vložitve, temeljijo na učenju statičnega modela besednih vložitev za vsak časovni segment korpusa in nato na poravnavi teh modelov, da postanejo primerljivi. To je mogoče doseči z uporabo *inkrementalnega posodabljanja* (Kim in sod., 2014), kjer se model najprej uči na prvi časovni rezini korpusa in nato posodobi na vsaki naslednji časovni rezini. Druga možnost je uporaba *poravnave vektorskega prostora* (Hamilton in sod., 2016). Tu se modeli učijo neodvisno na vsaki časovni rezini, na koncu pa se izvede poravnava modelov z optimizacijo geometrijske transformacije.

Vse metode, ki uporabljajo statične besedne vložitve, imajo omejitve glede občutljivosti in interpretabilnosti zaradi dejstva, da je vsaka beseda predstavljena samo z eno vektorsko reprezentacijo znotraj časovne rezine. Te omejitve je mogoče odpraviti z uporabo kontekstualnih besednih vložitev, ki jih lahko npr. pridobimo iz sistema BERT (Devlin in sod., 2019), kjer se za vsak kontekst, v katerem se pojavi beseda, ustvari drugačna vektorska reprezentacija, kar omogoča tudi modeliranje večpomenskih besed. Vsi ti pristopi vsebujejo nek postopek agregacije, v katerem so kontekstualne vložitve posameznih pojavitev besed v določenem časovnem obdobju v korpusu združene v smiselne časovne reprezentacije. V različnih študijah se uporabljajo različne metode, od preprostega povprečenja (Martinc, Kralj Novak in Pollak, 2020) do primerjave parov časovnih vektorjev (Kutuzov in Giulianelli, 2020) in združevanja v gruče, kjer se predvideva, da posamezna gruča reprezentacij ustreza posameznemu pomenu dane besede (Giulianelli in sod., 2020; Martinc, Montariol in sod., 2020; Montariol in sod., 2021). Najbolj priljubljena metoda za primerjavo distribucij gruč iz različnih časovnih obdobij, in s tem pridobitev kvalitativne ocene spremembe v rabi določene besede, je Jensen-Shannonova divergenca (JSD) (Lin, 1991), ki se uporablja v študijah Giulianelli in sod. (2020) in Martinc, Montariol in sod. (2020). V Montariol in sod. (2021) predlagajo zamenjavo JSD z Wassersteinovo razdaljo (WD) (Solomon, 2018), kar motivirajo s pojasnilom, da WD poleg primerjave distribucije gruč upošteva tudi položaj gruč v semantičnem prostoru, kar vodi do boljše učinkovitosti sistema.

Še en zanimiv pristop, ki prav tako uporablja kontekstualne besedne vložitve, je predlagal Rosin in sod. (2022). Da bi bil model, temelječ na BERT arhitekturi, občutljiv na specifično časovno rabo besed, predlagajo dodajanje posebnega časovnega žetona vsakemu vhodnemu besedilnemu zaporedju med učenjem modela s pomočjo naloge napovedi zamaskiranih besed v tekstu (masked language modelling). S tem želijo neposredno vključiti časovno informacijo v proces učenja, tako naučen jezikovni model pa lahko med drugim uporabimo za datiranje stavkov.

Raziskave na temo sprememb v rabi besed v slovenščini so zelo redke. Za nas je relevantna predvsem pred kratkim izvedena študija Pranjić in sod. (2024), kjer so naredili prvo testno množico za testiranje slovenskih modelov za zaznavanje sprememb v rabi besed na slovenskih korpusih. V študiji predstavijo nov model za zaznavanje semantičnih premikov s pomočjo optimalnega transporta, hkrati pa testirajo tudi metodologijo, ki jo opisujemo v tej študiji. Za razliko od te študije, se študija od Pranjić in sod. (2024) ne posveča izdelavi prvega javno dostopnega spletnega sistema za zaznavanje sprememb v rabi besed v slovenščini in njegovi uporabnosti za sociolingivistične raziskave.

3 OPIS SISTEMA ZA ZAZNAVANJE SPREMEMB V RABI BESED

Kot je bilo povedano že v poglavju 1, javno dostopen sistem za zaznavanje sprememb v rabi besed temelji na metodi nenadzorovanega gručenja besednih vložitev, ki se je izkazala za zelo uspešno na več različnih jezikih (Montariol in sod., 2021). Metoda je bila izbrana predvsem zato, ker za delovanje ne potrebuje učenja, saj v slovenščini ne obstaja učna množica za učenje modelov za detekcijo sprememb v rabi besed, in pa zato, ker kot ena redkih metod omogoča vizualizacijo in interpretacijo sprememb posameznih besed. V nadaljnje podrobno opišemo vhodne, izhodne in računske specifikacije sistema ter izdelan cevovod za zaznavanje časovnih sprememb v rabi besed.

3.1 Vhodne specifikacije

Kot vhod sistem potrebuje korpus, ki vsebuje besedila iz različnih časovnih obdobij in ga je mogoče razdeliti na različna časovna obdobja. Dolžina posameznih časovnih obdobij in razmejitve med obdobji so v teoriji poljubne, v praksi pa so pogojene z raziskovalnim vprašanjem, ki nas zanima in količino podatkov, ki jo imamo. V idealnem primeru mora vsaka časovna rezina korpusa vsebovati vsaj 5 milijonov besed. To omogoča sestavo obsežnega besedišča lem, za katere lahko določimo časovno spremembo v rabi. Vsaka beseda, za katero želimo določiti spremembo v rabi, se mora za veljavnost rezultatov pojaviti v vsaki časovni rezini korpusa vsaj 20-krat (v idealnem primeru pa vsaj 100 krat), saj manj kot 20 pojavitev določene besede ne omogoča izdelave dovolj kvalitetne distribucije rab besede za posamezno obdobje.

3.2 Računske specifikacije

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 pri večini potrebno 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 (Montariol in sod., 2021).

Medtem ko je največji korpus, na katerem je bil preizkušen sistem, vseboval približno 100 milijonov besed na časovno rezino in besedišče sestavljeno iz približno 8000 lem, teoretično ni zgornje meje za velikost korpusa. Vendar pa je treba upoštevati nekatere praktične omejitve, saj se zahteve glede diskovnega spomina povečujejo z velikostjo besedišča in številom časovnih obdobij. Za vsako časovno obdobje je potrebno na disk shraniti do 200 kontekstualnih besednih vložitev za posamezno lemo, ena vložitev pa ima velikost približno 3072 bajtov. Na primer, če bi korpus vseboval 4 časovna obdobja in bi besedišče vsebovalo 50.000 lem, bi to zahtevalo 4 * 50.000 * 200 * 3072 = 122,8 GB prostora na disku. Po drugi strani zahteve po RAM-u niso odvisne od števila časovnih obdobij, kar pomeni, da bi zgornji primer zahteval 50.000 * 200 * 3072 = 30,7 GB RAM-a.

3.3 Cevovod za zaznavanje časovnih sprememb v rabi besed

Cevovod za zaznavanje sprememb v rabi besed je sestavljen iz več zaporednih korakov:

1. **Predprocesiranje korpusa:** Tokenizacija in lematizacija korpusa s pomočjo orodij za predprocesiranje (v primeru slovenščine uporabljamo orodje CLASSLA-Stanza (Terčon in Ljubešić, 2023)).

2. **Domenska adaptacija nevronskega modela:** Nevronski jezikovni model 5 epoh učimo na celotnem korpusu na nenadzorovan način, tj. na nalogi napovedovanja naključno zamaskiranih besed v tekstu.

3. **Razdelitev korpusa na časovne rezine:** Korpus razdelimo na časovne rezine, ki se ločeno vnesejo v model v serijah (ang. batch) po 32 tekstovnih sekvenc naenkrat. Tekstovne sekvence omejimo na dolžino 256 žetonov.

4. **Ekstrakcija kontekstualnih vložitev:** Za vsako sekvenco ustvarimo reprezentacijo sekvence tako, da vzamemo in seštejemo zadnje štiri izhodne plasti nevronov kodirnika nevronske mreže. Nato vsako reprezentacijo zaporedja razdelimo na 256 delov, da dobimo kontekstualno vložitev velikosti 768 nevronov za vsako besedo. Beseda je lematizirana in vektor, ki predstavlja njeno kontekstualno vložitev, se bodisi shrani na seznam kontekstualnih vložitev za določeno besedno lemo, ali pa se združi z enim od predhodno pridobljenih kontekstualnih vektorjev za isto lemo v istem časovnem obdobju. Da bi izboljšali skalabilnost sistema, omejimo število kontekstualnih vložitev, ki se hranijo v pomnilniku za dano lemo in časovno obdobje, na vnaprej določen prag določen pri 200, ki je bil empirično izbran iz nabora kandidatov za prag (20, 50, 100, 200, 500) in ponuja razumen kompromis med skalabilnostjo sistema in točnostjo reprezentacije za posamezno lemo v posameznem časovnem obdobju. Novo pridobljeno besedno vložitev se združi z najbližjo shranjeno besedno vložitvijo na seznamu glede na kosinusno razdaljo, če 1.) je nova vložitev preveč podobna eni od prejšnjih shranjenih vložitev (tj., kosinusna podobnost je večja ali enaka 0.99) ali 2.) če seznam že vsebuje vnaprej določeno največje število vektorjev (200 v našem primeru). Na ta način za vsako lemo v besedišču pridobimo do 200 kontekstualnih vložitev, ki predstavljajo posamezno (ali združeno) pojavnico s to lemo v kontekstu.

5. **Gručenje kontekstualnih vložitev:** Da bi pridobili posamezne rabe za lemo v določenem časovnem obdobju, kontekstualne vložitve, ki pripadajo posamezni lemi, gručimo s pomočjo algoritma k-means. Združevanje v gruče za dano lemo se izvede na množici vseh vektorjev iz vseh časovnih obdobij skupaj, da so gruče v različnih časovnih obdobjih primerljive med sabo. Vrednost k v k-means algoritmu je določena pri 5, saj ima večina besed manj kot 5 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 v teoriji vodijo v največje razlike med besednimi vložitvami), temveč na podlagi sintaktičnih in slovničnih razlik, kar ponavadi poslabša učinkovitost metode. Zato po zgoraj opisanem postopku združevanja vložitev pojavnic v 5 gruč s pomočjo k-means algoritma izvedemo doda-

tno združevanje (če sta si dve gruči zelo podobni) in odstranitev gruč (če je v gruči manj kot 10 pojavov leme, kar kaže na precej obrobno rabo leme).

6. **Izdelava distribucij rab za vsako časovno obdobje:** Za vsako lemo v vsakem časovnem obdobju v zgornjem koraku pridobimo množico gruč, ki predstavljajo posamezne rabe besede. Iz teh gruč nato pridobimo distribucijo rab tako, da število pojavnic leme v vsaki gruči delimo s frekvenco leme v časovnem obdobju.

7. **Merjenje sprememb v rabi:** Distribucije rab, ki jih za določeno lemo pridobimo za vsako časovno obdobje v korpusu, primerjamo med sabo s pomočjo mere Jensen Shannon Divergence (JSD) (Menéndez et al., 1997), ki se uporablja za merjenje razlik med distribucijami. S pomočjo mere JSD lahko vse besede v besedišču razporedimo glede na velikost izmerjene spremembe v rabi med zaporednimi obdobji in na ta način poiščemo besede, katerih raba se je med različnimi časovnimi obdobji najbolj spremenila.

3.4 Izhodne specifikacije

Predlagani sistem nam omogoča, da razumemo, kako se raba posamezne leme spreminja med časovnimi obdobji s pomočjo metode za interpretacijo na podlagi mere Term frequency - inverse document frequency (TF-IDF). Za vsako rabo posamezne leme imamo na voljo kontekst (tj. stavek, v katerem se pojavnica, ki pripada določeni lemi, pojavi). Stavke, ki vsebujejo posamezne rabe besede, ki pripadajo isti gruči, združimo v “dokument”, nato pa izluščimo najbolj diskriminatorne unigrame, bigrame in trigrame za vsakega od teh dokumentov s pomočjo algoritma “term frequency - inverse document frequency” (TF-IDF), kjer skupek vseh “dokumentov” (tj. množica vseh stavkov v korpusu, v katerih se pojavi posamezna lema) obravnavamo kot korpus. Funkcijske besede brez pomena (ang. stopwords) in besede, ki se pojavljajo v več kot 80% gruč, so izključene, da se zagotovi, da so izbrane ključne besede za vsako gručo čim bolj specifične. Na koncu dobimo seznam do 7 ključnih besed za vsako gručo, ki nudijo vpogled v posamezno rabo besede.

3.5 Implementacija uporabniškega vmesnika

S pomočjo odprtokodnega JavaScript ogrodja Vue.js smo implementirali uporabniški vmesnik, ki omogoča interpretacijo in analizo sprememb v rabi. Uporabniški vmesnik je sestavljen iz dveh ločenih komponent. Prva komponenta ponuja globalni pogled na korpus razdeljen na časovne rezine s prikazom tabele, v kateri so vse besede v korpusu, ki se pojavijo več kot 20-krat (glej sliko 1). Privzeto so besede razvrščene glede na skupno izmerjeno spremembo v rabi med prvim in zadnjim časovnim obdobjem v korpusu, vendar pa tabela omogoča razvrščanje po poljubnem stolpcu.

S klikom na posamezno besedo pridemo do druge komponente uporabniškega vmesnika, ki omogoča analizo sprememb v rabi za posamezno besedo (glej sliko 2). Komponenta vizualizira distribucijo rab za določeno časovno obdobje, s klikom na posamezno gručo (tj. barvo) na sliki, pa se izpiše seznam kontekstov (tj. stavkov), ki sodijo v določeno gručo. Uporabniški vmesnik je zasnovan na način, da lahko uporabnik z bolj splošnih informacij (na korpusni ravni), ki jih prikazuje prva komponenta, hitro (s pomočjo klika na posamezno besedo) prehaja na bolj podrobne informacije (na besedni ravni), ki jih prikazuje druga komponenta, kar omogoča hitro časovno analizo sprememb v rabi besede.

4 PRIMER UPORABE: ISKANJE BESED, KI SO NAJBOLJ SPREMENILE RABO V RAZLIČNIH ČASOVNIH OBDOBJIH IN NJIHOVA ANALIZA

Za slovenščino smo nevronski model SloBERTa (Ulčar in Robnik-Šikonja, 2021), ki smo ga uporabili za ekstrakcijo kontekstualnih besednih vložitev, naučili na delu korpusa Gigafida 2.0 (Krek in sod., 2020). Gigafida 2.0 je referenčni korpus standardne pisane slovenščine, in vsebuje besedila iz časopisov (47,8 odstotkov besedil), revij (16,5 odstotkov), internetnih vsebin (28,0 odstotkov)[1], stvarnih besedil (3,8 odstotkov), leposlovja (3,5 odstotkov) in drugih zvrsti.

Iz besedil iz Gigafide 2.0 smo zgradili tri različne korpuse. Prvi korpus[2] omogoča merjenje dolgoročnih sprememb v rabi med dvema obdobjema, obdobjem, ki pokriva 8 let med 1990 in 1997 (to obdobje vsebuje najstarejša besedila v

[1] Internetni teksti vsebujejo tudi novice iz novičarskih portalov, ki so po vsebini zelo podobne časopisnim besedilom.
[2] http://kt-nlp-demo.ijs.si:8080/semanticshifttable/2

GigaFidi 2.0, za razpon 8 let pa smo se odločili predvsem zato, da smo pridobili dovoljšnjo količino teksta za učenje modela) ter zadnjim letom v Gigafidi 2.0, ki vsebuje besedila iz leta 2018. V tem korpusu nas zanimajo predvsem dolgoročne spremembe v rabi besed, ki so nastale v časovnem obdobju več kot 20 let.

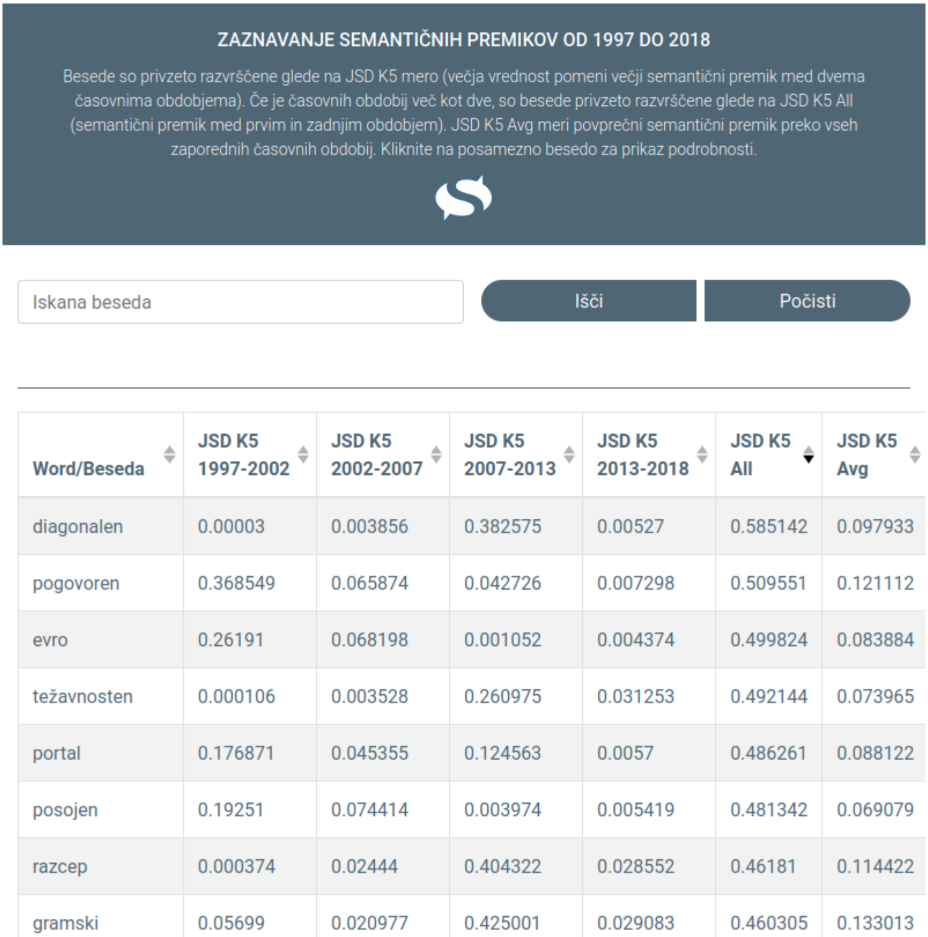
Tabela 1: Število dokumentov, besed in virov po letih v treh korpusih za merjenje sprememb v rabi besed.

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

Drugi korpus[3] omogoča merjenje sprememb v rabi med dvema letnima obdobjema, 2017 in 2018. V tem korpusu želimo meriti kratkoročne spremembe v rabi besed, ki so nastale v časovnem obdobju enega leta. Tretji korpus[4] po drugi strani za razliko od prvih dveh korpusov pokriva pet obdobij, tj. obdobja 1990-1997, 2002, 2007, 2013, 2018. Tu želimo meriti spremembe v rabi besed med več zaporednimi obdobji in na ta način bolje razumeti celotno dinamiko spreminjanja rabe besed, ki ne poteka vedno linearno in v eni smeri. Vsi korpusi so predstavljeni v Tabeli 1.

Primer uporabniškega vmesnika za vhodni korpus sestavljen iz štirih časovnih obdobjih predstavljen v Tabeli 1 je predstavljen na sliki 1. Besede so privzeto razvrščene po meri *JSD K5 All*, ki s pomočjo mere JSD izmeri razliko med distribucijama (pridobljenima s pomočjo algoritma k-means s k=5) v rabi besede med prvim in zadnjim obdobjem v korpusu (ang. beseda "All" označuje, da gre

[3] http://kt-nlp-demo.ijs.si:8080/semanticshifttable/3
[4] http://kt-nlp-demo.ijs.si:8080/semanticshifttable/1



ZAZNAVANJE SEMANTIČNIH PREMIKOV OD 1997 DO 2018

Besede so privzeto razvrščene glede na JSD K5 mero (večja vrednost pomeni večji semantični premik med dvema časovnima obdobjema). Če je časovnih obdobij več kot dve, so besede privzeto razvrščene glede na JSD K5 All (semantični premik med prvim in zadnjim obdobjem). JSD K5 Avg meri povprečni semantični premik preko vseh zaporednih časovnih obdobij. Kliknite na posamezno besedo za prikaz podrobnosti.

Iskana beseda

Išči

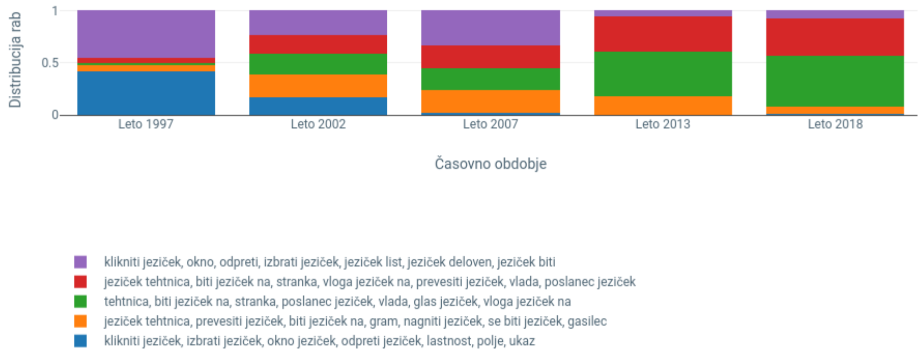
Počisti

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
gramski	0.05699	0.020977	0.425001	0.029083	0.460305	0.133013

Slika 1: Uporabniški vmesnik za prikaz in iskanje vseh besed v besedišču korpusa. Besede so razporejene glede na zaznano spremembo v rabi.

za spremembo v rabi besede od prvega do zadnjega obdobja). Vidimo, da se je glede na ta kriterij, najbolj spremenila raba besede *diagonalen*. Hkrati lahko s pomočjo vrednosti v drugih stolpcih, ki prikazujejo spremembe (merjene z JSD) med zaporednimi obdobji, opazimo, da je do izrazite spremembe v rabi prišlo med obdobjema 2007 in 2013 (JSD vrednost okoli 0.38).

Če kliknemo na določeno besedo, pridemo do druge komponente uporabniškega sistema. Primer izhoda sistema naučenega na vhodnem korpusu sestavljenem



Slika 2: Primer izhoda sistema za besedo jeziček.

iz štirih časovnih obdobjih za besedo *jeziček* (tj., trinajsta najbolj spremenjena beseda v besedišču glede na mero *JSD K5 All*), je predstavljen na Sliki 2. Vidimo, da se je v obdobju do leta 1997 beseda jeziček v veliki večini primerov nanašala na določen element spletnih strani, ki se v angleščini imenuje dropdown. To rabo pokrivata dve gruči na sliki, vijolična in modra. Sčasoma, do leta 2013, se ta specifična raba besede skoraj popolnoma preneha, beseda jeziček pa se v letih 2013 in 2018 v veliki večini uporablja znotraj fraze "jeziček na tehtnici". To rabo pokrivata rdeča in zelena gruča.

Seveda pa je pri interpretaciji samo na podlagi slike potrebno biti previden. Za dokončno analizo je nujno treba upoštevati sestavo časovnega podkorpusa in preveriti besedila, iz katerih specifična raba izhaja. Npr., v zgornjem primeru bi prenehanje računalniške rabe besede jeziček lahko bilo povezano tudi s sestavo virov podkorpusa, tj. v podkorpusu, ki vsebuje besedila do 1997 bi lahko bilo znatno več računalniških virov kot v podkorpusu iz leta 2018, kar bi lahko vplivalo na distribucijo rab. Vire besedil za posamezno gručo lahko preverimo s klikom na posamezno gručo. S klikom se pod sliko izpiše seznam kontekstov (tj. stavkov), ki sodijo v določeno gručo, njihov vir ter obdobje, iz katerega izhajajo.

5 PRIMER UPORABE: TEMA MIGRACIJ

V poglavju 4 smo pokazali, da lahko sistem uporabimo za analizo sprememb v rabi besed v različnih obdobjih, same meje med obdobji pa so bile določene glede na razpoložljive podatke, tj., korpus GigaFida 2.0 smo razdelili na dva ali štiri med seboj najbolj oddaljena obdobja, da smo preverili, kako uspešen je sistem pri zaznavanju dolgoročnih premikov, ter na dve najmlajši letni obdobji v korpusu, da smo preverili, kako uspešno sistem zazna kratkoročne spremembe v rabi.

V tej sekciji nas po drugi strani zanima, kako sta teroristični napad v ZDA 11. septembra 2001 in obdobje "begunske krize" vplivali na reprezentacijo fenomena migracije v slovenski družbi. V ta namen smo korpus GigaFida razdelili na pet obdobij, predstavljenih v tabeli 2[5]:

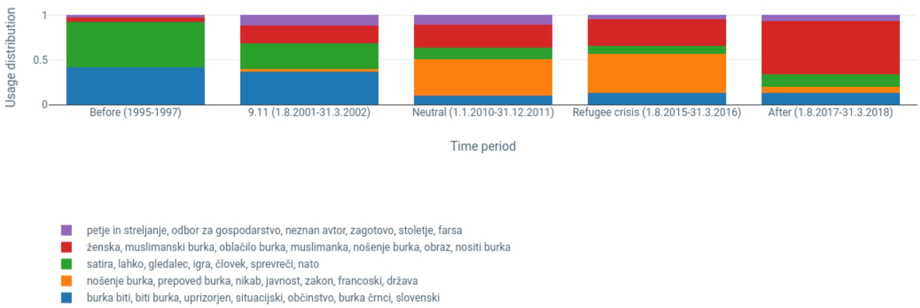
1. pred-obdobje (1995-97),
2. čas terorističnega napada v ZDA 11. septembra 2001, ki mu sledi načeloma
3. nevtralno obdobje (2010-11),
4. čas množičnih migracij v Evropi zaradi Zahodno-balkanske poti, najpogosteje poimenovan "begunska kriza" (2015-16) in
5. po-obdobje (2017-18).

Tabela 2: Število dokumentov, besed in virov v posameznih obdobjih določenih na podlagi dogodkov povezanih s fenomenom migracije.

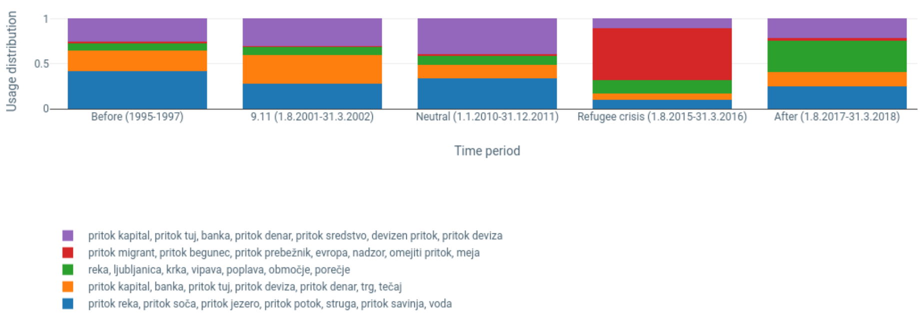
Obdobje	Št. dokumentov	Št. besed	Št. virov
Pred-obdobje (1995-1997)	4.577	43.300.937	18
9.11 (1.8.2001-31.3.2002)	1.198	47.554.430	22
Nevtralno (1.1.2010-31.12.2011)	1.473	33.707.077	90
Begunska kriza (1.8.2015-31.3.2016)	212	21.123.369	8
Po-obdobje (1.8.2017-31.3.2018)	273	26.775.448	8

Med besedami, ki so spremenile rabo med temi petimi obdobji, obravnavamo dva specifična primera, "burka" in "pritok".

[5] Sistem za analizo sprememb v rabi besed med temi petimi obdobji je dostopen na http://kt-nlp-demo.ijs.si:8080/semanticshifttable/6.



Slika 3: Sprememba v rabi besede burka.



Slika 4: Sprememba v rabi besede pritok.

Slika 3 ponazori spremembo v rabi besede “burka”. Medtem ko je v obdobju pred napadi v ZDA 11. septembra 2001 razumevanje besede povezano predvsem z norčavim vedenjem oziroma gledališčem v smislu dramskega dela s šaljivo vsebino, raba besede v smislu muslimanskega ženskega oblačila oziroma pokrivala skozi pet obravnavanih obdobij narašča in je najpogostejša v zadnjem obdobju (2017-2018). V tem primeru je potrebno 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 posledic 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 socilingvistično pogojen, zaznava tega prevzema prostora s strani sistema pa pokaže, da je sistem mogoče uporabiti tudi za sociološko analizo.

Uporaba besede burka v smislu ženskega oblačila 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 npr. velja v Franciji vse od leta 2011). Ta vidik je bil pričakovano najbolj izpostavljen v obdobju po napadu na ZDA, kateremu je sledila napoved t. i. vojne proti terorizmu (ang. the war on terror), ter v času t. i. begunske krize, ko je ozemlje Slovenije prečkalo skoraj 500.000 beguncev, za katere se je predvidevalo, da so muslimanske veroizpovedi. Najbolj obsežna pa je uporaba besede burka v smislu ženskega oblačila v obdobju po ključnih dveh časovnih točkah v po-obdobju, kar sovpada z globalnim porastom razprave o migracijah kot problemu, predvsem zaradi domnevne nezdružljivosti islama z zahodno oziroma evropsko (in slovensko) kulturo (Kundnani, 2015). V zadnjem obdobju tako pri rabi besede prevladuje vidik spola, razprava se osredotoči na muslimansko žensko (Farris, 2017).

Zanimiva je tudi sprememba v rabi besede "pritok" (gl. sliko 4). Od prevladujoče povezave »pritoka« z vodo (npr. pritok reke) v drugi polovici 1990. let, preko navezave na denar, banke in devize (npr. pritok kapitala) v drugem (in tudi tretjem) obdobju, je očiten porast v rabi v povezavi z migracijami skozi "pritok migrantov, beguncev, prebežnikov" v času "begunske krize". V tem obdobju je sprememba v rabi povezana s političnim dogajanjem v Evropi, kjer v ospredje pride problematika omejevanja in upravljanja migracij ter preprečevanje vstopa beguncem, kar potrjujejo vse obstoječe raziskave medijskega poročanja (gl. npr. Pajnik (2017)). V po-obdobju tega več ni, se pa spet okrepi povezava z vodo in rekami.

6 ZAKLJUČEK IN NADALJNJE DELO

V članku smo predstavili prvi spletni sistem za zaznavanje sprememb v rabi besede v slovenščini. S pomočjo uporabniškega vmesnika sistem omogoča temeljito časovno analizo sprememb na ravni korpusa in na ravni posameznih besed. Medtem ko trdimo, da je sistem uporaben za sociolingvistično analizo, je potrebno opozoriti tudi na omejitve sistema. Kot prvo, sistem v veliko primerih prikaže več gruč pogostih rab, kot jih dejansko obstaja. Ta problem izhaja iz vektorskih vložitev, ki poleg semantike besed upoštevajo tudi sintakso, zaradi česar se v veliko primerih ustvari več gruč, ki pokrivajo semantično isto rabo besede, razlike med njimi pa temeljijo izključno na sintaksi.

Druga omejitev sistema izhaja iz podatkov, saj je povsem možno, da na spremembe v rabi določenih besed vplivajo razlike v sestavi virov posameznih časovnih podkorpusov. Iz teh razlogov je potrebno biti pri interpretaciji rezultatov, ki jih poda sistem, previden.

Prav tako bi radi opozorili na dejstvo, da je glavni namen članka predstavitev novega javno dostopnega sistema za zaznavanje in analizo sprememb v rabi besed v slovenščini, zaradi česar je na sami analizi posameznih izbranih besed manj poudarka. Prikazani primeri spremembe rabe so bili izbrani predvsem zato, ker dobro prikažejo, kakšne spremembe je sistem sposoben zaznati. Za celovito sociolingivstično analizo bi po drugi strani bilo potrebno primere izbrati bolj sistematično (tj. glede na dejanski rang spremembe) in raziskati sistematičnost sprememb v rabi kot globlje značilnosti obravnavanega obdobja. Bolj celovito sociološko raziskavo načrtujemo v prihodnosti.

Poleg tega, bomo v prihodnje sistem uporabili na novejših tekstovnih korpusih v slovenščini in na ta način pridobili podatke o spremembah o rabi besed po letu 2018 ter na novih študijah primerov (npr., kako je na evolucijo raznovrstnih konceptov v slovenščini vplivala pandemija COVID, ki je glede na raziskave imela odločilen vpliv na evolucijo medijskega poročanja v angleško govorečih državah (Montariol in sod., 2021)). Prav tako bomo preizkusili nove metode za zaznavanje in interpretacijo sprememb v rabi besed, in na ta način poskušali izboljšati delovanje sistema. In nazadnje, osredotočili se bomo tudi na metode za odkrivanje skupine konceptov, ki izražajo podobno spremembo v rabi.

ZAHVALA

Delo je bilo izvedeno v okviru projekta RSDO ter programov in projektov Javne agencije za raziskovalno dejavnost Republike Slovenije: Sovražni govor v sodobnih konceptualizacijah nacionalizma, rasizma, spola in migracij (J5-3102), Tehnike vektorskih vložitev za medijske aplikacije (L2-50070), Računalniško podprta večjezična analiza novičarskega diskurza s kontekstualnimi besednimi vložitvami (J6-2581), Tehnologije znanja (P2-0103).

LITERATURA

Aitchison, J. (2001). Language change: Progress or decay? V (pogl. The ever-whirling wheel: The inevitability of change). Cambridge: Cambridge University Press.

Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M. in Kamps, J. (2017). Words are malleable: Computing semantic shifts in political and media discourse. V *Proceedings of the 2017 acm on conference on information and knowledge management* (str. 1509–1518).

Basile, P., Caputo, A., Caselli, T., Cassotti, P. in Varvara, R. (2020). Diacr-ita @ evalita2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task. *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*. https://api.semanticscholar.org/CorpusID:229292864

Devlin, J., Chang, M.-W., Lee, K. in Toutanova, K. (2019, June). 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)* (str. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1423 doi: 10.18653/v1/N19-1423

Farris, S. R. (2017). *In the name of women's rights: The rise of femonationalism*. Duke University Press. Pridobljeno 2024-05-31, http://www.jstor.org/stable/j.ctv11sn2fp

Gantar, P., Holdt, Š. A. in Pollak, S. (2018). Leksikalne novosti v besedilih računalniško posredovane komunikacije. *Slavisticna Revija*, *66*(4), 459–472.

Gillani, N. in Levy, R. (2019). 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* (str. 94–99).

Giulianelli, M., Del Tredici, M. in Fernández, R. (2020, July). Analysing lexical semantic change with contextualised word representations. V *Proceedings of the 58th annual meeting of the association for computational linguistics* (str. 3960–3973). Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.acl-main.365

Hamilton, W. L., Leskovec, J. in Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. V *Proceedings of the 54th annual meeting of the association for computational linguistics* (str. 1489–1501). http://aclweb.org/anthology/P16-1141 doi: 10.18653/v1/P16-1141

Hilpert, M. in Gries, S. T. (2008). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, *24*(4), 385–401.

Hopper, P. J. in sod. (1991). On some principles of grammaticization. *Approaches to* *grammaticalization*, *1*, 17–35.

Juola, P. (2003). The time course of language change. *Computers and the Humanities*, *37*(1), 77–96.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D. in Petrov, S. (2014). Temporal analysis of language through neural language models. V *Proceedings of the acl 2014 workshop on language technologies and computational social science* (str. 61–65). http://aclweb.org/anthology/W14-2517 doi: 10.3115/v1/W14-2517

Krek, S., Holdt, Š. A., Erjavec, T., Čibej, J., Repar, A., Gantar, P., ... Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard slovene. V *Proceedings of the 12th language resources and evaluation conference* (str. 3340–3345).

Kundnani, A. (2015). *The muslims are coming: Islamophobia, extremism, and the domestic war on terror*. Verso.

Kutuzov, A. in Giulianelli, M. (2020, December). UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. V *Proceedings of the fourteenth workshop on semantic evaluation* (str. 126–134). Barcelona (online): International Committee for Computational Linguistics. https://www.aclweb.org/anthology/2020.semeval-1.14

Kutuzov, A., Pivovarova, L. in sod. (2021). RuShiftEval: a shared task on semantic shift detection for russian. V *Computational linguistics and intellectual technologies: Papers from the annual conference dialogue.*

Kutuzov, A., Velldal, E. in Øvrelid, L. (2017, August). Tracing armed conflicts with diachronic word embedding models. V *Proceedings of the events and stories in the news workshop* (str. 31–36). Vancouver, Canada: Association for Computational Linguistics. https://www.aclweb.org/anthology/W17-2705 doi: 10.18653/v1/W17-2705

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, *37*(1), 145–151.

Martinc, M., Kralj Novak, P. in Pollak, S. (2020, May). Leveraging contextual embeddings for detecting diachronic semantic shift. V *Proceedings of the twelfth language resources and evaluation conference* (str. 4811–4819). Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.592

Martinc, M., Montariol, S., Zosa, E. in Pivovarova, L. (2020). Capturing evolution in word usage: Just add more clusters? V *Companion proceedings of the web conference 2020* (str. 343–349). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3366424.3382186 doi: 10.1145/3366424.3382186

Martinc, M., Perger, N., Pelicon, A., Ulčar, M., Vezovnik, A. in Pollak, S. (2021). 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* (str. 121–126).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. in Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Montariol, S., Martinc, M., Pivovarova, L. in sod. (2021). 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.*

Pajnik, M. (2017). Medijsko-politični paralelizem. legitimizacija migracijske politike na primeru komentarja v časopisu delo. *Dve domovini / Two Homelands*, *45*, 169-184.

Pranjić, M., Dobrovoljc, K., Pollak, S. in Martinc, M. (2024). Semantic change detection for slovene language: a novel dataset and an approach based on optimal transport. *arXiv preprint arXiv:2402.16596*.

Rosin, G. D., Guy, I. in Radinsky, K. (2022). Time masking for temporal language models. V *Proceedings of the fifteenth acm international conference on web search and data mining* (str. 833–841).

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H. in Tahmasebi, N. (2020, December). SemEval-2020 task 1: Unsupervised lexical semantic change detection. V *Proceedings of the fourteenth workshop on semantic evaluation* (str. 1–23). Barcelona (online): International Committee for Computational Linguistics. https://www.aclweb.org/anthology/2020.semeval-1.1

Solomon, J. (2018). Optimal transport on discrete domains.

Stewart, I., Arendt, D., Bell, E. in Volkova, S. (2017). Measuring, predicting and visualizing short-term change in word representation and usage in VKontakte social network. V *Eleventh international aaai conference on web and social media.*

Tahmasebi, N., Borin, L. in Jatowt, A. (2018). Survey of computational approaches to diachronic conceptual change. *arXiv preprint arXiv:1811.06278*.

Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y. in Hengchen, S. (Ur.). (2021). *Computational approaches to semantic change* (Št. 6). Berlin: Language Science Press. doi: 10.5281/zenodo.5040241

Tang, X. (2018). A state-of-the-art of semantic change computation. *Natural Language Engineering*, *24*(5), 649–676.

Terčon, L. in Ljubešić, N. (2023). Classla-stanza: The next step for linguistic processing of south slavic languages. *arXiv preprint arXiv:2308.04255*.

Ulčar, M. in Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model.

Zamora-Reina, F. D., Bravo-Marquez, F. in Schlechtweg, D. (2022, May). LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. V N. Tahmasebi, S. Montariol, A. Kutuzov, S. Hengchen, H. Dubossarsky in L. Borin (Ur.), *Proceedings of the 3rd workshop on computational approaches to historical language change* (str. 149–164). Dublin, Ireland: Association for Computational Linguistics. https://aclanthology.org/2022.lchange-1.16 doi: 10.18653/v1/2022.lchange-1.16

A SYSTEM FOR WORD USAGE CHANGE DETECTION AND ITS USAGE FOR SOCIOLINGUISTIC ANALYSIS

We present the first web-based publicly available system for detection of word usage changes in Slovene, a less resourced Slavic language with two million speakers. We present its specifications, graphical user interface, and the methodology for usage change detection. We explain how the system can be used for ranking words in the vocabulary according to the usage change, and for finding and analysing words that exhibit significant change in usage. By employing the system on several different Slovenian corpora, we showcase that the system is capable of detecting long-term and short-term usage changes. Finally, we show that the system can also be used for analysis of social representation of migrations, by employing it on the reference Slovenian corpus GigaFida 2.0 with manually defined splits according to the events that have significantly influenced the attitude towards migration and migrants in Slovenia. The system is publicly available at http://kt-nlp-demo.ijs.si:8080.

Keywords: Word usage change detection, sociolinguistics, migration

UNLOCKING HISTORY: A REDESIGN OF THE SISTORY 5.0 PORTAL

Katja MEDEN,[1,3], Ana CVEK[1], Vid KLOPČIČ[2], Matevž PESEK[2], Mihael OJSTERŠEK[1], Mojca ŠORN[1], Andrej PANČUR[1]

[1]Institute of Contemporary History

[2]Faculty of Computer and Information Science, University of Ljubljana

[3]Department of Knowledge Technologies, Institute "Jožef Stefan"

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 comprises a wide range of historical information, including publications, images and extensive databases as well as comprehensive metadata describing 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 wider audience. This paper presents the process of redesigning the portal in its technical and content enhancements, followed by an in-depth analysis of the content offered by the portal in its current form.

Ključne besede: SIstory, redesign, information systems, metadata, historiography

1 INTRODUCTION

The beginnings of the Research Infrastructure of Slovenian Historiography (RI INZ) date back to September 2008, when the online research and education portal of Slovenian historiography, *History of Slovenia - SIstory*, was launched. Its main content at the time consisted of a combination of historiographical research, historical sources and technical infrastructure outputs. Its main aim was to provide digitised and freely accessible research results and sources, with a focus on preserving older and not-so-easily accessible historical sources and thus preserving cultural and scientific heritage (Šorn et al., 2011). The operation and design of the SIstory portal were based on supporting the research process of the Institute's research community members. With the very successful initial response from the public and related institutions, SIstory gradually moved beyond the boundaries of “written history” and developed interactive presentations of historical content, supported by the development of the then-new technologies (Šorn et al., 2011). While the content and initial design of the portal seemed to be synonymous with the concept of a “digital library”, SIstory quickly became more than that. With the data preservation and integration of the research community at the forefront of SIstory’s development, it became the main venue of a newly established Slovenian national node of the DARIAH research infrastructure family, DARIAH-SI, with the Institute (and more specifically RI INZ) as the national coordinating institution (Pančur & Šorn, 2019).

The redesign presented in this paper focuses on the technological advances that are now more accessible in order 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 visual exploratory analysis of the collected data. Finally, we are focussing our efforts on the transparency and reuse of the data. In accordance with the European Commission’s Data Act (*Data Act*, 2024), we have redesigned the portal to make all collected data fully accessible and to make the interaction more user-friendly.

The rest of the paper is structured as follows: Section 2 outlines the state of the portal before and after the redesign, focussing on the state of the portal, the reasons for the redesign and the components that were included in the redesign. Section 3 then provides some basic statistics about the content and presents the analysis of the portal content. Section 4 discusses the results, highlighting notable trends and their significance for the portal. Finally, Section 5 provides an overview of the paper and presents some options for future work.

2 HISTORY OF THE SISTORY

The SIstory portal can look back on a relatively long history of development. As already mentioned, the first steps were taken in 2008 with the initial release and since then several versions of the portal have been published as individual upgrades of the portal. In 2011, the first software and technological upgrade of the portal was carried out, which was a major step forward in Slovenian digital historiography. With this, SIstory not only established the latest standards and enabled faster and more stable operation of the system, but also played an important role in establishing a national digital infrastructure for the humanities and arts (Cvek et al., 2022). The first upgrade consisted of several components (Rožman & Marolt, 2011):

- Content administration in SOLR[1] and upgrading folder structures and file names.
- Implementation of the Dublin Core metadata standard (the schema contained all 15 basic elements of DCMES[2], a year later the original schema was upgraded with elements of the qualified DCMI Metadata Terms. (DCTERMS)[3].
- Creation of a unique and permanent URN—Uniform Resource Name.
- Introduction of the Sphinx metadata search engine[4]. Two search engines were implemented: basic and advanced.
- Upgrading of the portal administration.
- Design of the structure and access levels for users.

As the DCTERMS element set was no longer sufficient to adequately describe the different types of information sources, the portal was updated in 2013 to develop and integrate the SIstory metadata schema, a customised set of metadata elements that better matched the nature of the then-current content. The next major update of the portal took place in 2016. A mapping between the SIstory metadata schema and the DC was created to enable better interoperability of the data. The SIstory metadata schema was then extended with elements and structures from the HOPE application profile (a well-established profile in the GLAM community) (Lemmens et al., 2011) to develop the *SIstory application profile* (Pančur, 2013), which has remained the metadata set of choice for the SIstory portal over the years. The application profile and its implementation led to the structure, syntax and semantics of the metadata input tool.

Besides the metadata enrichment, new frameworks and graphical templates of the system, the graphical interface of both the administration and the user interface were also installed. In addition, the search engine (filtering and sorting of results; full-text search) was also taken into account in the updates (Cvek et

[1] Apache Solr
[2] Dublin Core Metadata Element Set
[3] DCMI Metadata Terms
[4] Sphinx search engine

al., 2022). Overall, since the portal's inception in 2008, a series of upgrades have been made, each improving the portal's functionalities and features. This brings us to today and to new steps in the portal's development – the decision to redesign the portal from the ground up.

3 SISTORY: THE REDESIGN

Immediately before the redesign, the basis of the portal consisted of various modules that were responsible for entering, processing and storing a collection of entities enriched with metadata. More specifically, the base consisted of:

- MySQL relational database
- the non-relational Elasticsearch database to search and index the content of collections and sources
- PHP 7
- File management module (extraction of raw text from PDF, HTML, XML, DOCX files, generation of previews)
- The metadata schema followed the previously mentioned SIstory application profile schema, developed for the purpose of flexible descriptions of various entities.

However, as the portal was updated several times, with each update based on the new versions of the same technology, this posed a problem as the code became too vast to manage efficiently. Furthermore, the concepts and various other solutions that were developed over the years were very ambitious and necessary for the time. Nevertheless, they did not prove to be as useful in the practical day-to-day operation of the portal as originally thought. This along with the outdated appearance of the user interface were decisive factors in our decision to start from scratch. When planning the redesign, we took into account the legacy issues and solutions from previous versions of the portal in order to improve the functionality of the system and provide a familiar user experience.

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

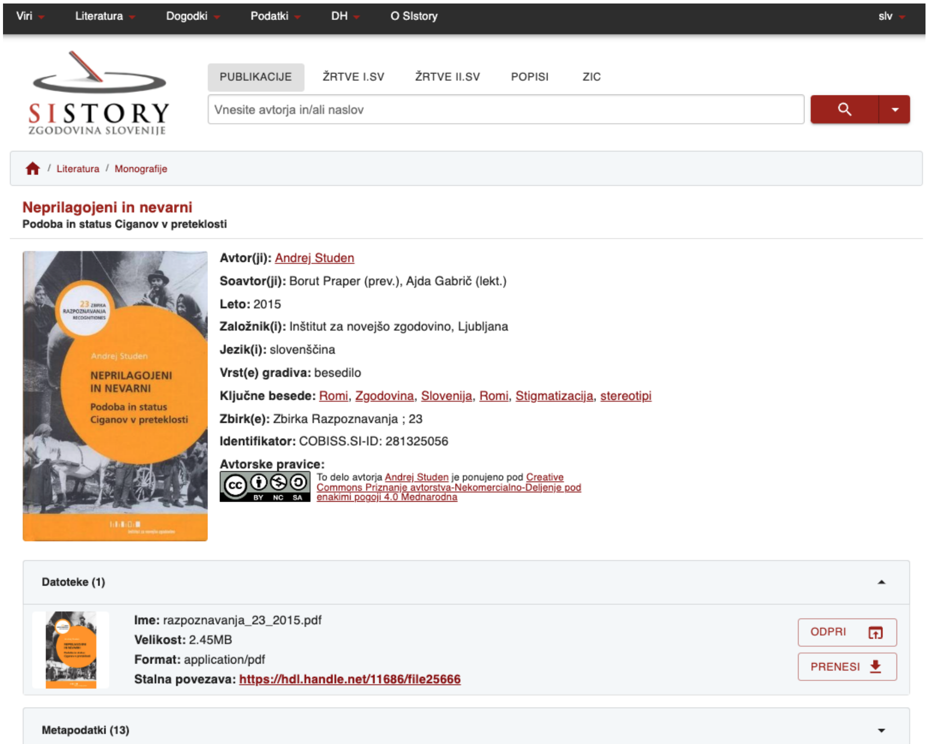


Figure 1: The new SIstory UI

3.1 Technical design

In terms of the technical composition, the redesigned SIstory 5.0 portal is based on a robust technical framework, while the backend utilizes Django version 5.0.4, based on Python 3.11, for efficient data management and content delivery. The server infrastructure is supported by nginx version 1.25.

On the frontend, SIstory employs Next.js version 14.1.0 and React version 18.2.0 in combination with Node.js v18 for dynamic and interactive user interfaces. This modern frontend stack enables smooth navigation and responsive design across various devices and improves accessibility for users accessing historical content. Figure 1 shows an example of the redesigned UI.

The portal's database architecture is based on PostgreSQL version 16 and provides a robust foundation for storing and retrieving large volumes of historical data with high speed and reliability. In addition, SIstory integrates Matomo version 5.0.3, an analysis function that provides administrators to gain valuable insights into user behaviour and interaction patterns, thus forming the basis for future developments and improvements.

For an efficient search functionality, SIstory incorporates Elasticsearch version 8.9.0 and Kibana version 8.9.0, so that users can quickly locate relevant historical documents and sources. The use of Elasticsearch ensures fast and accurate search results and improves the overall usability of the portal.

In addition, SIstory employs Handle[5] system to enable permanent identifiers that provide reliable and permanent access to specific historical documents and sources. This ensures that users can reliably reference and cite the materials, 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.

3.2 Metadata design

The portal previously used the SIstory application profile (SIstory AP) as the basis for encoding the metadata, with the HOPE application profile serving as the basis. In practice, this posed a problem as SIstory AP contained several elements (and element groups) that were not used as frequently as originally assumed. This in turn led to a simplification of the profile. To this end, an analysis of the existing SIstory AP was carried out, with the main aim of identifying metadata elements that should be retained and addressing elements 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 (i.e. DC is the base standard for the OAI-PMH metadata harvesting protocol)[6]. A very limited number of elements of the

[5] Handle System identifier
[6] OAI-PMH protocol for SIstory 5.0 is at the time of writing still in the implementation phase.

namespace *sistory* have been retained[7], mainly due to the remnants of the older publications described with these specific metadata elements. The overview of the major metadata groups is presented in Table 1.

Metadata	**Unique Values**	**Nr. of Usages**
No. of entries	57,263	
Creator	5,146	24,908
Subject	23,720	318,321
Publisher	1,036	53,699
Collection	426	1,861
Contributor	1,290	42,885
Type	12	58,868
Language	60	69,490

Table 1: Overview of the most important metadata groups, the number of unique instances and the total number of occurrences in the SIstory portal (at the time of creation).

In total, SIstory comprises over 57,000 unique entries, and over 5,000 unique authors/physical persons (under the category Creator)[8], while "*Subject*" contains the keywords that describe the publications. The secondary forms of authorship are described in the category "*Contributor*" (e.g. editor, translator...), while the type of publication on the basis of the controlled vocabulary (DCMI Type [9]) encompassing 12 categories. Finally, the portal includes publications in 60 different languages, which are presented in more detail in Section 4.3.

4 SISTORY UNVEILED: CONTENT ANALYSIS

In the efforts to present the redesign of the SIstory portal, it became clear that focusing on the mainly 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

[7] For example, *SIstory Unstored* – an element/field for storing metadata that cannot be stored in any other metadata field due to its content.

[8] In the metadata mask there are two separate fields for a Creator, which according to the definition of Dublin Core can be either a physical person or a legal "organisation". Under the category "*Creator*" in the Table 1 only the occurrences for a physical person/author are counted

[9] DCMI Type Vocabulary

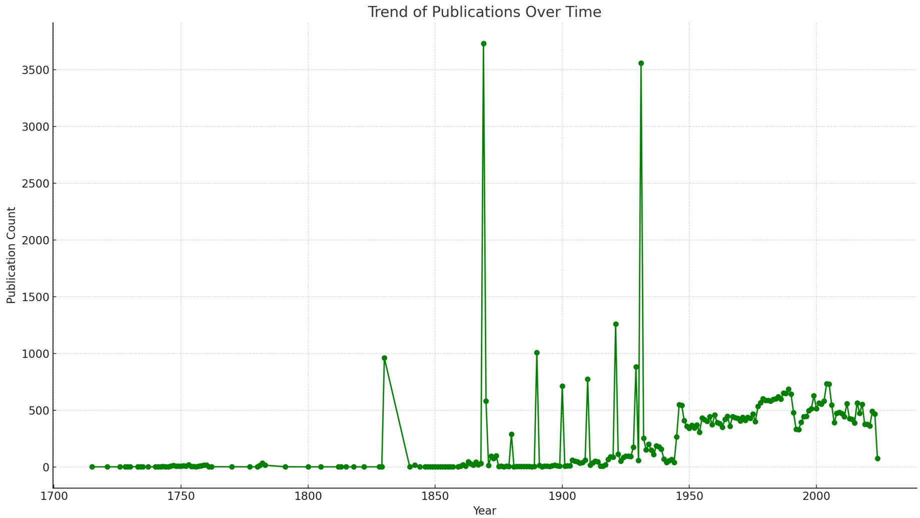


Figure 2: Trends of publications over time.

content analysis to gain a better understanding of the content available on the portal. In the analysis, we focused on different aspects of the portal's content, from the basic statistics of the main metadata groups (not the individual metadata elements) to analysing the keywords and language of the publications available on the portal. This in turn allows us to show not only the scope of content available, but also to highlight the different types and variations of this content.

In order to gain insight into the content available on the portal, we have focused on analysing the metadata of the publications. More precisely, we analysed the following specific metadata groups:

- Publication date (date when a specific publication was first published, i.e. in general, not on the portal itself),
- content menus (especially the first and second menu levels),
- language of the publication (according to ISO 639-2 standard),
- and the publication keywords (e.g. Jugoslavija, učbeniki...).

This in turn allowed us to capture some of the trends of publications for different time periods, more specifically, the distribution of publications over time, keyword analysis and language trends of publications.

4.1 Trends of publications

One of the trends analysed in this section is the distribution of publications over time according to their publication date, in order to check which years are best represented in terms of published content. The results are shown in Figure2.

One of the first trends to emerge is the distribution of publications within the period from around 1715 to the mid-20th century, which showed several severe spikes in the number of publications. Conversely, the post-World War II period shows a much more even flow. The spikes in the timeline are most likely due to several large works or specific publications types (e.g. textbooks, population censuses) being published on the portal, while the steadiness of the publication flow in the post-war period to the present day seems to indicate a greater variance in the types of publications (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 (presented in Section 4.2) to provide additional insight and substantiate the reasons for the trends identified.

4.2 Keyword analysis

The keyword analysis of the portal content, in which the 10 most frequent/representative keywords used to describe the sources in the individual menus were examined for each of the 20 years (with the exception of the period 2010 - 2024). This was done 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[10].

[10] Some of the keywords are very similar to one another, which is mostly due to slightly different versions of the notation format. For example, "*uradni list*" and "*uradni listi*" are singular and plural versions of the same keyword, but are counted separately.

4.2.1 SOURCES—TOP 10 KEYWORDS

Table 2 gives an overview of the 10 most frequent keywords found for each 20-year period for the menu “Sources”. This category covers different types of resources, specifically archival, oral, and printed resources as well as digitized versions of physical objects. The latter are mostly images of physical objects, such as statues or death masks) and printed sources.
The keywords in the 18th century period (or, more precisely, in the period from 1710 – 1790), the trend seems to consist of several repeating keywords, such as “*patenti*” (patents), “*odloki*” (decrees), “*norme*” (norms), “*Marija Terezija*”, which refer directly to one of the more extensive collections of publications entitled “*Collection of various patents, decrees, ordinances, norms, instructions, etc., issued by Charles VI, Maria Theresa and Joseph II*”[11], which was acquired through our cooperation with the related institution, the Central Judicial Library. In addition to this collection, some of the keywords also refer to a large number of population censuses (as indicated by the keywords “*popisi prebivalstva*” (population censuses) and “*občina*” (municipality)) published for this particular period. This also applies to most of the 19th century, where, in addition to the census, large amounts of theatre lists of various regional theatres (e.g. the Regional Theatre in Ljubljana), which in the past provided information on current plays or other related events. In addition to theatre lists and censuses, the calendars of the Society of St. Mohor (annual publication containing a calendar, religious prayers, illustrations, poetry, etc.)[12] were found in the keywords “*koledar*” or “*Družba sv. Mohorja*” (Society of St. Mohor), which were very popular at the time. Lastly, several keywords indicate a high number of minutes of the Carniolan Regional Assembly, as shown by the keywords “*Kranjska*”, “*Carniola*”, “*deželna avtonomija*” or “*provincial autonomy*”.

At the beginning of the 20th century, the most frequent publications uploaded to the portal are initially indicated with the already known keywords *popisi prebivalstva* and *občina*, which represent censuses, with several others, such as *uradni listi* (official gazettes), which refer to the official gazettes from different periods, republics and countries[13] (e.g. Slovenia, Yugoslavia, Serbia, Bosnia and Herzegovina, all of which are included in the keywords for this period) and

[11] Example of a Josef II directed patent
[12] Example of the St. Mohor calendar
[13] Official gazettes

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
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

Table 2: Top 10 Keywords by Two-Decade Period for Viri (Sources)

stenografski zapisniki (stenographic records), which represent the minutes of various executive and legislative bodies, which were among the more prevalent publications and for which, the number of such publications only intensified within this period.

Finally, for the more recent period (2010 – 2024), the keywords refer mainly to studies carried out in connection with the censuses of Slovenia from 1830 – 1931, which are the result of cooperation with the Historical Archive of Ljubljana.

4.2.2 LITERATURE—TOP 10 KEYWORDS

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
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

Table 3: Top 10 Keywords by Two-Decade Period for Literatura (Literature)

Similarly, Table 3 shows the 10 most frequent keywords for a single 20-year period for the Literature menu, which consists of publications such as research monographs, (Slovenian) serial publications on historiography — together 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 predominantly dominated by the textbooks produced as part of the projects "*Šolski listi*" and "*Schools and Imperial, National, and Transnational Identifications: Habsburg Empire, Yugoslavia, and Slovenia*"[14], an extensive digitization project of textbooks primarily intended for schools on various school subjects identified in the table with the following keywords: *učbeniki* (textbooks), *abecedniki* (abecedarium), *matematika* (mathematics), *čitanke* and *berila* (reading material), etc. In the early to mid-20th century, however, the topics are then expanded to include additional material on the topics of politics, political programmes and political parties, indicated by the keywords *politični programi* (political programmes), *katoliški tabor* (Catholic camp), *liberalni tabor* (Liberal camp). For the second half of the 20th century, the themes shift to the Second World War, more precisely to the role of Yugoslavia (and Slovenia) in the Second World War (keywords). Directly related to this is also a considerable amount of literature referring to archival sources (keyword *arhivsko gradivo*)— mostly in connection with a specific journal, *The Gazette of the Archival Association and Archives of Slovenia*. Lastly, a very prominent keyword, *ocene in poročila* (reviews and reports), refers to a very specific form of contributions to various Slovenian (scientific) journals, in which the authors of the contributions give their reviews of various published works on the topics of the journal (in this case, mainly history).

4.2.3 EVENTS—TOP 10 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 and digitization of various exhibitions related to the field of historiography, the Institute or related institutions.

The first difference between the Tables 2 and 3 is the significantly shorter time period, which is not particularly surprising, however, the portal only exists from

[14] Schools and Imperial, National, and Transnational Identifications: Habsburg Empire, Yugoslavia, and Slovenia

Decade Range	Top Keywords
1990-2009	Kranjska, Ljubljana, Slovenija, dekleta, dokumentarni filmi, izobraževanje, univerze, zgodovina, študenti
2010-2024	Središče za javno zgodovino, Filozofska fakulteta, Oddelek za zgodovino, zgodovina, Slovenija, Jugoslavija, šolstvo, muzej, druga svetovna vojna, video

Table 4: Top 10 Keywords by Two-Decade Period for Dogodki (Events)

2008 onwards—this is also directly related to the relatively unrepresentative keywords within the period 1990 – 2009, as there are only a few individual publications related to this period (specifically, there are only 2 such publications). The topics of these publications are directly related to the topic of girls' education in Ljubljana[15] and Slovenian students abroad[16]. However, the number of publications increases 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).

4.2.4 DATA AND DH—TOP 10 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 than of a large part of the portal.

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

Table 5: Top 10 Keywords by Two-Decade Period for Podatki (Data)

[15] Šola naših babic: izobraževanje deklet v Ljubljani
[16] Študenti s Kranjske na avstrijskih in nemških univerzah 1365 – 1917

The Table 5 is mainly limited to research data in the field of historiography, more specifically, rich data on the old place names in Slovenia[17] and censuses of the Jewish population in Slovenia[18].

While the categories examined are generally research data, the DH category relates primarily to the expanding vision of the RI INZ at the time – expanding into the field of digital humanities and providing data and tools to support research activity in these (related) fields. This later led to the development of a separate repository for digital humanities, the SI-DIH repository, another product of RI INZ and DARIAH-SI.

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

Table 6: Top 10 Keywords by Two-Decade Period for DH

As these were the first steps of the infrastructure towards DH, there are only limited publications and tools available, but they incorporate the technologies of the time— this is also reflected in the keywords in Table 6, such as *nadgrajena resničnost* (augmented reality), *XML shema* (XML schema), *metapodatki* (metadata), *HTML*, *TEI*.

4.3 Language trends

In addition to the keyword analysis, we also took a look at 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 *History of Slovenia – SIstory.si* is Slovene, with a total of 46937 occurrences, most common in the period 1970 – 2010, especially between 1990 – 2010. The second most frequent language is Serbian (8072 occurrences), although an explicit distinction must be made here, as Serbian also belongs to two other language categories: Serbian (Cyrillic) for publications in Cyrillic script (912 in total) and the Bosnian/Croatian/Serbian category for publications where the language could not be explicitly identified (mostly publications referring to the *Official Gazettes* of Yugoslavia), which are the most frequent in publications.

[17] Place names in Slovenia 1.0

[18] List of Jews in Slovenia (Dravska banovina), 1937

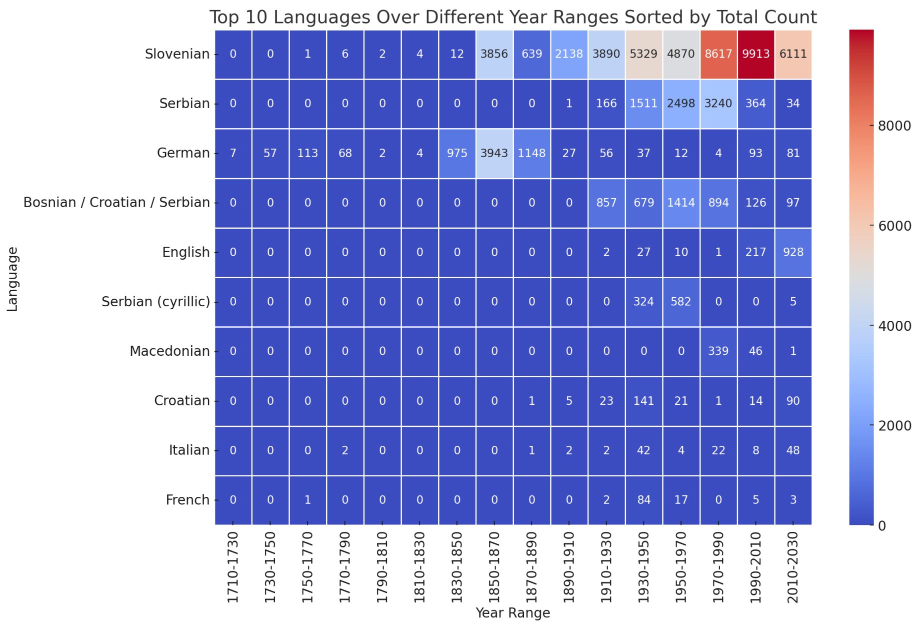


Figure 3: Language trends – distribution of publication languages over time.

Publications with all of the above categories for Serbian were published most frequently in the period from 1950 to 1990, again mainly in connection with the official gazettes. The last very frequent publication language is German, especially for the 19th century, more precisely for the period 1830 – 1890, which includes older publications related to the history of Slovenia in the mentioned period, when Slovenia was a part of the Austrian Empire (between 1804 – 1867) and Austria-Hungary (1867 – 1918), where the publications were written in both German and Slovene. It is somewhat less prominent in the 18th and the beginning of the 19th century (1710 – 1830), but it is much more emphasised in relation to Slovene). While Table 4.3 highlights the 10 most common languages, the portal also includes publications (and languages) that are not so common (i.e. less than 5 times): Spanish - Castilian, Latin, Polish, Arabic, Albanian, Ukrainian, Esperanto. In addition, several publications are multilingual (42 publications), while some units of the portal have no language. This applies in particular to our digitised collection of death masks entitled “*The Casting of Death*”[19].

[19] The Casting of Death

5 CONCLUSIONS

In this paper, we have presented the technical, visual and content-related changes that have been implemented in the new SIstory 5.0 portal. The technical changes have been noted in the feedback from portal users as a better user experience due to its responsiveness and speed in terms of reduced page load.

The purpose of visualising the data in this paper served as a medium to show the current trends in data collection to bring the collected data closer to the general public. While the portal's content has recently been the focus of a study (Šorn & Cvek, 2023), the emphasis there was on the chronological additions to the SIstory portal over the course of its existence. Still, no metadata-based analysis has been conducted to help us understand the content coverage and themes of the portal. The redesign therefore presented us with the perfect opportunity to familiarise ourselves better 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 of publications over time, provided an overview of the distribution of content and outlined the likely reasons for this. These were then further explored within the keyword analysis, which revealed to some extent that the type of publications within the peaks of the graph corresponded well with the hypothesised reason for such a distribution of content over time (i.e. large volumes of publications of the same type such as textbooks and censuses). This also applies to the language analysis, which mainly served to give us 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 19th century publications were predominantly in German, with some exceptions in Slovene, while in the 20th century the language coverage started to expand to other South Slavic languages (again with Slovene as the dominant language).

Our future work will consist of an in-depth analysis of the internal technical process of data collection and a possible expansion of the scope of the metadata. The current state of the redesigned portal will serve as a basis for our future work, directly involving our community in the process, which has already been one of the cornerstones in the development of the portal. In the future, we want to involve our users directly in the development process from the very beginning (and gather their feedback) regarding the planned future features. One of these possibilities is the integration of visualisations as tools that allow researchers and other users to more easily interact and work with the data within the portal. In addition, we will focus on existing collections to expand the data and provide an equivalent platform for accessible and reusable modern historical sources.

ACKNOWLEDGMENTS

The paper was written in the framework of the Slovenian Research and Innovation Agency-funded research infrastructure DARIAH-SI, the infrastructure programme I0-0013 "The Research Infrastructure of Slovenian Historiography infrastructure programme", and research programme P6-0436 "Digital humanities: resources, tools and methods".

REFERENCES

Cvek, A., Ojsteršek, M., & Šorn, M. (2022). *Izhodišča metapodatkovnih sistemov portala Zgodovina Slovenije – SIstory (2008–2016)*. Ljubljana: Inštitut za novejšo zgodovino. (ISBN 0 471 40300)

Data act. (2024). Retrieved 28. maj 2024, from https://digital-strategy.ec.europa.eu/en/policies/data-act

Lemmens, B., Janssens, J., Dyck, R. V., Bardi, A., Manghi, P., Beving., E., ... Straube, A. (2011). *Hope - the common HOPE metadata structure, including the harmonisation specifications (d2.2)* (Tech. Rep.). Project report, HOPE, Deliverable D2.2, 2011.

Pančur, A. (2013). *Metapodatki portala Zgodovina Slovenije - SIstory* (Tech. Rep.). Inštitut za novejšo zgodovino.

Pančur, A., & Šorn, M. (2019). Na začetku je bil SIstory: Raziskovalna infrastruktura slovenskega zgodovinopisja. In *Inštitut za novejšo zgodovino : 60 let mislimo preteklost* (pp. 47–58). Ljubljana, Slovenia. (https://hdl.handle.net/11686/46230)

Rožman, B., & Marolt, G. (2011). *Analiza podatkov in postavitev standardov in infrastrukture za nadgradnjo portala SIstory.si.*

Šorn, M., Pančur, A., & Sunčič, M. (2011). SIstory: arhivsko gradivo in e-humanistika. *Arhivi*, *34*(1), 145.

Šorn, M., & Cvek, A. (2023). *Vsebine in njihovo razporejanje na portalu Zgodovina Slovenije - SIstory (s poudarkom na publikacijah)*. Inštitut za novejšo zgodovino.

ODPIRANJE ZGODOVINE: PRENOVA PORTALA SISTORY 5.0

Portal Zgodovina Slovenije - SIstory.si predstavlja pomembno interdisciplinarno zbirko publikacij, podatkov, zbirk in metapodatkov, predvsem na področju zgodovinopisja. V zbirki portal zajema širok spekter zgodovinskih publikacij ter metapodatke, ki podatke opisujejo. Nedavna prenova portala SIstory je bila osredotočena na prizadevanja, da bi podatke ponudili ne le kot zbirko zgodovinskih publikacij, temveč tudi omogočili večjo preglednost, interoperabilnost in dostopnost raziskovalnih podatkov širšemu občinstvu, tako raziskovalcem kot laični javnosti. V tem prispevku je predstavljen proces prenove portala v njegovih tehničnih in vsebinskih izboljšavah, ki mu sledi poglobljena analiza vsebin, ki jih ponuja portal v sedanji obliki.

Keywords: SIstory, prenova, podatkovni sistemi, metapodatki, zgodovinopisje

"PARLAMENT JE PO TEORIJI POLJE KONTROLIRANEGA KONFLIKTA": SLOVENSKI PARLAMENTARNI KORPUS SIPARL 4.0

Katja MEDEN,[1,2,3] Tomaž ERJAVEC,[1] Andrej PANČUR[3]

[1] Odsek za tehnologije znanja, Inštitut Jožef Stefan

[2] Mednarodna podiplomska šola Jožefa Stefana

[3] Inštitut za novejšo zgodovino

Razprave nacionalnih parlamentov in drugih zakonodajnih teles, ter s tem parlamentarni korpusi, predstavljajo pomemben vir podatkov za raziskovanje političnega in predvsem parlamentarnega govora ter drugih povezanih pojavov na področju humanističnih in družboslovnih ved. Slovenski parlamentarni korpus siParl vsebuje nabor parlamentarnih razprav slovenskih parlamentarnih teles, ki pokriva obdobje med 1990 in 2022, ter s tem zajema ne le zgolj različna obdobja zgodovine Slovenije, temveč tudi prehod med različnima političnima sistemoma. Korpus ima bogato zgodovino razvoja in je tekom svojega nastajanja močno vplival na razvoj sorodnih korpusov, iniciativ in projektov. V članku predstavimo novo različico 4.0 korpusa siParl, podrobno opišemo razvoj korpusa in novosti, strukturo in označevanje podatkov v korpusu, ter predstavimo naslednje korake pri razvoju korpusa siParl.

Ključne besede: parlamentarni korpus, parlamentarne razprave, slovenski parlament, TEI XML, Parla-CLARIN

1 UVOD

Parlamentarne razprave predstavljajo pomembno povezavo med predstavniki parlamenta (poslanci) in splošno javnostjo, njihova prosta dostopnost pa igra pomembno vlogo v zagotavljanju transparentnosti parlamentarnega procesa in uresničevanju pravice do splošne obveščenosti javnosti. V tem pogledu parlamentarni korpusi predstavljajo bogat in obsežen vir za analizo političnega in parlamentarnega diskurza v okviru različnih raziskovalnih področij. Parlamentarni korpusi obsegajo tako prepise govorjenega jezika kot tudi bogate sociodemografske metapodatke govorcev (Fišer in Pahor de Maiti, 2020, 2021), kot je na primer njihova pripadnost različnim političnim strankam in zakonodajnim organom. Poleg tega parlamentarne razprave v večini držav niso predmet zakonodaje o avtorskih pravicah, oziroma so iz njih izvzete, saj veljajo za javne dobrine. To je v zadnjem obdobju povzročilo hitro rast števila parlamentarnih korpusov, zlasti nacionalnih. Eden izmed njih je tudi *Slovenski parlamentarni korpus siParl*, katerega začetki segajo v leto 2016 pod takratnim imenom *SlovParl*. Skupaj z razvojem sorodnih iniciativ, kot je projekt ParlaMint[1] (Erjavec in sod., 2022), je siParl postavljal pomembne mejnike pri označevanju in razvoju parlamentarnih korpusov.

V tem prispevku predstavljamo novo različico korpusa siParl (4.0), ki obsega transkripte parlamentarnih razprav in razprav delovnih teles Državnega zbora Republike Slovenije v obdobju 1990 – 2022 (1. – 8. mandatno obdobje). V primerjavi s predhodno verzijo korpusa 3.0 (Pančur in sod., 2022) so bile v novo različico dodane manjkajoče razprave delovnih teles za 8. mandatno obdobje. Poleg razširitve korpusa z dodatnimi podatki smo se v tej različici posvetili tudi posodobitvi delotoka priprave korpusa. V sklopu posodobitve smo poskušali zagotoviti čim bolj avtomatiziran proces označevanja in priprave korpusa, od začetne ekstrakcije podatkov do priprave jezikoslovno označene različice korpusa.

Prispevek je strukturiran na naslednji način: poglavje 2 se posveča pregledu zgodovine razvoja korpusa siParl in ob tem opiše in poudari pomembnost sorodnih iniciativ, na katere je korpus siParl pomembno vplival, medtem ko so te iniciative prav tako pomembno prispevale k razvoju korpusa siParl. Poglavje 3 predstavi delotok razvoja korpusa, v poglavju 4 pa predstavimo njegov rezultat, novi korpus siParl 4.0. V zaključku se (poglavje 5) posvetimo naslednjim korakom pri razvoju in uporabi korpusa siParl.

2 ZGODOVINA PARLAMENTARNEGA KORPUSA SIPARL

Razvoj slovenskih parlamentarnih korpusov ima dokaj dolgo tradicijo, ki se je začela z prvimi koraki postavitve prve različice korpusa siParl. V naslednjem delu tako predstavljamo osnovne značilnosti slovenskega parlamentarnega sistema in parlamentarnih razprav, zgodovino razvoja korpusa ter predstavimo sorodne iniciative in projekte.

[1] https://www.clarin.eu/parlamint

2.1 Slovenski parlamentarni sistem

Slovenski parlament, ki predstavlja večino parlamentarnih razprav v korpusu siParl, je sestavljen iz dveh domov, pri čemer *Državni zbor* (DZ) predstavlja najvišje predstavniško in zakonodajno telo Republike Slovenije. Poleg zakonodajne funkcije Državni zbor opravlja tudi volilno in nadzorno funkcijo, sestavlja pa ga 90 poslancev, od katerih sta dva poslanca predstavnika italijanske in madžarske narodne skupnosti.

Plenarnim zasedanjem Državnega zbora so v korpusu že skoraj od samega začetka razvoja pridružena tudi zasedanja *delovnih teles Državnega zbora*, ki so bila ustanovljena za spremljanje razmer na posameznih področjih, pripravo političnih odločitev na teh področjih in obravnavo predlogov zakonov ali zakonodajnih aktov. Kadar Državni zbor nima plenarnega zasedanja, poslanci delujejo v teh delovnih telesih (Državni zbor Republike Slovenije, 2020), ki vključujejo odbore in komisije za posamezna področja (npr. odbor za zunanje zadeve ali komisija za narodni skupnosti) ter preiskovalne komisije.

Nazadnje podatke za obdobje modernega slovenskega parlamentarizma dopolnjujejo tudi zasedanja Kolegija predsednika Državnega zbora, ki je posvetovalno telo predsednika DZ in ki ima v določenih primerih tudi pristojnost odločanja. Sestavljajo ga predsednik in podpredsedniki Državnega zbora, predsedniki poslanskih skupin in poslanci narodnostnih skupin (tj. poslanci italijanske in madžarske narodnostne skupine).

V korpus so že od samega začetka poleg zasedanj dobro poznanega modernega parlamenta vključene tudi seje 11. mandatnega obdobja Skupščine Republike Slovenije. Seje skupščine predstavljajo zelo specifičen in pomemben zgodovinski kontekst za razvoj sodobnega parlamentarnega sistema v Sloveniji, ki je bila med letoma 1945 in 1991 del zvezno urejene Jugoslavije in v kateri je bil parlament odraz socialističnega sistema. Prve večstrankarske volitve so potekale aprila 1990, leta 1992 pa so poslanci skupščine sprejeli novo ustavo, ki je formalno končala obdobje Socialistične skupščine Slovenije in uvedla nov klasični parlament. Seje parlamenta tako ponujajo edinstven vpogled v obdobje pred, med in po osamosvojitvi Slovenije (Pančur in sod., 2018).

2.2 Od SlovParl do siParl

Razvoj korpusa siParl se je začel leta 2016, ko se je v okviru Raziskovalne infrastrukture Slovenskega zgodovinopisja (RI INZ) na Inštitutu za novejšo zgodovino pričel projekt digitalizacije obstoječih parlamentarnih dokumentov, ki bi zgodovinarjem omogočil boljši dostop do (in posledično omogočil tudi analizo) čedalje večjega števila digitalnih zgodovinskih virov v Sloveniji (Pančur in Šorn, 2016).

Celotni proces in posledični razvoj prvih različic korpusa siParl je že od samih začetkov temeljil na upoštevanju naslednjih načel (Pančur in Erjavec, 2020):

1. *Multidisciplinarnost*: Korpus mora biti uporaben ne le za zgodovinarje, temveč tudi za druge sorodne discipline.
2. *Inkluzivnost* dokumentov: Poleg parlamentarnih razprav je v bodoče potrebno vključevati tudi druge vrste parlamentarnih dokumentov.
3. *Dolgoročnost*: Načelo dolgoročnosti se v tem primeru nanaša predvsem na zagotavljanje stabilnega in dolgoročnega financiranja takšnih projektov (saj takšnih načrtov ni mogoče uresničevati v obdobju kratkoročnih raziskovalnih projektov), predvsem v okviru raziskovalnih infrastruktur.
4. *Upoštevanje načel odprte znanosti*: Vsa predhodna načela se lahko optimalno uresničuje le v skladu z načeli načeli odprte znanosti (npr. odprtost in zagotavljanje prostega dostopa podatkov, upoštevanje FAIR načel itd.)

V skladu z zgoraj naštetimi načeli korpus siParl že od svojih začetkov nastaja v tesnem sodelovanju med slovenskima raziskovalnima infrastrukturama DARIAH-SI in CLARIN.SI ter njunima evropskima infrastrukturama, DARIAH-EU in CLARIN.EU.

Prvi rezultat omenjenega procesa je bil korpus *SlovParl 1.0* (Pančur in sod., 2016), eden izmed prvih parlamentarnih korpusov slovansko govorečih držav. Korpus je vseboval 2,7 milijonov besed parlamentarnih razprav v Zboru združenega dela Skupščine Republike Slovenije (1990 – 1992). To obdobje je v slovenski zgodovini še posebej zanimivo, saj zajema obdobje pred, med in po osamosvojitvi Slovenije leta 1991 ter s tem povezan prehod iz socializma v demokracijo, ki se je odražal tudi v strukturi in delovanju parlamenta. SlovParl 1.0 je bil kodiran v skladu s smernicami za označevanje besedil Text Encoding Initiative Guidelines (TEI Consortium, 2020), prvotno z uporabo modula TEI za izvedena dela[2] (TEI Drama), ki je bil nato preoblikovan v uporabo elementov modula TEI za transkribirani govor[3] (TEI Speech) (Pančur, 2016), da bi bolje omogočil kasnejše jezikoslovno označevanje.

SlovParl 1.0 je tako postavil temelje za razvoj slovenskih parlamentarnih korpusov, ki so mu kmalu sledile druge različice:

- *SlovParl 2.0* (Pančur in sod., 2017), 1990 - 1992, dodani so bili prepisi zapisnikov Družbeno političnega zbora in Zbora Občin Skupščine Republike Slovenije (Pančur in sod., 2018).
- *siParl 1.0* (Pančur in sod., 2019), 1990 - 2018, dodani so bili prepisi Državnega zbora Republike Slovenije, delovnih teles in Kolegija predsednika Državnega zbora RS za obdobje od 1. do 7. mandata.
- *siParl 2.0* (Pančur in sod., 2020), 1990 - 2018, korpus vsebuje identične podatke kot siParl 1.0, vendar so pri tem bile odpravljene marsikatere nepravilnosti v označevanju in (ročno preverjenih) metapodatkih, podatki pa so ponovno označeni s takrat najsodobnejšimi orodji za jezikoslovno označevanje (Pančur in Erjavec, 2020).
- *siParl 3.0* (Pančur in sod., 2022; Meden in sod., 2024), 1990 - 2022, dodani so bili prepisi plenarnih sej Državnega zbora za 8. mandat.

Do korpusa siParl 2.0 je razvoj potekal precej linearno, kjer je vsaka različica dopolnila korpus z novimi podatki in odpravljala morebitne napake. Sočasno z razvojem siParl 2.0 pa se je v tem času vzpostavil mednarodni projekt ParlaMint (bolj detaljno predstavljen v poglavju 2.3), kjer se je priprava korpusa razdelila na dva sorodna korpusa, *siParl* in *ParlaMint-SI*.[4] Korpusa sta si med seboj sicer precej podobna, vendar se med seboj razlikujeta predvsem v načinu označevanja parlamentarnih podatkov. Hkrati podatki, pripravljeni za razvoj korpusa siParl, služijo neposredno tudi kot osnova za pripravo korpusa ParlaMint-SI.

[2] Performance Texts TEI module
[3] Transcriptions of Speech TEI module
[4] V času pisanja prispevka je trenutna najnovejša različica ParlaMint-SI 4.1, del nabora primerljivih korpusov ParlaMint 4.1.

Korpus	**Leto**	**Vsebina**	**Kodiranje**
SlovParl 1.0	2016	Zbor zduženega dela (1990 – 1992)	TEI Drama, TEI Speech
SlovParl 2.0	2017	Skupščina Republike Slovenije (1990 – 1992)	TEI Drama, TEI Speech
SiParl 1.0	2019	Skupščina RS (1990 – 1992) Državni zbor RS (1990–2018)	TEI Drama, TEI Speech
SiParl 2.0	2020	Skupščina RS (1990 – 1992) Državni zbor RS (1992 – 2018) Delovna telesa DZ (1996 – 2018) Kolegij predsednika DZ (1996–2018)	Parla-CLARIN
ParlaMint-SI 1.0	*2020*	*Državni zbor RS (2015–2020)*	*Parla-CLARIN, ParlaMint*
ParlaMint-SI 2.0/2.1	*2021*	*Državni zbor RS (2015 – 2020)*	*Parla-CLARIN, ParlaMint*
SiParl 3.0	2022	Skupščina RS (1990 – 1992) Državni zbor RS (1992 – 2018) Delovna telesa DZ (1996 – 2022) Kolegij predsednika DZ (1996 – 2018)	Parla-CLARIN
ParlaMint-SI 3.0	*2023*	*Državni zbor RS (2000–2022)*	*ParlaMint*

Tabela 1: Zgodovina razvoja korpusa siParl od svojih začetkov leta 2016 (SlovParl 1.0) do predzadnje različice korpusa (siParl 3.0) ter sočasen razvoj sorodnega korpusa ParlaMint-SI (besedilo, označeno z *ležečo pisavo*).

Celotna zgodovina razvoja korpusa siParl (in posredno korpusa ParlaMint-SI) do predzadnje različice korpusa (siParl 3.0, ki predstavlja izhodiščno točko nadaljnjega razvoja nove različice), je predstavljena v Tabeli 1.

Nova, trenutna različica korpusa siParl (4.0) je omenjeni korpus razširila še z manjkajočimi prepisi sej delovnih teles in Kolegija predsednika Državnega zbora za 8. mandat, ki ni bil vključen v različico 3.0. Poleg tega v na tej točki zelo dobro uveljavljen delovni postopek uvaja več posodobitev, povezanih z avtomatizacijo in standardizacijo določenih delov postopka. Nova različica in posodobljen delotok sta podrobneje predstavljena v poglavju 3 in 4.

2.3 siParl4.0 in povezane iniciative

Sočasno z razvojem korpusa siParl je vzporedno potekal tudi razvoj različnih sorodnih iniciativ in priporočil za označevanje parlamentarnih podatkov. V tem pogledu je potrebno omeniti predvsem priporočila za označevanje parlamentarnih razprav Parla-CLARIN, prilagoditev smernic *Text Encoding Initiative* (TEI) (TEI Consortium, 2020), ki omogočajo označevanje širokega nabora korpusov parlamentarnih razprav in drugih sorodnih dokumentov (Erjavec in Pančur, 2022). Korpus siParl 2.0 je tako postal prvi korpus, ki je bil kodiran v skladu s priporočili Parla-CLARIN.

Poleg priporočil Parla-CLARIN, ki so se kmalu uveljavile kot referenčna shema za označevanje parlamentarnih podatkov, je v letu 2020 temu sledil začetek projekta ParlaMint I. Primarni cilj projekta ParlaMint je bil zagotoviti ustrezne mehanizme za pretvorbo posameznih nacionalnih parlamentarnih korpusov v nabor medsebojno primerljivih in poenotenih parlamentarnih korpusov. Ob zaključku prve faze projekta (ParlaMint I, 2019 - 2021) je nabor obsegal 17 nacionalnih parlamentarnih korpusov. Kot osnova za enotno kodiranje parlamentarnih razprav pa so bila uporabljena prej omenjena priporočila Parla-CLARIN, ki jih je bilo potrebno prilagoditi (in predvsem zaostriti) za zagotavljanje uniformiranosti kodiranja podatkov. Projekt se je nadaljeval v letih od 2022 do 2023, kjer se je končni nabor korpusov razširil na 29 evropskih korpusov, tako nacionalnih parlamentov, kot tudi parlamentov avtonomnih pokrajin. V tej fazi so bili korpusi tudi dodatno obogateni z razširjenim naborom ne le podatkov, temveč tudi metapodatkov (Ogrodniczuk in sod., 2022). Kot omenjeno v poglavju 2.2, je v sklopu projekta ParlaMint nastajal tudi korpus ParlaMint-SI, kjer korpus siParl predstavlja osnovo pri razvoju vsake različice ParlaMint-SI.

Poleg ParlaMint-SI pa je v sklopu pregleda sorodnih inciativ in korpusov potrebno omeniti tudi druge nedavno objavljene korpuse, ki pokrivajo starejša obdobja slovenske parlamentarne zgodovine. *Parlamentarni korpus prve Jugoslavije Yu1Parl 1.0* (Kavčič in sod., 2023b), ki pokriva obdobje Kraljevine SHS v letih 1919 – 1939 vsebuje večjezične prepise Narodnega predstavništva Kraljevine Jugoslavije. Dokumenti so zapisani v srbohrvaškem in slovenskem jeziku (odvisno od govorca), kjer je bil srbohrvaški jezik zapisan v cirilici. Drugi zgodovinski korpus je *Korpus kranjske deželne skupščine Kranjska 1.0* (Kavčič in sod., 2023a), ki vsebuje prepise z obdobja 1861 – 1913. Dokumenti so prav tako dvojezični, v slovenščini in nemščini, pri čemer je bila nemška pisava najprej zapisana v gotici in šele pozneje v latinici. Omenimo lahko tudi (s korpusom siParl precej povezan) *Korpus programov političnih strank za slovenske parlamentarne volitve leta 2022 Programi2022* (Polanič in Dobranić, 2022), ki vsebuje 19 programov slovenskih političnih strank, in s katerim je bil narejen prvi korak k izpolnjevanju v poglavju 2.2 omenjenega načela inkluzivnosti oziroma dodajanju različnih vrst parlamentarnih dokumentov.

3 RAZVOJ KORPUSA SIPARL4.0

siParl 4.0 pokriva parlamentarne razprave Državnega zbora in njegovih delovnih teles za obdobje 1990 – 2022, pri katerem so bile v primerjavi z verzijo 3.0 dodane razprave delovnih teles DZ za 8. mandatno obdobje (2018 – 2022) ter posodobljen delotok priprave korpusa. Celoten postopek gradnje korpusa je v sestavljen iz naslednjih faz:

1. Ekstrakcija podatkov posameznih sej parlamentarnih razprav
2. Čiščenje in priprava podatkov
3. Označevanje podatkov v skladu s priporočili Parla-CLARIN:
   (a) Dvostopenjsko označevanje podatkov (po modulih TEI Drama in TEI Speech)
   (b) Razločevanje govorcev
   (c) Popis in dodajanje metapodatkov
   (d) Priprava kolofonov korpusa (ang. corpus header) in posameznih mandatov (ang. term headers) za dokumentiranje korpusa
4. Jezikoslovno označevanje korpusa za pripravo jezikoslovno označene različice korpusa
5. Generiranje izpeljanih formatov korpusa in dodajanje korpusa v konkordančnike.
6. Objava korpusa v repozitoriju CLARIN.SI.

Čeprav je bil potek dela pri sestavljanju korpusa siParl dobro uveljavljen že pri gradnji starejših različic korpusov siParl (in SlovParl), se je skozi leta pokazala potreba po bolj avtomatiziranem pristopu obdelave korpusa in s tem minimiziranju ročnih intervencij. Te so zaradi narave podatkov do določene mere neizogibne, predvsem pri popisu in dodajanju metapodatkov. Posodobljena verzija delotoka sicer ohranja zgoraj orisano strukturo izdelave korpusa, vendar pri posameznih fazah doda določene novosti, ki poskušajo pripomoči k prej omenjenim glavnim ciljem te posodobitve (tj. čim večja avtomatizacija in standardiziranost postopka, seveda ob ustreznem ročnem pregledu pravilnosti podatkov za čim višjo kakovost metapodatkov).

V tem razdelku se tako posvečamo posameznim fazam gradnje, izpostavimo posodobitve ter predstavimo primere in rezultate vsake posamezne faze.

3.1 Ekstrakcija in čiščenje podatkov

Začetno točko gradnje korpusa predstavlja postopek ekstrakcije podatkov. V primeru parlamentarnega korpusa so to posamezne seje parlamenta, v primeru korpusa siParl, so to seje Državnega zbora, ki so za obdobje 1992 – 2022[5] na voljo na spletni strani Državnega zbora[6] v formatu HTML. Ena izmed večjih novosti delotoka je avtomatizacija faze ekstrakcije podatkov, ki je bil do sedaj zgolj pol-avtomatiziran (določen del ekstrakcije je bil opravljen ročno, predvsem iz vidika kompleksne strukture spletne strani). Ravno zaradi nenavadne strukture spletne strani smo za ekstrakcijo sej razvili skripte za ekstrakcijo podatkov na podlagi Python knjižnice *Selenium*. Rezultati ekstrakcije so tako dokumenti v formatu HTML, ter ločena datoteka, ki vsebuje vse spremljajoče metapodatke posameznih dokumentov (npr. naslov seje, ime delovnega telesa, datum, tip seje). Ti dokumenti vsebujejo marsikatere nezaželene (HTML) oznake ali celo strukturne napake, ki onemogočajo nadaljnje procesiranje podatkov.[7] Pri čiščenju podatkov se tako v tem delu osredotočamo na odpravo strukturnih napak, pretvorbo HTML dokumentov v XML format ter druge problematične vidike originalnih podatkov, da s tem podatke pripravimo na nadaljnjo obdelavo.

[5] Prepisi so na spletni strani Državnega zbora na voljo le za obdobje 1992 - 2022, medtem ko so seje Skupščine Republike Slovenije (1990 – 1992) rezultat digitalizacijskega projekta z leta 2016.
[6] https://www.dz-rs.si/wps/portal/Home
[7] Večina nadaljnje obdelave temelji na skriptah XSLT, ki zahtevajo pravilno strukturirane XML dokumente (ang. valid XML structure).

3.2 TEI kodiranje podatkov – TEI Drama in TEI Speech

Naslednja faza delotoka je namenjena označevanju parlamentarnih podatkov v skladu s priporočili za označevanje parlamentarnih podatkov Parla-CLARIN (Erjavec in Pančur, 2019), ki temelji na *de-facto* standardu za označevanje strukture besedil, Text Encoding Initiative (TEI Consortium, 2020). Ta med drugim ponuja veliko število modulov, ki vsebujejo nabor označevalcev za različne namene. V primeru označevanja parlamentarnih razprav se v večini primerov uporabljata dva specifična modula: modul za označevanje izvedbenih del (*Preformance Texts*, pogosteje poznan kot TEI Drama), ter modul za označevanje prepisov govora (*Transcriptions of Speech*, oziroma TEI Speech). Priporočila Parla-CLARIN tako vsebujejo elemente modula za označevanje prepisov, z dodanimi nabori elementov iz drugih modulov, kot je na primer modul za označevanje imen, datumov, ljudi in krajev (Names, Dates, People, and Places).

V TEI skupnosti že vrsto let poteka obsežna razprava o uporabi modulov TEI Drama in TEI Speech za kodiranje parlamentarnih razprav, predvsem v smeri tega, kateri izmed modulov je bolj primeren kot osnova kodiranja parlamentarnih razprav. Modul TEI Drama predstavlja parlamentarne razprave kot gledališki perfomans. Parlament je v tem pogledu predstava, kjer vsi poslanci in drugi akterji v parlamentu igrajo določeno vlogo oziroma predstavljajo igralce, njihovi govori pa dramsko besedilo. Ta logika je v večini primerov vezana na poslanske govore, ki so lahko (in v večini primerov tudi so) predhodno pripravljeni, vendar pa se jih poslanci ne nujno tudi strogo držijo in se v svojih govorih pogosto odmaknejo od njih. Modul TEI Drama tako vključuje elemente za kodiranje govorcev kot seznam (`<castList>`), ime govorca (`<speaker>`) in “odrske napotke”(`<stage>`) (Pančur in sod., 2018). V nasprotju z dramskim modulom se modul TEI Speech osredotoča na poslanski govor, kjer se kot glavni argument uporablja dejstvo, da parlamentarne razprave (predvsem poslanski govori) v parlamentu predstavljajo govorjeni jezik (ang. spoken language). Čeprav se v korpusu nahajajo prepisi parlamentarnih razprav, te vsebujejo značilnosti govorjenega jezika (npr. premori med govorom, prekinitve govora, itd) (Truan in Romary, 2022).

Kljub temu, da za oba vidika obstajajo argumenti, je v večini primerov označevanja parlamentarnih korpusov na koncu prevladal modul TEI Speech, predvsem zaradi lažjega končnega jezikoslovnega označevanja parlamentarnih razprav.

V primeru korpusa siParl smo tako, kot že omenjeno v poglavju 2 v svojih začetkih (še pod imenom SlovParl 1.0 in 2.0) podatke označevali tako v obliki TEI Drama kot v TEI Speech. S prehodom v različico 2.0 pa je siParl postal prvi korpus, ki ima podatke označene v skladu s priporočilom Parla-CLARIN. Parla-CLARIN je postal zelo uveljavljen način kodiranja (slovenskih in tujih) parlamentarnih razprav, vendar pa v primeru korpusa siParl zaradi obsežne količine podatkov označevanje poteka v dveh stopnjah. Prva stopnja obsega označevanje podatkov v skladu z modulom TEI Drama, medtem ko v drugi stopnji podatke pretvorimo iz TEI Drama v format Parla-CLARIN.

Na to odločitev je vplivala tudi zgoraj omenjena razprava o naravi in ustreznem kodiranju parlamentarnih podatkov (TEI Drama proti TEI Speech). Ker je bil SlovParl primarno namenjen zgodovinarjem in humanistiki, je v tem oziru prevladal pogled na parlament kot performans in s tem povezano označevanje podatkov z modulom TEI Drama (Pančur in Šorn, 2016). Temu je kasneje sledila pretvorba v TEI Speech, ko se je korpus začelo tudi jezikoslovno označevati, da bi se s tem razširilo uporabnost korpusa tudi za druge vede (predvsem jezikoslovje). Pretvorba podatkov v oba formata je tako omogočilo ohranjanje obeh pogledov na parlamentarne podatke[8].

Ker poleg ohranjanja narave podatkov takšen način kodiranja omogoča tudi postopno kodiranje relativno kompleksnih podatkov, smo ta način dvostopenjskega označevanja ohranili tudi v trenutni različici korpusa. V nadaljevanju bolj podrobno predstavimo potek kodiranja podatkov v obeh stopnjah.

3.2.1 KODIRANJE PODATKOV V SKLADU S TEI DRAMA

Prva stopnja kodiranja podatkov vključuje transformacijo podatkov (s pomočjo skript XSLT) iz osnovnega formata podatkov v kodiranje v skladu z modulom TEI Drama. V tej stopnji tako zajamemo in kategoriziramo (primerno označimo) vse vzorce, ki se pojavljajo v podatkih:

- imena in priimki govorcev (npr. DR. MIRO CERAR:)
- točke dnevnega reda (npr.., 1. TOČKA DNEVNEGA REDA)
- vrzeli v prepisih, ki so posledica nerazumljivega dela govora ali iz drugih vzrokov manjkajočega dela besedila v prepisu (označeni kot “...”)

[8] Pretvorbe podatkov so na voljo v GitHub repozitoriju korpusa.

- posamezni govori in segmenti transkripta
- prekinitve in pripombe (označene kot besedilo v oklepaju /.../ ali (...)).

Potrebno je poudariti, da so to zgolj najbolj pogosti primeri vzorcev, ki se pojavljajo v prepisih in da se zaradi pogostih napak in napačnega označevanja v originalnih prepisih lahko pojavljajo v najrazličnejših oblikah (ki jih najpogosteje zajamemo z različnimi oblikami regularnih izrazov). To pomeni, da ostajajo podatki tudi po preliminarnem čiščenju in strukturiranju še vedno relativno neurejeni.

V posodobljenem delotoku v to fazo obdelave podatkov vključimo tudi proces razločevanja govorcev, ki je v preteklosti temeljil predvsem na ročnem pregledu in urejanju podatkov. Ta del procesa smo zato do določene mere uspeli avtomatizirati s pomočjo knjižnic v programskem jeziku Python. Proces razločevanja govorcev tako sedaj temelji na računanju Levenštajnove razdalje (ang. *Levenshtein distance*:

> Levenštajnova razdalja (ang. *Levenshtein distance*) med dvema nizoma je število modifikacij, potrebnih za pretvorbo enega niza (s1) v drugi niz (s2). Pokriva enoznakovne spremembe, kot so brisanje, vstavljanje in zamenjavo.

Na primeru parlamentarnih razprav smo tako na podlagi Levenštajnove razdalje ter frekvence, s katero se posamezna oblika pojavlja v besedilih, lahko oblikovali seznam vseh oblik imen in priimkov govorcev in najverjetnejšo pravilno obliko govorca (npr. za razločevanje oblik imen *"Alenka Bratušek"*, ki se v besedilih pojavlja 1000-krat, *"Alena Bratušek"* ki se v tej obliki zapisa pojavi trikrat, in *"Alenkra Bratušek"*, ki se pojavi zgolj enkrat, je kot najverjetnejša oblika pravilnega zapisa imena ponujena *"Alenka Bratušek"*). S tem uspešno minimiziramo čas, potreben za ročni pregled podatkov[9]. S tem oblikujemo osnovni seznam govorcev (`<listPerson>`), ter v razpravah posamezne govore opremimo z ustreznim identifikatorjem govorca (npr. `<sp who="#ŽižaFelice">`).

Končni produkt prve transformacije podatkov obsega dokumente, ki vsebujejo vse (v skladu z TEI Drama) označene vzorce, ki se pojavljajo v parlamentarnih razpravah, ter ročno pregledan seznam govorcev in njihove identifikatorje. Ta

[9] Ročni pregled je seveda še vedno nujen, predvsem v primeru manj pogostih oblik imena, govorcev z enakim imenom in priimkom itd.

```xml
<head>Vsebina zapisa seje</head
<stage type="title">REPUBLIKA SLOVENIJA,
DRŽAVNI ZBOR, KOLEGIJ PREDSEDNIKA
DRŽAVNEGA ZBORA</stage>
<stage type="session">1. seja</stage>
<stage type="date">(29. junij 2018)</stage>
<stage type="chairman">Sejo je vodil mag. Matej Tonin,
predsednik Državnega zbora.</stage>
<stage type="time">Seja se je pričela ob 10.01.</stage>
<sp who="ToninMatej">
    <speaker>PREDSEDNIK MAG. MATEJ TONIN:</speaker>
    <p>Dober dan, spoštovani vodje poslanskih skupin.</p>
    <p>Začenjam 1. sejo Kolegija predsednika
    Državnega zbora.</p>
    <p>S tem lahko preidemo kar na
    <title>1. TOČKO DNEVNEGA REDA</title>.</p>
</sp>
```

(a) TEI Drama

```xml
<list type="agenda">
   <item xml:id="KPDZ-001-Redna-2018-06-29.toc-item0">
       <title>Pred dnevnim redom</title>
       <ptr target="#KPDZ-001-Redna-2018-06-29.seg1"/>
       <ptr target="#KPDZ-001-Redna-2018-06-29.seg2"/>
       <ptr target="#KPDZ-001-Redna-2018-06-29.seg3"/>
   </item>
   ...
<note type="time">Seja se je pričela ob 10.01.</note>
<note type="speaker">PREDSEDNIK MAG. MATEJ TONIN:</note>
<u who="#ToninMatej"
   xml:id="KPDZ-001-Redna-2018-06-29.u1"
   ana="#chair">
  <seg xml:id="KPDZ-001-Redna-2018-06-29.seg1">Dober dan,
   spoštovani vodje poslanskih skupin.</seg>
   <seg xml:id="KPDZ-001-Redna-2018-06-29.seg2">
    Začenjam 1. sejo Kolegija predsednika
    Državnega zbora.</seg>
  <seg xml:id="KPDZ-001-Redna-2018-06-29.seg3">S tem lahko
   preidemo kar na 1. TOČKO DNEVNEGA REDA</seg>
</u>
```

(b) TEI Speech

Slika 1: Primer razlik med TEI Drama in TEI Speech kodiranimi parlamentarnimi razpravami.

pretvorba tako služi kot osnova za pretvorbo podatkov v skladu z priporočili Parla-CLARIN.

3.2.2 KODIRANJE PODATKOV V SKLADU Z MODULOM TEI SPEECH OZ. PRIPOROČILI PARLA-CLARIN

Druga stopnja kodiranja podatkov temelji na pretvorbi podatkov iz kodiranja TEI Drama v TEI Speech. V tej stopnji podatke s pomočjo XSLT (in nekaj Python) skript pretvorimo v končno obliko, ki je v skladu s priporočili Parla-CLARIN (ki vsebuje predvsem nabor oznak modula TEI Speech).

V tem delu poteka tudi združevanje posameznih govorov in njihovih segmentov (odstavkov). Del delotoka predstavlja tudi dodajanje identifikatorjev, označevanje vlog govorcev (`chair` za predsedujoče govorce in `regular` za poslance in druge govorce), označevanje opomb (`kinesic`, `vocal`, `incident` in gap) in priprava seznama točk dnevnega reda (ang. *agenda*), ki vsebujejo povezave na določene dele seje, ki se nanašajo na specifično točko dnevnega reda. Slika 1 prikazuje primer istega besedila, označenega z naborom oznak TEI Drama (Slika 1a) in TEI Speech (Slika 1b).

Ta faza tako predstavlja zadnjo stopnjo kodiranja podatkov in jih pripravi na naslednje stopnje gradnje korpusa – pripravo tako imenovanih korenskih dokumentov (ang. *root*, bolj podrobno predstavljenih pri opisu strukture korpusa v razdelku 4) ter jezikoslovno označevanje.

3.3 Priprava metapodatkov

Čeprav že sami transkripti parlamentarnih razprav predstavljajo bogat vir podatkov za raziskave parlamentarnega diskurza, z dodajanjem metapodatkov govorcev parlamentarni korpusi omogočajo raziskovanje drugih socialno-demografskih vidikov. V tem delu je velika večina dela sestavljena iz ročnega pregledovanja različnih virov, da tako zagotovimo natančen popis govorčevih in drugih metapodatkov, ki v grobem obsegajo:

1. Metapodatke o govorcih (`<listPerson>`):
   (a) ime in priimek govorca
   (b) spol govorca (generirano avtomatično na podlagi vzorcev ženskih in moških imen, skupaj z ročnim pregledom)
   (c) datum rojstva (če je le ta na voljo)
   (d) afiliacija govorca z različnimi parlamentarnimi (ali drugimi političnimi organizacijami, npr. politično stranko ali v primeru poslanca, njegovo/njeno afiliacijo s poslansko skupino)

2. Metapodatke o organizacijah (`<listOrg>`):
   (a) podatki o Državnem zboru (ali v primeru SSK11, Skupščini, `xml:id="SK"`, `xml:id="DZ"`)
   (b) seznam delovnih teles (`xml:id="workingBodies"`)
   (c) seznam političnih strank, poslanskih skupin, nepovezanih (poslanskih skupin ali posameznih poslancev), ter drugih organizacij (npr. koalicija DEMOS)
   (d) seznam Vlad RS (`xml:id="GOV"`)
   (e) Vsak seznam vsebuje nadaljnje metapodatke:
      i. ime organizacije (`<orgName>`)
      ii. podatek o časovnem obdobju, v katerem organizacija obstaja (`<event>`)
      iii. povezava na Wikipedijo ali drug vir podatkov, ki je bil uporabljen za sestavo tega seznama (`idno`).
      iv. seznam "dogodkov"(`<listEvent>`/`<event>`), ki označuje posamezna časovna obdobja obstoja organizacije.

Metapodatki tako obsegajo ne le osnovne identifikacijske podatke govorca, ki jih pridobimo že pri obdelavi prepisov (npr. ime in priimek govorca), temveč tudi informacije o govorčevih afiliacijah (npr. poslanske skupine, politične stranke) ter natančnih obdobjih v katerih je deloval. Slednje je potrebno pridobiti iz zunanjih virov, kot so mandatna poročila Državnega zbora Republike Slovenije[10] ali popisa sestav preteklih vlad.[11] Hkrati so ti metapodatki na voljo tako v korenskih dokumentih posameznih mandatov in celotnega korpusa, kot tudi v datotekah, ki spremljajo korpus. Te so primerne predvsem za analizo podatkov v tabelaričnih formatih, kot sta CSV in TSV ter predstavljajo pomemben del za analizo podatkov v konkordančnikih. Slika 2 predstavlja primer metapodatkov o govorcu, kodiranih v korpusu.

3.4 Jezikoslovno označevanje korpusa

Korpus siParl je poleg ravnokar predstavljene osnove različice na voljo tudi v jezikoslovno označeni različici (*TEI.ana*). Ta različica je po vsebini, metapodatkih in kodiranju enaka osnovni, vendar besedilu doda jezikoslovne označbe (anotacije). Dodane jezikoslovne oznake obsegajo tokenizacijo besedila, segmentacijo na povedi, oblikoskladenjsko označevanje (ang. *PoS tagging*, lema-

[10] Mandatna poročila Državnega zbora RS
[11] Pretekle Vlade Republike Slovenije

```
<person xml:id="BandelliMarko">
    <persName>
        <surname>Bandelli</surname>
        <forename>Marko</forename>
    </persName>
    <sex value="M"/>
    <birth when="1967-11-19">
        <placeName ref="https://www.geonames.org/
        3165185/trieste.html">Trst</placeName>
    </birth>
    <affiliation role="MP"
        ref="#DZ"
        from="2018-06-22"
        to="2018-09-12"
        ana="#DZ.8"/>
    <affiliation role="member"
        ref="#party.SAB"
        from="2018-06-22"
        to="2018-09-12"
        ana="#DZ.8"/>
        ...
    <affiliation role="minister"
        ref="#GOV"
        from="2018-09-13"
        to="2018-11-19"
        ana="#GOV.13"/>
</person>
```

Slika 2: Primer kodiranja metapodatkov poslanca, ki vsebuje osnovne osebne podatke in govorčevo afiliacijo z Državnim zborom, politično stranko in vlado.

tizacijo, skladenjske odvisnosti (ang. *syntactic dependencies*) in prepoznavanja imenskih entitet (ang. *named-entity recognition*). Besedilo v tej verziji korpusa je bilo jezikoslovno označeno s pomočjo orodja CLASSLA-Stanza (verzija 2.1) [12] (Ljubešić in Dobrovoljc, 2019; Terčon in Ljubešić, 2023). Slika 3 prikazuje primer jezikoslovno označenega stavka "*Hvala za besedo.*".

4 SLOVENSKI PARLAMENTARNI KORPUS SIPARL4.0

Korpus siParl 4.0 zajema parlamentarne razprave in razprave delovnih teles Državnega zbora Republike Slovenije med leti 1990 in 2022 ter obsega več kot

[12] https://github.com/clarinsi/classla

```
<s xml:id="KPDZ-redna-1-2014-08-22.seg10.1">
<w xml:id="KPDZ-redna-1-2014-08-22.seg10.1.1"
    msd="UPosTag=NOUN|Case=Nom|Gender=Fem|
    Number=Sing" ana="mte:Ncfsn"
    lemma="hvala">Hvala</w>
<w xml:id="KPDZ-redna-1-2014-08-22.seg10.1.2"
    msd="UPosTag=ADP|Case=Acc"
    ana="mte:Sa" lemma="za">za</w>
<w xml:id="KPDZ-redna-1-2014-08-22.seg10.1.3"
    msd="UPosTag=NOUN|Case=Acc
    |Gender=Fem|Number=Sing" ana="mte:Ncfsa"
    lemma="beseda" join="right">besedo</w>
<pc xml:id="KPDZ-redna-1-2014-08-22.seg10.1.7"
msd="UPosTag=PUNCT" ana="mte:Z">.</pc>
<linkGrp type="UD-SYN" targFunc="head argument">
    <link ana="ud-syn:amod"
        target="#KPDZ-redna-1-2014-08-22.seg10.1.2
        #KPDZ-redna-1-2014-08-22.seg10.1.1"/>
    <link ana="ud-syn:root"
        target="#KPDZ-redna-1-2014-08-22.seg10.1
        #KPDZ-redna-1-2014-08-22.seg10.1.2"/>
    <link ana="ud-syn:cc"
        target="#KPDZ-redna-1-2014-08-22.seg10.1.4
        #KPDZ-redna-1-2014-08-22.seg10.1.3"/>
   </linkGrp>
</s>
```

Slika 3: Primer jezikoslovno označenega stavka "*Hvala za besedo.*"

13 tisoč sej, milijon govorov in 230 milijonov besed. Vsebinsko, korpus obsega parlamentarne razprave naslednjih parlamentarnih teles:

- Skupščina Republike Slovenije, 1990 - 1992, 11. mandatno obdobje
- Državni zbor Republike Slovenije, 1992 - 2022, 1. - 8. mandatno obdobje
- Delovna telesa Državnega zbora Republike Slovenije, 1996 - 2022, 2. - 8. mandatno obdobje
- Kolegij predsednika Državnega zbora, 1996 - 2022, 2. - 8. mandatno obdobje

V primerjavi z drugimi sorodnimi parlamentarnimi korpusi (npr. korpusi ParlaMint) ima siParl 4.0 bolj kompleksno strukturo, razlog za to pa je delno povezan tudi s strukturo priporočil za označevanje parlamentarnih razprav Parla-CLARIN:

1. Korenski dokument korpusa (ang. corpus root, `<teiCorpus>`):
   (a) Metapodatki za celoten korpus (`<teiHeader>`):
      i. Bibliografski podatki o celotnem korpusu (`<fileDesc>`)
      ii. Taksonomije za organizacije, odbore, izraze, vrste sestankov, vrste govornikov (`<taxonomy>`)
      iii. Seznam govorcev za celoten korpus (`<listPerson>`)
      iv. Seznam organizacij za celoten korpus (`<listOrg>`)
   (b) XInclude datoteke posameznih sej za celoten korpus (`<xi:include>`)
2. Korenski dokument posameznih mandatov (ločeno po vrstah parlamentarnih teles – DZ, delovna telesa) (`teiCorpus`):
   (a) Metapodatki za posamezen mandat (`<teiHeader>`):
      i. Enaka struktura kot za celoten korpus, vendar vsebuje samo informacije, ki so relevantne za posamezen mandat.
   (b) XInclude datoteke posameznih sej za posamezen mandat (`<xi:include>`)
3. Posamezne seje, ločeno po tipu sej (npr. redne, izredne) in dnevu (`<TEI>`)
   (a) Metapodatki posamezne seje (`<teiHeader>`)
      i. Bibliografski podatki posamezne seje (`<fileDesc>`)
      ii. Točke dnevnega reda (`<agenda>`)
   (b) Transkript seje (`<text>`):
      i. bibliografske opombe, npr. naslov seje, datum, ura, predsednik (`<head>`, `<note>`)
      ii. Transkript govora (`<u>`):
         A. Opombe prepisovalca (`<note>`, `<gap>`, `<vocal>`, `<kinesic>`, `<incident>`)
         B. Paragrafi (`<seg>`)

Struktura vsebuje korenske dokumente korpusa (ang. *corpus root*) in mandatov (ang. *term root*) na isti ravni, dokumenta sta medsebojno neodvisna. Z drugimi besedami, bodisi celoten korpus bodisi posamezen mandat se lahko obravnava kot ločen dokument XML. Kar zadeva datoteke in strukturo direktorijev in datotek, korenski dokument korpusa in vsak korenski dokument mandata ustrezata eni datoteki. Opozoriti je potrebno, da imamo za vsak mandat dva korenska dokumenta, odvisno od tega, katero parlamentarno telo je zasedalo, bodisi Državni zbor (*SDZ*), bodisi eno izmed delovnih teles Državnega zbora (*SDT*).

<table>
<tr><th colspan="2">Mandat</th><th>Obdobje</th><th>Dnevi/ Mandat</th><th>Dnevi- Skupaj</th><th>Org.</th><th>Govor.</th><th>Posl.</th></tr>
<tr><td colspan="2">SSK11</td><td>1990-05-05 - 1992-12-23</td><td colspan="2">608</td><td>3</td><td>512</td><td>242</td></tr>
<tr><td>1</td><td>DZ</td><td>1992-12-23 - 1996-11-28</td><td colspan="2">462</td><td>1</td><td>315</td><td>101</td></tr>
<tr><td rowspan="2">2</td><td>DZ</td><td rowspan="2">1996-11-28 - 2000-10-27</td><td>430</td><td rowspan="2">2.141</td><td>1</td><td>359</td><td>105</td></tr>
<tr><td>DT</td><td>1711</td><td>28</td><td>1268</td><td>108</td></tr>
<tr><td rowspan="2">3</td><td>DZ</td><td rowspan="2">2000-10-27 - 2004-10-22</td><td>303</td><td rowspan="2">1708</td><td>1</td><td>296</td><td>106</td></tr>
<tr><td>DT</td><td>1405</td><td>26</td><td>1291</td><td>102</td></tr>
<tr><td rowspan="2">4</td><td>DZ</td><td rowspan="2">2004-10-22 - 2008-10-15</td><td>283</td><td rowspan="2">1583</td><td>1</td><td>237</td><td>103</td></tr>
<tr><td>DT</td><td>1300</td><td>26</td><td>1886</td><td>101</td></tr>
<tr><td rowspan="2">5</td><td>DZ</td><td rowspan="2">2008-10-15 - 2011-12-15</td><td>233</td><td rowspan="2">1541</td><td>1</td><td>191</td><td>103</td></tr>
<tr><td>DT</td><td>1308</td><td>26</td><td>2144</td><td>102</td></tr>
<tr><td rowspan="2">6</td><td>DZ</td><td rowspan="2">2011-12-16 - 2014-08-01</td><td>196</td><td rowspan="2">1321</td><td>1</td><td>202</td><td>115</td></tr>
<tr><td>DT</td><td>1125</td><td>29</td><td>1994</td><td>111</td></tr>
<tr><td rowspan="2">7</td><td>DZ</td><td rowspan="2">2014-08-01 - 2018-06-22</td><td>288</td><td rowspan="2">1987</td><td>1</td><td>207</td><td>101</td></tr>
<tr><td>DT</td><td>1699</td><td>23</td><td>2697</td><td>100</td></tr>
<tr><td rowspan="2">8</td><td>DZ</td><td rowspan="2">2018-06-22 - 2022-05-13</td><td>269</td><td rowspan="2">1882</td><td>1</td><td>264</td><td>109</td></tr>
<tr><td>DT</td><td>1613</td><td>27</td><td>2077</td><td>104</td></tr>
<tr><td colspan="2">Skupaj</td><td>1990-05-05 - 2022-05-13</td><td colspan="2">13233</td><td>69</td><td>9568</td><td>720</td></tr>
</table>

Tabela 2: Osnovne kvantitativne značilnosti posameznih mandatov (**Mandat**), ki vsebujejo zajeta zakonodajna (oziroma mandatna) obdobja (**Obdobje**). Tu se podatki ločijo še na tip zakonodajnega telesa (**SSK11** – Skupščina Republike Slovenije, **DZ** – Državni zbor, **DT** – delovna telesa), število dni, ko so potekala zasedanja po posameznih zakonodajnih organih (**Dnevi/Mandat**) in skupno število dni za celotno obdobje (**Dnevi - Skupaj**). Nazadnje so predstavljena še števila drugih zakonodajnih organov ali organizacij (**Org.**), govornikov (**Govor.**) in poslancev v parlamentu (**Posl.**). Državni zbor sicer sestavlja 90 poslancev, vendar lahko spremembe poslancev med zakonodajnim obdobjem povzročijo, da skupno število preseže 90 poslancev. Poleg tega je SSK11 (Skupščina Republike Slovenije) zaradi drugačne strukture parlamentarnega sistema štela 242 poslancev (oziroma delegatov).

Posamezne seje se nato razlikujejo glede na njihovo vrsto in datum, ko so te potekale. Vrste sej se naprej delijo na *redne*, *izredne* ali *skupne* seje, pri čemer izredne seje skliče predsednik Državnega zbora na zahtevo najmanj četrtine poslancev ali predsednika republike in se običajno ukvarjajo s časovno občutljivimi in nujnimi zadevami (Državni zbor Republike Slovenije, 2020). Vsaka seja ima dnevni red, ki vsebuje točke, o katerih se bo razpravljalo, in o katerih se razpravlja v obliki govorov (ki jih včasih lahko prekine neverbalna vsebina, kot so informacije o izidu glasovanja, dejanja, kot je aplavz, itd.). Nazadnje sledijo označeni prepisi razprave, vključno z vsemi prekinitvami in drugimi (verbalnimi in neverbalnimi) dejavniki govora.

Številčno celoten korpus obsega 13.233 besedil, 1.221.133 posameznih govorov, 3.736.988 segmentov oz. odstavkov, 13.122.555 stavkov, 230.585.189 besed in 273.023.945 pojavnic (ang. *tokens*). Tabela 2 prikazuje bolj podrobno statistiko posameznih mandatov glede na posamezen tip zakonodajnega telesa ter celotnega korpusa.

Celotni korpus je na voljo v različnih formatih. Osnovna oblika korpusa vključuje TEI XML (oziroma v skladu z Parla-CLARIN označene) parlamentarne razprave. Poleg omenjene različice je korpus na voljo tudi v jezikoslovno označeni različici (TEI.ana). Nazadnje pa je korpus dostopen tudi v drugih izpeljanih formatih, in sicer kot navadno besedilo (.txt), v formatu CoNLL-U (.connlu) ter v vertikalnem formatu (.vert), ki je namenjen uporabi korpusa v konkordančnikih.

Korpus siParl je na voljo na repozitoriju CLARIN.SI[13] (Pančur in sod., 2024), GitHub repozitoriju[14], prav tako pa je dostopen tudi na CLARIN.SI konkordančnikih NoSketch Engine[15].

5 ZAKLJUČKI

V prispevku smo predstavili novo različico Slovenskega parlamentarnega korpusa siParl 4.0, ki obsega parlamentarne razprave Republike Slovenije med leti 1990 in 2022 in s tem pokriva različna obdobja zgodovine Slovenije ter politične sisteme. Kot tak predstavlja bogat vir podatkov za analizo političnega in parlamentarnega diskurza. Korpus siParl je že v svojih začetkih še kot digitalizacijski projekt pomembno prispeval k usmerjenemu naboru parlamentarnih razprav in s tem k kasnejšemu razvoju slovenskih parlamentarnih korpusov. Hkrati se je korpus razvijal vzporedno z sorodnimi iniciativami, kot so priporo-

[13] http://hdl.handle.net/11356/1936
[14] https://github.com/DARIAH-SI/siParl
[15] https://www.clarin.si/ske/#dashboard?corpname=siparl40

čila za označevanje parlamentarnih razprav Parla-CLARIN in projektom Parla-Mint, na katere je tudi (ne)posredno vplival.

V primerjavi s prejšnjo različico smo obstoječi korpus siParl 3.0 razširili z manjkajočimi razpravami delovnih teles in razpravami Kolegija predsednika Državnega zbora Republike Slovenije za 8. mandat (2018 – 2022). Poleg omenjene razširitve podatkov, ki jih korpus pokriva, smo se v tej verziji posvetili tudi posodobitvi delotoka razvoja korpusa, z namenom poenostavitve in čim večje avtomatizacije procesa.

V prihodnosti se želimo posvetiti razširitvi korpusa s plenarnimi razpravami in razpravami drugih delovnih teles 9. mandata, ki je trenutno še v teku. Poleg vsebinskih razširitev si želimo korpus nadgrajevati tudi na metapodatkovni in tehnični ravni. V tem pogledu na dolgi rok načrtujemo konsistentno dodajanje novih prepisov in povezovanje le teh z zunanjimi podatkovnimi bazami ali drugimi dodatnimi parlamentarnimi gradivi. Tu primarno izpostavljamo povezovanje parlamentarnih razprav s Poročevalcem Državnega zbora[16], ki vsebuje besedila zakonov in drugih dodatkov ki jih sami prepisi razprav ne vsebujejo, in ki je prosto dostopen v okviru portala Zgodovina Slovenije - SIstory[17]. Nazadnje želimo v prihodnosti postopno razširiti nabor metapodatkov, ki trenutno opisujejo parlamentarne razprave v korpusu.

Poleg nadaljnjega tehničnega in vsebinskega razvoja korpusa želimo v prihodnosti pripraviti poglobljeno analizo vsebine Slovenskega parlamentarnega korpusa siParl. Kot je navedeno v citatu v naslovu tega prispevka,[18] parlament in parlamentarne razprave predstavljajo nenehne kontrolirane polemike oziroma razprave med poslanci in drugimi govorci – v ta namen želimo korpus siParl v prihodnosti obogatiti z označevanjem sentimenta, ki nam bodo skupaj s prepisi in njihovimi metapodatki omogočile poglobljen vpogled v "polje kontroliranega konflikta", ki ga parlament predstavlja.

[16] Poročevalec Državnega zbora
[17] Portal Zgodovina Slovenije - SIstory
[18] Citat pripada poslanki Maši Kociper, izrečen na 90. izredni seji Državnega zbora (22. 12. 2021) v 8. mandatu.

ZAHVALA

Avtorji se zahvaljujemo anonimnim recenzentom za koristne pripombe ter dr. Nikoli Ljubešiću za jezikoslovno označevanje korpusa siParl 4.0. Prispevek je nastal v okviru slovenskih raziskovalnih infrastruktur DARIAH-SI in CLARIN.SI, raziskovalnega programa "Tehnologije znanja (P2-0103)"ter infrastrukturnega programa "Raziskovalna infrastruktura Slovenskega zgodovinopisja (I0-0013)", financiranih s strani Javne agencije za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije.

LITERATURA

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., ... Fišer, D. (2022). The Parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 1–34. (https://doi.org/10.1007/s10579-021-09574-0)

Erjavec, T. in Pančur, A. (2022). The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. *Journal of the Text Encoding Initiative*(14). (https://doi.org/10.4000/jtei.4133)

Erjavec, T. in Pančur, A. (2019). *Parla-CLARIN: A TEI Schema for Corpora of Parliamentary Proceedings.* (https://clarin-eric.github.io/parla-clarin/)

Fišer, D. in Pahor de Maiti, K. (2020, Aug). Voices of the parliament. *Modern Languages Open,*. (https://doi.org/10.3828/mlo.v0i0.295)

Fišer, D. in Pahor de Maiti, K. (2021). *Voices of the parliament: A corpus approach to parliamentary discourse research.* (https://sidih.github.io/voices/index.html)

Kavčič, A., Mundjar, A. in Marolt, M. (2023a). *Carniolan provincial assembly corpus kranjska 1.0*. Faculty of Computer and Information Science, University of Ljubljana. (http://hdl.handle.net/11356/1824)

Kavčič, A., Mundjar, A. in Marolt, M. (2023b). *Parliamentary corpus of first yugoslavia (1919-1939) yu1Parl 1.0*. Faculty of Computer and Information Science, University of Ljubljana. (http://hdl.handle.net/11356/1845)

Ljubešić, N. in Dobrovoljc, K. (2019, August). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. V *Proceedings of the 7th workshop on Balto-Slavic Natural Language Processing* (str. 29–34). Florence, Italy: Association for Computational Linguistics. (https://doi.org/10.18653/v1/W19-3704)

Meden, K., Erjavec, T. in Pančur, A. (2024). Slovenian parliamentary corpus siParl. *Language Resources and Evaluation*, 1–21.

Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešić, N., Çöltekin, Ç., ... Katja, M. (2022, June). ParlaMint II: The show must go on. V D. Fišer, M. Eskevich, J. Lenardič in F. de Jong (Ur.), *Proceedings of the workshop parlaclarin iii within the 13th language resources and evaluation conference* (str. 1–6). Marseille, France: European Language Resources Association. https://aclanthology.org/2022.parlaclarin-1.1

Pančur, A. (2016). Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI. V *Zbornik konference jezikovne tehnologije in digitalna humanistika* (str. 142–48).

Pančur, A., Erjavec, T., Meden, K., Ojsteršek, M., Šorn, M. in Blaj Hribar, N. (2022). *Slovenian parliamentary corpus (1990-2022) siParl 3.0*. Institute of Contemporary History. (http://hdl.handle.net/11356/1748)

Pančur, A., Erjavec, T., Ojsteršek, M., Šorn, M. in Blaj Hribar, N. (2019). *Slovenian parliamentary corpus (1990-2018) siParl 1.0*. Institute of Contemporary History. (http://hdl.handle.net/11356/1236)

Pančur, A., Erjavec, T., Ojsteršek, M., Šorn, M. in Blaj Hribar, N. (2020). *Slovenian parliamentary corpus (1990-2018) siParl 2.0*. Institute of Contemporary History. (http://hdl.handle.net/11356/1300)

Pančur, A., Meden, K., Erjavec, T., Ojsteršek, M., Šorn, M. in Blaj Hribar, N. (2024). *Slovenian parliamentary corpus (1990-2022) siParl 4.0*. Institute of Contemporary History. (http://hdl.handle.net/11356/1936)

Pančur, A. in Šorn, M. (2016). Smart big data: Use of Slovenian parliamentary papers in digital history. *Contributions to Contemporary History*, *56*(3), 130–146.

Pančur, A., Šorn, M. in Erjavec, T. (2016). *Slovenian parliamentary corpus (1990-1992) SlovParl 1.0*. Institute of Contemporary History. (http://hdl.handle.net/11356/1075)

Pančur, A., Šorn, M. in Erjavec, T. (2017). *Slovenian parliamentary corpus (1990-1992) SlovParl 2.0*. Institute of Contemporary History. (http://hdl.handle.net/11356/1167)

Pančur, A., Šorn, M. in Erjavec, T. (2018). Slovparl 2.0: The collection of Slovene parliamentary debates from the period of secession. V *D. fišer, & ad j. maria eskevich (eds.) parlaclarin 2018 workshop proceedings may* (Zv. 7, str. 2018).

Pančur, A. in Erjavec, T. (2020). The siParl corpus of Slovene parliamentary proceedings. V *Proceedings of the second parlaclarin workshop* (str. 28–34).

Polanič, P. in Dobranić, F. (2022). *Corpus of political party programs Programi2022*. Institute of Contemporary History. (http://hdl.handle.net/11356/1734)

Sestava in organiziranost. (2020). https://www.dz-rs.si/wps/portal/Home/odz/pristojnosti/organiziranost. (Dostop: 2023-02-25)

TEI P5: Guidelines for Electronic Text Encoding and Interchange. (2020). https://tei-c.org/guidelines/p5/. (Dostop: 27.2.2023)

Terčon, L. in Ljubešić, N. (2023). *CLASSLA-Stanza: The next step for linguistic processing of South Slavic Languages.* (https://doi.org/10.48550/arXiv.2308.04255)

Truan, N. in Romary, L. (2022). Building, Encoding, and Annotating a Corpus of Parliamentary Debates in TEI XML: A Cross-Linguistic Account. *Journal of the Text Encoding Initiative*(14).

"ACCORDING TO THEORY, PARLIAMENT IS A FIELD OF CONTROLLED CONFLICT": SLOVENIAN PARLIAMENTARY CORPUS SIPARL 4.0

The debates of national parliaments and other legislative bodies, and thus the parliamentary corpora, represent an important source of data for the study of political and especially parliamentary speech and other related phenomena in the humanities and social sciences. The siParl Slovenian Parliamentary Corpus contains a series of parliamentary debates of Slovenian parliamentary bodies from the period between 1990 and 2022, covering not only different periods of Slovenian history, but also transitions between different political systems. The corpus has a long history of development and has influenced the development of related corpora, initiatives and projects. In this article, we present the new version of siParl 4.0, describe the development of the corpus, the structure and the encoding of parliamentary data in the corpus. We also outline the next steps in the further development of siParl.

Keywords: parliamentary corpora, parliamentary debates, Slovenian Parliament, TEI-XML, Parla-CLARIN

EXPANDING THE FRENK CORPUS OF SOCIALLY UNACCEPTABLE DISCOURSE TO FRENCH

Kristina PAHOR DE MAITI TEKAVČIČ,[1,2] Nikola LJUBEŠIĆ,[3] Darja FIŠER[2]
[1]Faculty of Arts, University of Ljubljana
[2]Institute of Contemporary History, Ljubljana
[3]Jožef Stefan Institute, Ljubljana

This paper outlines the creation of the French part of the FRENK corpus, which contains socially unacceptable comments posted in response to news articles on the topics of LGBT and migrants which were published on Facebook by prominent media outlets. These comments were manually annotated for the type and target of socially unacceptable comments. Out of 10,239 comments with around 300,000 tokens in FRENK-fr, around a third of the comments represent socially unacceptable discourse, of which around 1% are violent. These are most often aimed at migrants, who together with the LGBT community and their supporters represent the most prominent target group of socially unacceptable comments. FRENK-fr is fully comparable to other language-specific parts of the FRENK corpus, and can serve as a valuable resource for cross-cultural qualitative analyses of disrespectful online communication which can also inform actions by civil society and political institutions. Additionally, FRENK-fr provides essential data for training more generalizable language models to identify socially unacceptable discourse.

Keywords: socially unacceptable discourse, hate speech, French, migrants, LGBT

1 INTRODUCTION

The last two decades have seen a visible rise in the importance of social media platforms in influencing public opinion and actions. Social media have become a powerful hybrid space merging the public and private sphere in previously unseen ways which blurs the boundaries between information of general relevance, gossip and verified facts. The ease of content production enabled by social media platforms, allows for large amounts of messages to enter the virtual space. This saturated media landscape makes verifying and filtering information difficult, which is why the end users can be quickly faced with disinformation and socially unacceptable content. Such messages are highly problematic because they influence our reasoning and decision-making processes, but also because they negatively impact social cohesion and thus the possibilities for a better future.

Efforts to understand socially unacceptable discourse (SUD) are crucial in order to limit its propagation and nonconstructive or even dangerous effects on the recipients (López & López, 2017), but since SUD is a highly heterogenous phenomenon without clear register characteristics (Zhang & Luo, 2019), the task proves especially complicated. Moreover, it has been shown that negative impact can be triggered even by implicit inappropriate messages (Kopytowska & Baider, 2017) which is why the efforts cannot remain limited to the investigation of *hate speech* or the most explicit violent content alone. In fact, the researchers of SUD increasingly shift their attention to non-violent, but nevertheless SUD messages and explore the role of various rhetorical devices and strategies in the construction of SUD (Despot et al., 2023).

If the shift from explicitly violent to implicitly offensive messages has gained momentum, the research on SUD still shows an important limitation due to its bias toward English datasets (Piot et al., 2024).[1] This proves largely inadequate in the current state of events when different national internet safety agencies regularly report on the propagation of SUD online, and national and EU regulations are becoming more stringent in reference to online content moderation.[2] Imbalanced representation of SUD corpora in multiple languages hinders the development of robust and generalizable models for the moderation of SUD online which makes the efforts towards a less toxic online environment on the long run less successful. Moreover, where SUD datasets in languages other than English exist, the comparative research and model development is made difficult due to differences in definitions, terminology and annotation guidelines (Carneiro et al., 2023).

[1]This is not to say that hate speech datasets in other languages, in particular French and Slovene do not exist, for example Chiril et al. (2020); Vanetik and Mimoun (2022) or Kralj Novak et al. (2021).
[2]See, for example, the EU Digital Services Act, retrieved May 10, 2024, from https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

The FRENK dataset (Ljubešić et al., 2021) represents one of the rare multilingual comparable corpora of socially unacceptable online comments. So far, common data collection and annotation methodology has been used to create datasets of Croatian, English, Slovene and Dutch online comments that include socially acceptable and unacceptable content (Ljubešić et al., 2019; I. Markov et al., 2021). This paper is concerned with the extension of the FRENK corpus to French. A set of online comments in French was annotated according to the FRENK methodology thus creating a dataset that is comparable to other languages already included in the FRENK corpus. The new corpus FRENK-fr 1.0 is available for download from the CLARIN.SI repository (Pahor de Maiti et al., 2024).

The paper is structured as follows: in Section 2, we provide an overview of the corpus creation process with the explanation and examples related to the annotation schema, some information about the annotation campaign and the level of inter-annotator agreement. Section 3 outlines the characteristics of the new FRENK-fr focusing on the structure of the corpus with regard to the distribution of types of socially unacceptable discourse and its targets. When relevant, the results are put into context by comparing them to data from the Croatian, English or Slovene part of the FRENK corpus. The paper ends with Section 4 which summarizes the main features of FRENK-fr and addresses its comparative potential in relation to other language-specific parts of the FRENK corpus.

2 CORPUS ANNOTATION

The data for the FRENK corpus, including FRENK-fr, contain comments that were posted under Facebook posts of most popular national mainstream media outlets which often share their news articles via social media platforms. The three most trending news media outlets in France were selected according to the *Alexa* service, and include *Le Monde, Le Figaro* and *20minutes*.[3] The data were collected during the FRENK project (ARRS J7-8280; 2017–2020) and cover the period between 2010 and 2017 with the majority of data posted from 2015 on. Since the corpus was intended to be manually annotated, the first objective

[3] The Alexa service, retrieved in 2017/2018, from https://www.alexa.com/topsites/countries, is no longer available.

was to collect data that will likely include a relatively high share of potential SUD. Given the (then) prominent and controversial societal events related to stronger migration flows and several initiatives addressing the discrimination of the LGBT community, the decision was made to create a classifier that filtered the harvested Facebook posts of media outlets for two specific topics, namely LGBT/homophobia and migrants/Islamophobia (see Ljubešić et al. (2019) for more details on classifier creation).

2.1 Annotation schema

The French dataset is annotated according to the project-specific typology (Ljubešić et al., 2021) which includes *Types* and *Targets* of SUD. More specifically, SUD is defined as any form of communication that is vulgar, offensive or violence-evoking and/or that represents a disruption in communication by propagating negative claims that can hurt the addressee and do not represent any added value to the argument put forth (Vehovar et al., 2020). Thus, SUD differs from an argumented critique or literary/journalistic genres, like satire, by recurring to unfounded claims and language use that can reasonably be expected to trigger psychological or physical harm.

The *Type* level of the annotation scheme consists of four classes that indicate the type of discourse found in the comment, as listed below:

- **acceptable speech** (the comment does not contain vulgar lexis nor any claim that could be judged offensive or violent)
  - *Il me semble qu'il y ait la séparation de l'Eglise et de d'Etat en France. Que l'Eglise prenne des positions ne me choque pas, au final c'est l'Etat qui décide.* [I think there is a separation of Church and State in France. I'm not shocked that the Church takes a stand, in the end it's the State that decides.]
- **inappropriate speech** (the comment contains vulgar language that is not directed at anyone in particular or this connection cannot be clearly established)
  - *conneries!* [bullshit!]

- **offensive speech** (the comment contains discriminatory and defamatory claims that target an identifiable entity and is potentially punishable by civil law (Vehovar et al., 2020))
  - *Et c'est vrai que le mec au milieu qui ressemble a papillon de Tchernobyl a dus prendre son déguisement dans la derniere gay pride !!!* [And it's true that the guy in the middle, who looks like a Chernobyl butterfly, must have gotten his costume at the last gay pride parade!!!]
- **violent/threatening speech** (the comment contains calls to physical violence or threats and is potentially punishable by criminal law (Fišer et al., 2017))
  - *Faut les écrasés ces mecs* [They need to be run over those guys]

Furthermore, *Offensive* comments and *Violent/threatening* comments are given an additional tag according to the type of discrimination. Two tags are available to this end: *Background* – when the basis for discrimination is the target's background or any of the protected characteristics, such as gender, sexual orientation, nation, race, ethnicity, religion, disability, etc.; and *Other* – when the basis for discrimination are the target's interests or professional group affiliations, such as political groups, media publishing houses, civil protection groups, etc. In case the annotator finds several SUD elements that indicate different SUD types, the harshest type of SUD is selected as the final category. See the guidelines published in the CLARIN.SI repository at http://hdl.handle.net/11356/1462 (Ljubešić et al., 2021) for more details.

The *Target* level of the annotation scheme consists of five categories which were applied in a hierarchical order, meaning that when a comment offended or attacked multiple targets, the comment was assigned the first relevant target from the list below:

- **migrants/LGBTQ**
  - *OH, pauvre petit Syrien....* [OH, poor little Syrian....]
- **supporters of migrants/LGBTQ**
  - *Qu'ils arrêtent de nous faire chier avec cette connerie qui ne concerne qu'une minorité.* [Let them stop bothering us with this nonsense that

only concerns a minority.] (as a response to a news item post about a pro-LGBT marriage activist)

- **journalist/media**
  - *20minutes@ liberté d'expression juste quand ça vous arrange !! mais bon être honnête ne vous tuera pas croyez moi, vous devrez l'essayer* [20minutes@ freedom of expression just when it suits you!! But well, being honest won't kill you, believe me, you should try it]
- **commenter** (used when the comment targets the other commenter but the pro or against position of the targeted commenter cannot be clearly established)
  - *On voit bien ton intelligence povre mek* [Your intelligence is clearly evident, poor guy]
- **other** (anyone else, including the opponents of migrants/LGBTQ)
  - *Mais l'ONU ne sert à rien !* [But the UN is useless !]

2.2 Annotation campaign

Similarly to the campaigns for other languages in FRENK, the annotation campaign for FRENK-fr included a training period, i.e., training sessions with an expert annotator from the FRENK project, written guidelines, and an annotation period with expert support which included a mailing list exchange and regular feedback sessions to resolve difficult cases. The comments were annotated in context, which means they were delivered to the annotators in entire threads and linked to the original post. It should be noted that the entire thread does not mean a complete string of comments ever posted in absolute terms, since it was only possible to collect the data that existed at the time of collection. Thus, the comments and posts removed by Facebook (or even the information about the number of such posts) could not be retrieved.

As with other languages, the potential bias in annotation was addressed not only with training sessions, but also with ensuring multiple annotations for each comment and the support of an expert annotator that provided regular feedback. While the Slovene and English data were annotated by approximately eight annotators (Ljubešić et al., 2019), the French data, similar to Dutch data (I. Markov et al., 2021), received two annotations, with all disputed cases resolved by an expert annotator. The comments were annotated in a spreadsheet editor where they were listed in the original thread order and were linked to the relevant news item post and the information of their status as a reply or not. Each annotator worked independently of the other annotator, and needed around 100 hours each to complete the annotation task (this amount does not include the time needed to disambiguate the disputed cases, since this task was not timed).

A lower number of annotations per comment was due to difficulties in engaging the already trained annotators and lack of funding, but this annotation setting was deemed appropriate for two reasons. First, the student annotators were well familiar with the guidelines, the French annotation campaign being already their third or even fourth annotation task (after Slovene, English and Croatian); and second, all disputed cases were resolved by the expert annotator. The two annotators were paid for their work and were advanced students of French, but non-native speakers. To mitigate the potential influence of their lack of linguistic and sociohistorical knowledge on the interpretation of the comments, contact was established with a native French speaker for consultations when needed. While the sophistication of arguments in the comments was rather low, the main difficulty for comprehension was certainly the use of slang expressions, and occasional severe orthographic or syntactic errors that hampered the processing of the comment (which, however, is a common characteristic of the FRENK corpus).

The encountered difficulties in engaging the annotators warrant a short note. Annotating SUD proved to be a highly challenging task not only at the technical level with a complex annotation schema, but also at the psychological level, which was first anecdotally shown by relatively high withdrawal of annotators between the annotations campaigns, but later also confirmed empirically (Pahor de Maiti & Fišer, 2021). The dropout from one annotation campaign to the other. i.e., from Slovene to French, was of course influenced by the project team's judgment of annotator's performance, their self-reported level of linguistic knowledge and personal circumstances not related to the annotation task as such, but the analysis on the annotator's perception of this particular work clearly showed that the annotation of SUD is psychologically burdening and can lead to adverse effects on the people involved, but can also negatively impact the task at hand if these aspects are not addressed.

2.3 Inter-annotator agreement

Table 1 shows inter-annotator agreement for the French dataset which was assessed with Krippendorff's Alpha. The calculations were performed with the statistical package K-Alpha Calculator (Marzi et al., 2024) and considered SUD *Type* as an ordinal variable and SUD *Target* as a nominal variable as per observations reported in Ljubešić et al. (2019). The coefficient (2 annotators, 20694 values) for SUD *Type* is 0.651 [CI (95%, 1000 iterations): 0.635, 0.668], and 0.634 [CI (95%, 1000 iterations): 0.618, 0.648] for SUD *Target*. The result is slightly below the threshold for moderate agreement (0.67) (Krippendorff, 2019), but low-agreement is expected in SUD annotation because of the complexity of the phenomenon (Waseem, 2016).

Table 1: Inter-annotator agreement scores for SUD *Type* and SUD *Target* by topic and combined with 95% confidence interval over 1000 iterations for FRENK-fr.

Subset	*SUD Type*	*CI*	*SUD Target*	*CI*
LGBT	0.504	[0.472, 0.538]	0.488	[0.462, 0.515]
Migrants	0.810	[0.786, 0.834]	0.811	[0.785, 0.834]
Combined	0.651	[0.635, 0.668]	0.634	[0.618, 0.648]

There is, however, a noticeable difference in the coefficient between the two topics in the French dataset with a lower agreement on the annotations in the *LGBT* subset. Ljubešić et al. (2019) observe a similar pattern for English, i.e., lower agreement in the *LGBT* subset, and suggest that this can be possibly explained by the fact that the three most frequent annotation combinations in the *LGBT* dataset account for 91% of the annotations (compared to 75–80% in other subsets) which increases the possibility for the agreement by chance.

Although this is a possible explanation which might be due to overly detached annotators, a highly plausible risk in annotation of distressing content (Pahor de Maiti & Fišer, 2021), this result might simply reflect the specifics of the annotation schema which makes certain combinations impossible and other more likely. For instance, the *Acceptable speech* label can only be paired with *No target* label; *Background offensive* is most naturally paired with *LGBT/Migrants* as the target, and *Other offensive* is very likely to be paired with *Other* as a target. In FRENK-fr, the three most frequent *Type-Target* combinations represent 86% of all annotations in the *LGBT* subset and 89% in the *Migrants* subset, and in both subsets include *Acceptable speech – No target*, *Other Offensive – Other* and *Background offensive – LGBT/Migrants*. These combinations are expected both in terms of *Type-Target* pairs, but also in terms of their frequency, since violent comments are more likely to get removed by the platform. Therefore, the observed distribution might be a reflection of actual dataset characteristics rather than due to the chance.

In comparison to Slovene or English, as reported in Ljubešić et al. (2019), the outlier seems to be the high score for the French *Migrants* subset which shows high agreement both for *Type* and *Target*, whereas the French *LGBT* subset and other languages exhibit low to moderate agreement. Although the three most frequent annotation combination in the *Migrants* subset account for a large proportion (89%) of the annotations, this score can be, as with the *LGBT* dataset, rather than by chance, explained by annotators' experience, more easily interpretable comments (given their relative shortness, see Section 3), but also the annotators' personal convictions which might have been more aligned with regard to rights of migrants than that of the LGBT community. This also supports the observation that in French as well as in Slovene and English, the agreement is higher for the *Migrants* subset.

3 KEY CHARACTERISTICS OF FRENK-fr

This section quantitatively describes the structure of FRENK-fr focusing first on the overall size of the dataset and then more specifically on the *Type* and *Target* distribution. Table 2 indicates the number of Facebook posts posted by media outlets, the number of comments and tokens. FRENK-fr contains comments in entire threads posted under 66 posts made by the three selected media outlets on Facebook (see Section 2). The corpus consists of more than 10,000 annotated comments with around 300,000 tokens which all relatively evenly cover both topics. The number of comments is comparable to the Croatian (10,970 comments), Dutch (10,732 comments), English (11,661 comments) and Slovene dataset (10,164 comments) (Ljubešić et al., 2021; I. Markov et al., 2021).

Table 2: The number of media outlet posts, comments and tokens per topic in FRENK-fr.

	LGBT	*Migrants*	*Total*
Posts	31	35	66
Comments	5,182	5,057	10,239
Tokens	172,418	128,426	300,844

3.1 Type of socially unacceptable discourse

Table 3 gives information about the distribution of comments by their SUD *Type* in both topics. The data show that around two thirds of comments in FRENK-fr contain *Acceptable* discourse, and one third SUD with around 1% of the content representing *Violent* propositions. A low number of *Violent* comments observed is expected given the generally low share of violent comments on social media. The estimations vary, but are usually below 7% of the observed dataset (Berglind et al., 2019; Vidgen & Yasseri, 2020), and are probably the result of platform moderation efforts and of societal pressure which makes publishing violent content less appropriate.[4]

Table 3: The absolute and relative number of comments in FRENK-fr per topic reflecting the distribution of SUD *Types*.

	LGBT		*Migrants*		*Total*	
	#	%	#	%	#	%
Acceptable	3,476	33.95	3,692	36.06	7,168	70.01
Inappropriate	24	0.23	19	0.19	43	0.42
Offensive	1,664	16.25	1,248	12.19	2,912	28.44
–Background	437	4.27	308	3.01	745	7.28
–Other	1,227	11.98	940	9.18	2,167	21.16
Violent	18	0.18	98	0.96	116	1.13
–Background	3	0.03	83	0.81	86	0.84
–Other	15	0.15	15	0.15	30	0.29
Total	5,182	50.61	5,057	49.39	10,239	100.00

[4] See for example Facebook community standards on hate speech, retrieved May 10, 2024, from https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

Table 4: The absolute and relative number of tokens in FRENK-fr per topic reflecting the distribution of SUD *Types*.

	LGBT		*Migrants*		*Total*	
	#	%	#	%	#	%
Acceptable	104,146	34.62	78,473	26.08	182,619	60.70
Inappropriate	579	0.19	513	0.17	1,092	0.36
Offensive	67,216	22.34	47,796	15.89	115,012	38.23
–Background	16,417	5.46	11,656	3.87	28,073	9.33
–Other	50,799	16.89	36,140	12.01	86,939	28.90
Violent	477	0.16	1,644	0.55	2,121	0.71
–Background	56	0.02	1,289	0.43	1,345	0,45
–Other	421	0.14	355	0.12	776	0.26
Total	172,418	57.31	128,426	42.69	300,844	100.00

Although a 30% share of SUD is much smaller compared to, for example, the Slovene dataset where it amounts to a worrying 50% of all comments (Pahor de Maiti et al., 2020), this is still a clear indicator that an important share of discriminatory speech online is tolerated and not recognized as problematic enough to be reported and consequently removed from the platform. Furthermore, this observation is interesting from an interlingual perspective: we see that in the FRENK corpus, the French comments contain less SUD than Slovene, and the same is true of English comments when compared to Slovene comments in FRENK (Ljubešić et al., 2019), as well as to immigration-related online posts in Polish (Lewandowska-Tomaszczyk, 2017) or Cypriot Greek (Baider & Kopytowska, 2017). A contributing factor could be that in some of these studies, non-native speakers were employed for the annotation, and so their comprehension of comments might have lacked in precision which could have led them judge comments more or less harshly than a native speaker would. Nonetheless, a certain consistency across languages points towards an interesting avenue for future work since it is not clear whether this is just a result of potential annotation bias, more efficient algorithms for content moderation or if it indeed highlights a trait of communication culture linked to speakers of these languages.

Topic-wise, FRENK-fr contains a rather evenly distributed number of comments on the topic of LGBT and migrants which is a consequence of the data selection process. The interesting point, however, is that commenters post slightly more SUD comments on the LGBT topic than on migrants topic which is different from, for example, Slovene and English data where we can observe the opposite situation (Ljubešić et al., 2019). The *Violent* comments, however, are in all three languages more frequent in the *Migrants* subset. In FRENK-fr, in particular, we see five times more violent SUD in comments discussing migrants-related issues as opposed to those discussing LGBT-related issues.

This result is possibly impacted by the choice/availability of the media outlet posts, since certain subtopics related to LGBT or migrants trigger more disagreement then other subtopics, but it is also likely that it shows a lower level of tolerance of French commenters in FRENK toward LGBT issues in comparison to migrants. It should be noted, however, that this result cannot be interpreted as a clear indicator of more prominent discrimination of the LGBT community alone, since the topic subset of data includes SUD comments targeting different individuals or groups (see Section 3.2).

Table 4 provides information on the shares of tokens per topic. Length-wise, based on the comparison of the number of comments (see Table 3) and tokens produced (see Table 4) and their median value for the number of tokens (indicated in brackets), *Offensive* (26) comments appear to be longer than *Acceptable* (14) and *Violent* (13) ones, and comments on the LGBT topic (20) tend to be longer than those related to migrants (15). The two-tailed Mann-Whitney U statistics, evaluating whether the comment length differs between the *Offensive* and *Violent* comments on the one hand, and between *LGBT-related* and *migrants-related* comments on the other hand, is statistically significant in both cases at medium effect size (*LGBT* vs. *MIGR* subset: U = 15138751.0, p = .00 (3.07e-42), LGBT = 5,182, MIGR = 5,057, r = 0.135; *Violent* vs. *Offensive* comments: U = 240941.0, p = .00 (6.03e-15), Offensive = 2,912, Violent = 116, r = 0.142).[5] A statistically significant difference in the length of *Violent* and *Offensive* comments was also observed in the Slovene FRENK (Pahor de Maiti et al., 2020), and suggests a certain complexity inherent to *Offensive* comments. Based on qualitative analyses of Slovene data (Pahor de Maiti et al., 2023), it seems that in the *Offensive* comments, the commenters especially often use

[5] Calculated with the SciPy library created by Virtanen et al. (2020).

face-saving strategies in order to preserve their social image in view of possible damage brought about by their discriminatory propositions.

Long comments are less frequent in the *Violent* and *Acceptable* comments since there is less need to save face (Brown & Levinson, 1987). By posting an acceptable, i.e., a neutral/positive message, a person's public image is not in danger, while those commenters that post violent comments, which usually consist of short calls or allusion to violent actions, and thus intentionally break societal norms, do so by a lack of concern for their face, or more likely, to strengthen their position in a selected in-group. Authors of offensive comments, on the other hand, often attempt to present their views in an elaborate manner, and although they may genuinely be trying to address the complexity of the issue at hand, these does not override their discriminatory content. However, due to their complexity, such comments may give the impression of solid and acceptable argumentation, allowing them to be perceived as legitimate contributions to democratic discourse, despite being fundamentally socially unacceptable. Given that LGBT-related comments tend to be longer and possibly more complex than migrants-related ones, this might also contribute to the explanation why the LGBT-related comments might appear more difficult for the annotators, and hence a lower inter-annotator agreement for this topic across languages.

3.2 Target of socially unacceptable discourse

Table 5 provides information about the distribution of comments by their SUD *Target* in both topics. The shares are given for each topic separately. The data is provided only for the *Offensive* and *Violent* subset of data, since the *Acceptable* and *Inappropriate* comments did not receive the *Target* label.

We see that both *LGBT* and *Migrants* subset exhibit a similar distribution of SUD *Targets* across the comments. Namely, the topic-related minority group (LGBT or migrants) and their supporters represent the main SUD *Target* referenced in slightly more than 40% of the comments. Although this observation is not unique to FRENK-fr, and is in fact less pronounced than, for instance, in the Slovene or Croatian FRENK where the minority groups and their supporters are targeted in 50–70% of SUD comments (Pahor de Maiti, n.d.), we can see this as a clear indicator that more needs to be done to limit SUD aimed at vulnerable social groups in order to prevent the spread of discriminating ideas often advocated as free speech.

Table 5: The relative and absolute number of comments in FRENK-fr per topic and SUD Type reflecting the distribution of SUD *Targets*.

		Offensive	*Violent*	*Total*	
		%	%	%	#
LGBT	LGBT/related	44.05	0.59	44.65	751
	Other/opponent	24.02	0.42	24.44	411
	Commenter	21.34	0.06	21.40	360
	Media	9.51	/	9.51	160
MIGR	Migrant/related	35.14	6.24	41.38	557
	Other/opponent	29.12	0.89	30.01	404
	Commenter	21.77	0.15	21.92	295
	Media	6.69	/	6.69	90

Although SUD aimed at the minority groups is prevailing, an important share of SUD comments is also aimed at other individuals, namely the opponents and other commenters (20–30% each). This suggests that measures are needed at the level of general public in order to promote social cohesion and limit the back and forth of SUD comments that only aggravate the atmosphere. Moreover, migrants-related discussion in FRENK-fr appears slightly more polarizing for the French-speaking commenters then the LGBT-related topic which has a two times larger difference between comments targeting the minority group versus those targeting the opponents (approx. 10-point difference in the *Migrants* subset, and 20-point difference in the *LGBT* subset). In comparison to other languages, however, the French dataset appears more polarized than, for example, the Slovene and Croatian dataset where the commenters tend to be much more united in producing SUD mainly against the minority groups and their supporters (e.g., the Croatian dataset contains 70% of the comments targeting LGBT/supporters and only less than 10% of the comments targeting the opponents) (Pahor de Maiti, n.d.).

Media/journalists, on the other hand, are referenced in a relatively small amount of the comments, namely in around 8% of the French SUD comments with a higher count of such comments in the *LGBT* subset. This is still a higher share than in the Slovene or Croatian dataset where SUD targeting media/journalists amounts to around 5% of all comments, and is more frequent in the *Migrants* subset (Pahor de Maiti, n.d.). Therefore, this shows that the commenters more frequently express their dissatisfaction with the work of journalists in French than in Slovene or Croatian, and are more critical in the case of reporting on the topic of LGBT. Furthermore, despite the small frequency of the comments targeting media/journalists, and although expressing critique can, in general, be a positive thing, our observation pertains to disrespectfully-communicated criticism which can be much more damaging to journalists and the general perception of journalistic integrity than an argumented opposition, and should, therefore, be limited (Č. Markov & Đorđević, 2024).

On the positive note, however, media/journalists are never targeted in FRENK-fr with violent propositions. These are, as observed in Section 3.1, mostly found in the *Migrants* subset where they are in the great majority of cases aimed at migrants and their supporters. The same observation can be made for the Slovene dataset, but not for the Croatian one where the LGBT community and their supporters are the central group receiving violent verbal attacks. This indicates also a cultural difference in the perception of the two minority groups, and shows French commenters in FRENK as more violently intolerant towards migrants. Opponents and other commenters are less often targeted in violent comments both in the *Migrants* subset as in the *LGBT* subset, where *Violent* comments are altogether rather rare (around 1% of data). Unsurprisingly, the violent comments are especially rare in direct addresses of other commenters since the level of othering and dehumanization is usually higher for an external group of the discussion, like migrants, than for one of the active discourse participants.

4 CONCLUSION

This paper presented the creation process of the French part of the FRENK corpus of socially unacceptable comments, and its characteristics. FRENK-fr includes comments posted as a reaction to news posts related to the topics of LGBT/homophobia and migrants/Islamophobia which were published by well-known national media outlets on Facebook. These comments were manually annotated according to the FRENK project-specific schema and contain the type of socially unacceptable discourse, varying from acceptable to inappropriate, offensive and violent content, and the target of socially unacceptable comments which included the topic-related minority groups (LGBT, migrants), their supporters and opponents, other commenters and journalists.

The main aim of the creation of FRENK-fr was twofold: first, obtaining data for a qualitative comparative analysis of socially unacceptable communication practices on Facebook, and second, expansion of the FRENK corpora wit a new language for the needs of the development of a robust and generalizable model for data classification and hate speech detection.

The here presented FRENK-fr is a dataset that is fully comparable to other language-specific parts of the FRENK corpus, i.e., Croatian, English, Slovene and Dutch, since the data collection and filtering process, data size, as well as the annotation schema and annotation procedure were adopted from the Slovene FRENK which was created first. The main difference compared to the Croatian, English and Slovene, but not the Dutch FRENK dataset, is in the number of the annotations per comment. This means that the final annotation in the case of Croatian, English and Slovene was calculated as the mode of all the annotations, while in for Dutch and French, the final annotations are either the input of two annotators in the case of agreement, or the expert annotator label in the case of dispute. The inter-annotator agreement assessment shows similar scores across languages.

FRENK-fr is comparable to other languages both in the selection of media outlets as well as the relevance of the period covered. The selection of media sources, encompassing mainstream liberal, conservative and a more sensationalist outlet, is consistent across both French and other languages. Furthermore, the time frame of the comments coincides with similar societal changes across all languages/countries.

In particular, the 2015–2017 period was a time of an increased migration flow from the conflict zones in the Middle East and Africa that spread across Europe. France, like the other countries, faced difficulties processing asylum applications, had problems solving issues inside refugee camps, and faced increased intolerance towards the Muslim population in the country. Regarding the LGBT issues, all countries saw the organisation of the Pride Parade and other LGBT-related awareness-raising but also anti-LGBT events which received important media coverage, but also witnessed different campaigns advocating for legislative changes to address LGBT-discriminatory laws (e.g., *mariage pour* *tous* [marriage for all] campaign in France, or *čas je za* [time for yes] in Slovenia, advocating for the legalization of the same-sex marriage).

The analysis of the FRENK-fr annotation campaign showed that similarly to other language-specific FRENK datasets, the LGBT topic appears more difficult for the annotators which resulted in a lower inter-annotator agreement on that topic. FRENK-fr is, in general, characterized by a lower number of SUD comments compared to some of the other languages included in the FRENK corpus, with 30% of the comments representing SUD, of which only 1% is labeled as *Violent* speech. Furthermore, we observed that both topics attract a similar amount of SUD comments, and that *Offensive* comments tend to be longer than *Acceptable* or *Violent* ones, and LGBT-related comments longer than migrants-related comments. Target-wise, SUD comments are most often aimed at the two topic-specific minority groups, migrants and LGBT, and their supporters. With a lower, but not negligible share, SUD comments also target opponents of minority groups and other commenters. Violent comments, in particular, are in most cases aimed at migrants.

These results but also public reports on the current situation of communication culture online, clearly indicate that there is still much work to do in order to promote inclusive behaviour in public online environments. FRENK-fr can importantly contribute to this objective. It expands the FRENK corpus of socially unacceptable discourse to French, and because of its corpus creation design, enables fully comparable analyses with the other parts of the FRENK corpus. As such, it represent a highly valuable resource for inter-cultural qualitative analyses of disrespectful communication practices online that can inform the actions of the civil society and political institutions, but also provides crucial material for training language models created for the classification of socially unacceptable discourse.

ACKNOWLEDGMENTS

This work has been supported by the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (ARRS N06-0099), Digital Humanities research programme (Digital Humanities: resources, tools and methods; ARRS P6-0436) and the ARENAS project (EU Horizon Europe research and innovation programme; GA No. 101094731).

REFERENCES

Baider, F., & Kopytowska, M. (2017). Conceptualising the Other: Online discourses on the current refugee crisis in Cyprus and in Poland. *Lodz Papers in Pragmatics*, *13*(2), 203–233.

Berglind, T., Pelzer, B., & Kaati, L. (2019). Levels of hate in online environments. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 842–847).

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge University Press.

Carneiro, B. M., Linardi, M., & Longhi, J. (2023). Studying Socially Unacceptable Discourse Classification (SUD) through different eyes:" Are we on the same page?". *arXiv preprint arXiv:2308.04180*.

Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., & Coulomb-Gully, M. (2020). An annotated corpus for sexism detection in french tweets. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 1397–1403).

Despot, K. Š., Anić, A. O., & Veale, T. (2023). "Somewhere along your pedigree, a bitch got over the wall!" A proposal of implicitly offensive language typology. *Lodz Papers in Pragmatics*, *19*(2), 385–414.

Fišer, D., Erjavec, T., & Ljubešić, N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the first workshop on abusive language online* (pp. 46–51).

Kopytowska, M., & Baider, F. (2017). From stereotypes and prejudice to verbal and physical violence: Hate speech in context. *Lodz Papers in Pragmatics*, *13*(2), 133–152.

Kralj Novak, P., Mozetič, I., & Ljubešić, N. (2021). *Slovenian twitter hate speech dataset imsypp-sl*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1398

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology (4th Ed.)*. Sage publications. https://doi.org/10.4135/9781071878781

Lewandowska-Tomaszczyk, B. (2017). Incivility and confrontation in online conflict discourses. *Lodz papers in pragmatics*, *13*(2), 347–367.

Ljubešić, N., Fišer, D., & Erjavec, T. (2019). The FRENK datasets of socially unacceptable discourse in Slovene and English. In *Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings 22* (pp. 103–114).

Ljubešić, N., Fišer, D., Erjavec, T., & Šulc, A. (2021). *Offensive language dataset of Croatian, English and Slovenian comments FRENK 1.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1462

López, C. A., & López, R. M. (2017). Hate Speech in the Online Setting. In S. Assimakopoulos, F. H. Baider, & S. Millar (Eds.), *Online Hate Speech in the European Union – A Discourse-Analytic Perspective* (pp. 10–12). Springer.

Markov, Č., & Đorđević, A. (2024). Becoming a target: Journalists' perspectives on anti-press discourse and experiences with hate speech. *Journalism Practice, 18*(2), 283–300.

Markov, I., Ljubešić, N., Fišer, D., & Daelemans, W. (2021). Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the eleventh workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 149–159).

Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha Calculator–Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. *MethodsX*, *12*, 102545.

Pahor de Maiti, K. (n.d.). *Metaphor in socially unacceptable discourse online [doctoral dissertation]*. University of Ljubljana.

Pahor de Maiti, K., Fišer, D., & Erjavec, T. (2020). Grammatical footprint of socially unacceptable facebook comments. *Language Technologies & Digital Humanities*.

Pahor de Maiti, K., & Fišer, D. (2021). Working with socially unacceptable discourse online: Researchers' perspective on distressing data. In *Proceedings of the 8th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2021)* (pp. 78–82).

Pahor de Maiti, K., Franza, J., & Fišer, D. (2023). Haters in the spotlight: gender and socially unacceptable Facebook comments. *Internet Pragmatics, 6*(2), 173–196.

Pahor de Maiti, K., Ljubešić, N., & Fišer, D. (2024). *Offensive language dataset of French comments FRENK-fr 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1947

Piot, P., Martín-Rodilla, P., & Parapar, J. (2024). MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection. *arXiv preprint arXiv:2401.06526*.

Vanetik, N., & Mimoun, E. (2022). Detection of racist language in french tweets. *Information*, *13*(7), 318.

Vehovar, V., Povž, B., Fišer, D., Ljubešić, N., Šulc, A., & Jontes, D. (2020). Družbeno nesprejemljivi diskurz na Facebookovih straneh novičarskih portalov. *Teorija in praksa*, *57*(2), 622–645.

Vidgen, B., & Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics*, *17*(1), 66–78.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... Contributors, S. . (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waseem, Z. (2016). Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138–142).

Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, *10*(5), 925–945.

RAZŠIRITEV KORPUSA DRUŽBENO NESPREJEMLJIVEGA DISKURZA FRENK NA FRANCOŠČINO

Ta prispevek predstavlja francoski del korpusa FRENK, ki zajema družbeno nesprejemljive komentarje, napisane kot odziv na novice s tematiko LGBT in migracij, ki so jih na Facebooku objavile priljubljene medijske hiše. Ti komentarji so bili ročno označeni glede na vrsto in tarčo družbeno nesprejemljivega govora. Od 10.239 komentarjev z okoli 300.000 pojavnicami v korpusu FRENK-fr je približno tretjina komentarjev označena kot družbeno nesprejemljiva, od tega pa le 1% predstavljajo nasilni komentarji. Ti so najpogosteje usmerjeni proti migrantom, ki skupaj z LGBT skupnostjo in njihovimi podporniki predstavljajo napogostejšo ciljno skupino družbeno nesprejemljivih komentarjev. Korpus FRENK-fr je v celoti primerljiv z drugimi jezikovno-specifičnimi deli korpusa FRENK in lahko služi kot pomemben vir za medkulturne kvalitativne analize nespoštljive komunikacije na spletu, ki lahko podajo pomembne uvide za oblikovanje ukrepov na ravni civilne družbe in političnih institucij. Poleg tega korpus FRENK-fr zagotavlja tudi kakovostne podatke za treniranje jezikovnih modelov, namenjenih za prepoznavanje družbeno nesprejemljivega diskurza.

Keywords: družbeno nesprejemljivi diskurz, sovražni govor, francoščina, migranti, LGBT

A CORPUS LINGUISTIC CHARACTERISATION OF SPERIODIKA

Ajda PRETNAR ŽAGAR

Institute for Contemporary History

The paper provides a computational analysis of sPeriodika, a historical corpus of Slovenian periodicals spanning from 1771 to 1914. The authors focus on ten prominent newspapers within the corpus, employing keyword analysis, word frequency counts, and concordance analysis to characterise the content and historical development of the Slovenian language. The study describes newspaper characteristics through computational methods, relating the findings to the post-1848 period of intense nation-building. Additionally, it addresses the challenges posed by low-quality OCR (Optical Character Recognition) in historical documents. The results are threefold: 1) a quantitative description of the selected newspapers, 2) insights into the historical progression of the Slovenian language, 3) analysis of the nature of OCR errors within the corpus. The keyword analysis reveals specific thematic orientations of the newspapers, such as agriculture, pedagogy, feuilletons, and advertising. It also underscores the newspapers' roles in nation-building. The study contributes to the field of digital humanities by demonstrating how computational tools can unlock historical insights from digitised textual data, despite the limitations of OCR technology.

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

1 INTRODUCTION

The last decade saw a significant increase in academic research on historical newspaper processing (Ehrmann et al., 2023). The applications range from digitisation efforts and corpora production to computational analysis and the development of new methods.

sPeriodika (Dobranić et al., 2023) is a recently published corpus of historical Slovenian periodicals from 1771 to 1914. The corpus is extremely extensive and is based on the OCR-ed periodicals from the dLib digital library, maintained by the National and University Library of Slovenia (for the history of digital editions of periodicals, see (Eiselt, 2015)). It features some of the most important periodicals of the time, contributing to increased literacy and nation-building in Slovenia (Dović, 2006; Amon, 2008).

The paper is a corpus linguistic study as proposed in the original paper by Dobranić et al. (Dobranić et al., 2024). We selected the ten most prominent newspapers, those with the highest number of publications. We provide a basic computational overview of the corpus to characterise its content. Given that the OCR quality of the corpus is low (yet on par with similar historical OCR-ed newspapers, (Kettunen & Pääkkönen, 2016)), we were interested in whether we can extract meaningful newspaper characteristics using keyword analysis, word frequencies, and concordances. The results are threefold. We provide an overall quantitative-based description of the newspapers, give insight into the historical development of the Slovenian language, and present an overview of OCR errors. By providing an overview of the corpus that would take incredibly long to complete in the absence of digitisation and annotation, we argue that annotated historical editions are extremely valuable for the Slovenian research community.

The paper is structured as follows. First, we present related work on historical newspaper analysis and the historical context of the selected journals. Second, we describe the corpus and the selected subset of ten periodicals. We characterise the newspapers with keyword analysis, which shows the specifics of each newspaper, and the list of most frequent nouns, which shows the general orientation of the newspaper. We compare the periodicals in terms of their thematic, regional, and religious orientation. Third, we critically evaluate the results and suggest potential post-processing of the published corpus based on the keyword analysis. In conclusion, we sum up the findings and present the options for future research.

2 RELATED WORK

Historical newspapers are used extensively in digital humanities, mostly due to contemporary digitisation efforts, accessible interfaces for content exploration (Ehrmann et al., 2019), and open repositories. The studies range from diachronic and comparative analyses to discourse studies, with concept shift analysis being the most prominent.

Comparative studies focus on cross-country comparisons (Mayer et al., 2022) or exploring regional differences (Park & Cordell, 2023). Diachronic studies often focus on concepts shifts (Verheul et al., 2022; Marjanen et al., 2020; Pivovarova et al., 2019), semantic change (Pedrazzini & McGillivray, 2022), or topic shift over time (Marjanen et al., 2021). Another branch of studies entails a more content-oriented approach, focusing on the emergence of public discourses (Marjanen et al., 2019) or nation-building vocabularies (Schoots, 2023; Hengchen et al., 2021). Some studies also focus on multilingualism (Marjanen et al., 2019; Mayer et al., 2022), a common trait of historical newspapers that makes comparative analysis particularly challenging.

Outside of digital humanities, Slovenian historical newspapers are a popular research topic. The overwhelming share of the studies focus on the nation-building processes, particularly after the 1848 March Revolution[1] (Stergar, 1977). The most comprehensive study is done by Smilja Amon, who presents an overview of Slovenian journalistic efforts (Amon, 2008). Ljubljanski zvon itself provides a great overview of the newspapers in 1885 (Anonymous, 1885). It lists 34 papers published in Slovenian, with a description, editor, publisher, and price. The final overview found 8 political papers, 3 political-economic, 4 economic, 4 religious, 4 legal, 2 pedagogical, 5 literary, 1 political-literary, and 3 humorous-satirical.

Other studies primarily focus on Kmetijske in rokodelske novice (Mihelič, 1948), which pioneered journalism in the Slovenian language[2]. Linguistic analyses are similarly popular. Only a few studies focus on content analysis and comparison. One such study is done by Štepec (Štepec, 1987), who analyses reporting on crime in Slovenec and Slovenski narod. Štepec ascertains that the conservative Slovenec leaves the reporting on crime primarily to the liberal Slovenski narod, as they see the reporting on crime as un-Catholic and not serving any purpose. Other research on historical newspapers focused on the language question in Slovenski pravnik (Zorn, 1987), news about Istria (Marušič, 2007), fashion in women's journals (Ilich, 1999), and social-democrat periodicals (Kermavner, 1962).

[1] The period before the 1848 March Revolution is typically referred to as pre-March or Vormärz. In the paper, we refer to the subsequent period as post-March.

[2] The first Slovenian-language periodical was Lublanske novize by Valentin Vodnik in 1797, but they were short-lived.

3 SPERIODIKA

sPeriodika (Dobranić et al., 2023) is a corpus of Slovenian historical newspapers from the 18th, 19th and 20th century. The corpus was created by Dobranić et al. (Dobranić et al., 2024). As stated by the authors, the basis are the OCR-ed data produced by different technologies in varying periods by the National and University Library of Slovenia, upon which they performed additional cleaning and preprocessing. It is available on the CLARIN.SI repository and in the noSketch Engine concordancer.

3.1 Description

There are 216 newspapers in the sPeriodika corpus with varying number of publications (max 28406, min 1). The total number of publications is 148457. As there is a significant long tail in the distribution of publications per newspaper, we decided to analyse the ten newspapers with the highest sum of publications, which represents 78% of the corpus. We decided on such metric to capture the periodicals with the largest national presence and a sufficient time span (Figure 1). Table 1 shows the ten selected newspapers with the number and share of publications (rounded to two decimal points). The papers' titles carry meaning, which broadly defines their content: Agricultural and Artisan News (Kmetijske in rokodelske novice), Slovenian holder[3] (Slovenski gospodar), Teacher's Companion[4] (Učiteljski tovariš), Slovenian Nation (Slovenski narod), Home and World (Dom in svet), The Slovenian (Slovenec), Unity (Edinost), The Ljubljana Bell (Ljubljanski zvon), (Kinder)garten (Vertec), Soča[5].

3.2 Keyword comparison

We used noSketch Engine to extract keywords for all ten periodicals. We compared them to the entire corpus, meaning we extracted words (lemmas as extracted by noSketch) that are highly represented and thus statistically significant for a given subcorpus. Lemmatization was done with the CLASSLA-Stanza

[3] Gospodar can mean a holder, a lord, a master.
[4] Tovariš can mean a companion or a comrade. The newspaper became related to politics only after 1900.
[5] Soča is a river in Western Slovenia.

Newspaper	no. of publications	% publications	no. of tokens
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 (Ljubljana) (DS)	11073	7	32,326,449
Slovenec (1873) (SVN)	10897	7	137,506,802
Edinost (Trst) (ED)	8371	6	98,274,429
Ljubljanski zvon (LZ)	3923	3	15,590,800
Vertec (1871) (VT)	3515	2	3,170,465
Soča (SČ)	3367	2	38,879,707

Table 1: Newspapers with the highest number of publications in the sPeriodika corpus.

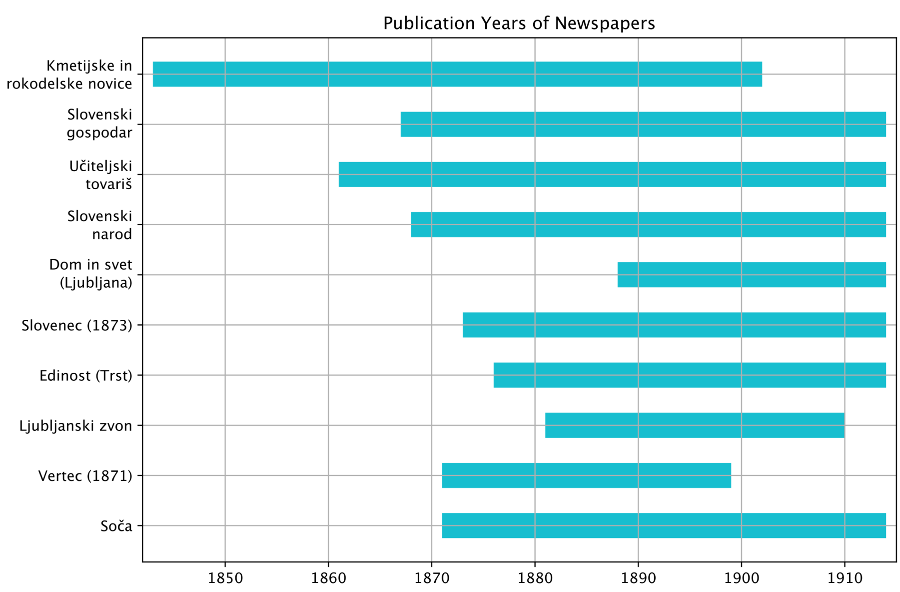


Figure 1: Publication years for the ten selected periodicals.

pipeline, as reported in the original sPeriodika paper (Dobranić et al., 2024). The keyness score, upon which the keywords are identified in noSketch, is computed with a simple maths method (Kilgarriff, 2009) and a smoothing parameter N=1 (default).

The formula for the keyness score, also known as ”simple maths” (Kilgarriff, 2009), is as follows:

$$\frac{fpm_{rmfoucs}+N}{fpm_{rmref}+N}$$

where $fpm_{rmfoucs}$ is the normalised (per million) frequency of the word in the focus corpus, fpm_{rmref} is the normalised (per million) frequency of the word in the reference corpus, and N is the smoothing parameter.

We analyse the top hundred words and present the first ten in Table 2. We omit the obvious OCR errors because we want to demonstrate the key content of the periodical, not the accidental errors. We report the number of OCR errors (per cent of errors in 100 hits) in the final row.

Kmetijske in rokodelske novice is true to its name. It discusses agricultural topics (kmetovavec, žlahen[6], žebec[7]) and regional news (Kranjska). It was the first full-fledged Slovenian language newspaper, and as such, it contains certain archaic words more than the other newspapers (onidan, en malo). The remaining words cover diverse categories, from newspaper sections (novičar) and finances (dnar) to news on Russia (rusovski) and national enlightenment topics (čitavnica[8]). Keyword analysis testifies to the wide variety of topics the newspaper covered and its longstanding central role in the cultural life of Slovenians in that period (Stergar, 1977).

Slovenski gospodar is the first paper in the list heavily affected by errors in OCR (94 %). The 94 % error rate refers to the keyword analysis results, not the entire periodical content. Inspection in a concordancer reveals that, typically, the letter n is transcribed as a (sloveaski –> slovenski, aaš –> naš, aemški –> nemški) and v as 7 (pra7). Other keywords reveal that mistaking č for 6 is also common. There are also mentions of Stajerc, which is a misliteration of Štajerc.

[6] Žlahen means noble and refers to different breeds, from cattle and bulls to fruit trees.
[7] Žebec is an archaic word for žrebec and means stallion.
[8] ”Ćitavnica” was more frequent in the earlier editions of KRN, where it was later substituted with ”čitalnica”.

It sometimes means a person from the Styria region, but most commonly, it refers to the Štajerc periodical, published between 1900 and 1918. The tone is very derogatory since Slovenski gospodar was a pro-Catholic and conservative periodical, while Štajerc was a more progressive pro-German newspaper (extensively described in (Jezernik, 2022). The few relevant keywords refer to fairs (sermon), happen (izgoditi), golden coin (fl), school union (šulverein), people (Dr Franc Radaj, MP; Franc Kosar), esteemed (vlč, velečastiti), and posilinemec (a mocking expression for pro-German Slovenians).

Učiteljski tovariš is also true to its name. Most keywords refer to pedagogy (zavezin[9], konvikt[10], učiteljstvo, učiteljski, lehrerbund, pedagoški,koleginja, ljudski). There is a political aspect to the debate with continuous mentions of ”Slomškar”, which refers to the competing ”Slomšek Union”, a union of Catholic teachers. As for ”tovarišica” (comrade, colleague, teacher), it is unclear whether the word has a political connotation or not, even from collocations. However, the two references to female colleagues (tovarišica and koleginja) are highly represented in Učiteljski tovariš, indicating the periodical perhaps treated female colleagues with a higher degree of equality. The periodical does have a much higher frequency of mentions of the two words relative to the general corpus. However, collocations do not reveal any special differences in context. Učiteljski tovariš also has a high degree of German loanwords (Lehrerbund, Lehrer, Volkschule, Lehrerschaft, Gesuche, Vorgeschriebenen) and mentions of people (Črnagoj[11], Jelenc, Maier, Strmšek, Režek, Požegar, Gangl).

Keyword analysis of Slovenski narod reveals many specific sections from the newspaper. The newspaper regularly published train schedules for Austrian railways (amstetten, pontabel, selzthal), reports from the Vienna stock exchange (prior oblig.), meteorological reports (wind directions), and specific advertisements (Moll Seidlitz powder, Revaliescere du Barry, Berger Kotran soap). Some words refer to the leading paragraph in the paper, which gave instructions for submissions to the paper (izvoti[12], četiristopne). There are a few OCR errors characteristic of the Slovenski narod, perhaps due to the font choice

[9]Zaveza refers to the Association of Austrian Yugoslav Teaching Unions
[10]Konvikt is an educational facility with a full board, mostly for priests.
[11]Fran Črnagoj was a teacher and businessman.
[12]This is a wrong lemma of ’izvole’, which means ’should it please them’.

(tuđi, tuđ[13], ćel[14]) Some of the results might be the result of over-correction, as Dobranić et al. (Dobranić et al., 2024) report statistic-based joining of split words (Trammwaydrušt, Stražatoplice).

Dom in svet (Ljubljana) is heavily literary and art-oriented. The periodical is characterised by the names of fictional characters (bodriški nadknez Gotšalk, Viljenica, Virida, Maruška, Ančka) or authors (Podgoričan) of stories the paper continually published. Much of their news mentions art pieces (spominiki, bilina, pasionski) and references publications (a text on cuneiform memorials, written by F. Sedej and published in the same newspaper). The most surprising is the heavy influence of the Slavic art world on the paper. Dom in svet regularly writes biographies of Central, Eastern, and Southern Slavic authors, and lists Slavic publications (especially in Russian, Serbian, and Croatian).

Similarly to Slovenski narod, keyword analysis of Slovenec revels specific sections of the newspaper, for example, reports from the Vienna stock exchange (vravnaven, salmov, dunavski, napoleondor, napoleond[15], waldsteinov), meteorology report, and a feuilleton Pismo Boltatovega Pepeta[16], written in a dialect (gespud, tku, kokr). There are some recurring advertisements, for example, for the Mercur Exchange Limited Company (kurzen), glass-making workshop, and an oil paint store. Several keywords refer to South-Eastern Europe (Croatia, Hungary, Bulgaria), slightly denoting the political orientation of the periodical. However, we expected a much higher ratio of political keywords due to the newspaper's importance in the Slovenian political space. Many keywords stem from the newspaper's header, where practical information on the subscription and distribution of the periodical was given. However, other periodicals, such as Slovenski narod, Slovenski gospodar, Edinost and Soča, also had a substantial header. The high prevalence of header words is perhaps due to the linguistic specifics of Slovenec's header.

Edinost (Trst), a paper published by the Slovenes in Italy, specifically Trieste, contains many marketing-related words. Many of them refer to streets or locations of business (barriera, nuova, vecchia, piazza, galatti), specifically, 68 %. Most are Italian street names, but there are also mentions of Istrian towns

[13] Both versions of tudi, meaning also.
[14] Correctly celo or čelo, meaning even or forehead.
[15] Both terms are literal transcription of napoléon d'or, a gold coin from France.
[16] A pseudonym for Srečko Magolič (Steska & Stelè, 2013).

(Pula, Rovinj). Edinost covered the Istrian region until 1902 when the Political Society of Croats and Slovenians in Istria was formed (Darovec, 2023, 66). When mentioning the Primorska region, mentions relate mostly to the weather forecast and the journal's subtitle (Glasilo političnega društva "Edinost" za Primorsko). There are also mentions of currencies (nvč is an abbreviation for "novčič", a coin at 1/100 of "goldinar") and advertisement space (inseratni refers to the newspaper's department for advertisement). Advertisement space is characterised by recurring ads for coffee (kava Santos good average), health services (izdiranje, plombiranje, ambulatorij), and food items (pekarna, butejka). Like other periodicals of the time, Edinost regularly published train schedules. "Medpostaja" and "Pula" are mostly used in the context of railway schedules, similar to Slovenski narod, but focused on Italian railways. Railway schedule news items show that the paper is very practical; it offers advertising space for local businesses and gives information on transportation. Many periodicals of the time had similar information (e.g. Slovenski narod).

Ljubljanski zvon was the leading literary publication of the time. Most of the top ten keywords contain references to literary characters (gojko, samorad, trenk, abadon, zdenka.). 29 % of keyword results are character names, highlighting the literary nature of the periodical. However, not all content was fictional. There are references to Slovniški razgovori (Grammatical discussions), where the periodical published lectures on proper Slovenian spelling and grammar (sedanjik, sgl, miklosich, dovršnik), and Štrekelj's Jezikoslovne mrvice (Linguistic nuggets), where the author was explaining the grammatical composition, meaning, and origin of certain words (subst). Many keywords are OCR errors, namely 36 %. The keyword issue with Ljubljanski zvon is somewhat particular. It is not only that, similarly to Slovenski gospodar, the top keywords are incorrectly transcribed (OCR-ed) words. The errors are linked intimately to the literary nature of the periodical. It is the only periodical selected for analysis that consistently uses diacritics on vowels. Diacritics are uncommon in Slovenian, but in the specific newspaper, it was likely used to stress the rhythm and correct pronunciation of the word. However, this stylistic choice causes issues for the OCR model.

Vertec (1871) contains many stories and is, thus, similar to Dom in svet and Ljubljanski zvon, characterised by literary characters (Marijca, Marijec[17], Katarinka, Ivanek). The ratio of literary character mentions in keyword results is 38 %. Unlike in other periodicals, the names are predominantly diminutives, reflecting the newspaper's orientation towards the youth. However, sometimes, a name refers not to a literary character but to a real person. The periodical listed the authors of correct solutions for its puzzles by name and location. Other keywords are bucolic, family- or nature-oriented (dedek, sestrica, ptičica, čmrlj, lisica). OCR error rate for this periodical is fairly high, at 36 %.

Soča published several translated works, including Alexandre Dumas' Three Musketeers (Athos, Porthos, Artagnan, Aramis) and Count Monte Cristo (Villefort), Henryk Sienkiewicz's Quo Vadis? (Vinicij) and The Knights of the Cross (Zbišek), and Maxim Gorky's Foma Gordeyev. The keywords, in total, include 23 % of character names. There are some regional specialities in the newspaper, for example, the word "nunc", which in the Gorizia dialect refers to an older familiar man. The regional character is also reflected in the mentions of local political figures, such as Alojzij Pajer-Monriva, a pro-Italian lawyer and politician, and Ivan Berbuč, a political and co-editor of Soča. A fun finding is the keyword "prismojenec"[18]. "Prismojenec" is a nickname for Primorski list, a conservative periodical standing in opposition to Soča, similar to how Slovenski gospodar stood in opposition to Štajerc. Soča, on the other hand, was content-wise more similar to Slovenec (Marušić, 2005, 326). The periodical contains 53 % OCR errors, making it one of the most difficult periodicals to analyse. A typical OCR error for this specific periodical is the omittance of the caron (uze[19], dezelni, drzaven, goriski, u2e). Moreover, the periodical has low-quality images, making OCR errors even more likely.

3.3 Characterisation by nouns

To further characterise the periodicals, we retrieved lists of the most frequent nouns for each journal from the noSketch Engine 3. While keywords describe the particularities of each journal compared to the entire corpus, they are often skewed towards OCR errors and coincidental recurring feuilleton full of literary

[17] Marijec is an erroneous lemma for the word Marijca
[18] Prismojenec in Slovenian means a wacko.
[19] Originally uže.

characters. To better understand the general nature of each journal, we observed the most frequent nouns. We decided on nouns to avoid having too many stopwords in the results. Nouns are, generally, a good indicator of the content.

It turned out, nouns were not very informative. All periodicals contain words pertaining to the journalistic form, i.e. dates and place names. Some results are OCR errors, which are the most frequent in Slovenski narod, Edinost, and Soča. It is expected the OCR error rate to drop, since we only asked for CLASSLA-identified nouns. However, Edinost was the only periodical where the error rate increased significantly. The newspaper is so highly characterised by place names and advertisements, that they overtook OCR errors in keyword analysis results. The errors in noun results are comparable to other periodicals. Errors aside, the most frequent 100 nouns reveal a general orientation of each newspaper.

Kmetijske in rokodelske novice prominently features the words country (dežela) and city (mesto), showing the newspaper's focus on the city and countryside relations. It contains references to politics (zbor, vlada, odbor, poslanec) and to national identity (narod, beseda, jezik). Due to its popularity, the periodical became a central publication for the Slovenian national movement (Dović, 2023). Slovenski gospodar similarly references politics (zbor, društvo, poslanec, volitev, okraj) and city-countryside relations while also showing its religious orientation (cerkev, nedelja). Učiteljski tovariš is highly focused on pedagogy (šola, učitelj, učiteljstvo, otrok, učiteljica, knjiga, učenec), with some organisational words inbetween (društvo, svet, zbor, odbor). Vertec is also domain-specific, generally focusing on family (mati, oče, otrok) and storytelling (človek, bog, čas, mesto, hiša). Slovenski narod, while littered with single letter "nouns" (i.e. errors), appears as a mostly political periodical (narod, zbor, vlada, Dunaj, Slovenec, stranka). Slovenec is another politically oriented periodical (vlada, mesto, društvo, zbor, narod). Dom in svet is more contemplative, mostly discussing the position of the man in the world (human, time, work, life, world, heart), with literary (knjiga, pisatelj, pesem, jezik) and religious emphasis (cerkev, bog, duša). Ljubljanski zvon is similarly contemplative but with much greater literary emphasis (knjiga, človek, beseda, življenje, delo, jezik, srce, narod) and a lack of religious themes. Regional identity is stressed in Edinost (Trieste being highly ranked) and Soča (Gorizia). Both periodicals show political/national orientation (društvo, narod, vlada).

Kmetijske in rokodelske novice, Slovenski narod, Učiteljski tovariš, Slovenec, and Ljubljanski zvon also frequently mention Ljubljana, showing their central geographical orientation. Conversely, Slovenski gospodar mentions Maribor instead, revealing its focus on the Styrian readers.

Strong nation-building forces in Slovenia defined the late 19th and early 20th centuries. The media landscape of the time greatly contributed to forming and expanding ideas of national identity, Slovenian culture, and political emancipation (Amon, 2008). Digitised editions of historical papers enable observing, comparing, and quantifying nation-building discourses. Below, we provide a quick glimpse into two nation-building aspects of Slovenian historical newspapers: the emergence of the ethnonym Slovenian and a comparison of post-March revolution discourses in Kmetijske in rokodelske novice.

7 of 10 periodicals have the word Slovenian (Slovenec) among the top 30 most frequent nouns. Upon inspection, the word Slovenian appeared only after 1843 when Bleiweis's Kmetijske in rokodelske novice was first published. The lack of the ethnonym "Slovenian" before 1843 can be partially attributed to many periodicals before this year being in German due to the strict pre-March censorship (Dović, 2006). However, as Dović argues (Dović, 2023), Novice pioneered the ethnonym Slovenia and Slovenians into Slovenian periodicals of the post-March period.

We examined collocations for the word Slovenec to determine whether the mentions mostly refer to the periodical Slovenec or the ethnonym. The most frequent collocates are Croats, Carinthian, and Trieste[20]. References to the periodical come in at fourth place, where the collocation is the quotation mark. Thus, mentions of the periodical appear in quotations, while the ethnonym appears without them.

[20]as an adjective for the Slovenes in Italy

Table 2: Top 10 keywords (lemmas) in selected periodicals. The cells contain a lemma and its frequency in the given periodical. The final row reports a percentage of OCR errors in top 100 keywords.

Rank	KRN	SG	UT	SN	DS	SVN	ED	LZ	VT	SČ
1	unidan (1,552)	sejmov (843)	zavezin (2,265)	amstetten (11,058)	nadknez (738)	vravnaven (3,299)	nvč (12,057)	gojko (889)	marijca (269)	athos (2040)
2	novičar (3421)	izgoditi (481)	konvikt (5,486)	izvoti (7,416)	virida (798)	gespud (3,447)	galatti (5,504)	samorad (679)	otiti (475)	porthos (1,411)
3	čitavnica (2,044)	fl (12,467)	učiteljstvo (54,905)	pontabel (6,225)	spominik (1,029)	tku (4,680)	barriera (7,162)	trenk (713)	štir (368)	artagnan (1,369)
4	rusovski (1,714)	šulverein (677)	učiteljski (58,083)	selzthal (8,551)	bodriški (631)	salmov (2,996)	inseraten (7,641)	abadon (549)	vrtčev (220)	aramis (1,253)
5	kmetova-vec (2,481)	radaj (541)	slomškar (1,244)	oblig (6,752)	viljenica (638)	kokr (3,535)	nuova (7,977)	zdenka (826)	katarinka (172)	nunec (1,946)
6	dnar (2,238)	vlč (903)	tovarišica (4,632)	franzensfe-ste (7,256)	juriš (912)	napoleon-dor (3,206)	konsorcija (5,091)	groga (1,046)	ivanek (181)	zbišek (1,004)
7	žlahen (1,433)	kosar (673)	koleginja (1,031)	četiristo-pen (3,690)	gotšalk (610)	kursen (2,771)	pula (7,343)	cetinovič (334)	pesenca (203)	meljavec (928)
8	krajnski (3,076)	posiline-mec (463)	lehrer-bund (902)	steyr (5,488)	maruška (670)	dunavski (4,189)	vecchia (6,292)	dramatiški (642)	marijec (155)	villefort (846)
9	žebec (632)	-	pedagoški (2,796)	osoben (28,671)	podgori-čan (996)	waldstei-nov (2,349)	medposta-ja (3,331)	obsezati (1,943)	vzpomlad (176)	vinicij (821)
10	enmalo (823)	-	črnagoj (779)	vara (13,567)	ančka (1,407)	napoleond (2,234)	piazza (14,364)	premec (381)	ivanko (170)	foma (916)
errors	5%	92%	12%	19%	1%	15%	0%	36%	36%	53%

Table 3: Top 10 keywords (nouns) in selected periodicals. The cells contain a noun and its frequency in the given periodical. The final row reports a percentage of OCR errors in top 100 keywords.

Rank	**KRN**	**SG**	**UT**	**SN**	**DS**	**SVN**	**ED**	**LZ**	**VT**	**SČ**
1	leto (82,258)	dan (69056)	šola (114,848)	leto (443,770)	leto (59,446)	leto (343,343)	Trst (258,913)	leto (39,468)	mati (6,266)	Gorica (80,144)
2	dan (68,933)	leto (51,176)	učitelj (97,051)	dan (408,203)	dan (48,748)	dan (321,052)	dan (216,971)	dan (27,830)	dan (5,835)	dan (75,524)
3	zbor (39,665)	zbor (28,349)	leto (67,049)	Ljubljana (302,647)	človek (42,447)	Ljubljana (229,561)	leto (157,861)	knjiga (20,532)	leto (5,744)	leto (69,817)
4	čas (37,683)	društvo (27,645)	dan (58,550)	ura (237,529)	čas (37,518)	ura (167,718)	ulica (140,912)	čas (19,748)	oče (4,515)	društvo (42,044)
5	človek (32,521)	poslanec (23,641)	učiteljstvo (54,905)	mesto (192,520)	delo (35,653)	vlada (153,937)	ura (130,535)	človek (16,778)	otrok (4,499)	zbor (41,046)
6	Ljubljana (30,918)	Slovenec (23,277)	društvo (50,566)	društvo (185,375)	življenje (34,640)	mesto (148,402)	društvo (124,234)	gospod (15,485)	človek (4,279)	Slovenec (35,437)
7	dežela (30,825)	kmet (23,255)	svet (40,158)	narod (181,277)	svet (33,444)	društvo (141,115)	cena (104,929)	mesto (14,896)	bog (3,788)	mesto (34,216)
8	mesto (30,551)	šola (23,213)	otrok (36,984)	zbor (166,316)	knjiga (33,096)	zbor (140,109)	mesto (91,023)	beseda (14,832)	čas (3,334)	čas (33,146)
9	šola (30,037)	človek (21,664)	Ljubljana (30,369)	vlada (164,447)	srce (31,636)	čas (134,440)	vlada (83,082)	nega (12,902)	gospod (3,060)	ura (32,886)
10	kraj (28,716)	mesto (21,129)	čas (29,185)	čas (163,074)	mesto (30,324)	narod (131,146)	ulica (81,713)	pisatelj (12,759)	roka (2,939)	gospod (31,052)
errors	5%	17%	13%	21%	3%	15%	22%	7%	18%	20%

4 DISCUSSION

We characterised the periodicals using two keyword analysis approaches, one with lemmas and one with nouns. This portrays a landscape of periodical at the turn of the 19th century, and supplements previous manual analysis of Slovenian newspapers. Periodicals are typically characterised by their proclaimed focus (KRN, Učiteljski tovariš), feuilletons and advertisements (Dom in svet, Slovenski narod itd.), or, alas, their OCR errors (Slovenski gospodar).

The identified importance of feuilletons and advertisements aligns with the previous research on historical Slovenian periodicals. Feuilletons, a part of a newspaper devoted to fiction, played an important role in the development of the Slovenian prose (Dović, 2006). Feuilletons were the first public venue for Slovenian authors to publish their work and reach a wider audience. Of course, keyword analysis only pointed to specific literary characters, which is expected as the technique determines words that appear uniquely in the subset. Thus, one cannot say that keyword analysis pinpointed the importance of feuilleton – the discovery was incidental.

On the other hand, the importance of advertising space was better characterised by the method. The ratio of editorial to advertisement space was at 4:1 in the late 19th century (Dović, 2006), making advertising space an extremely relevant part of the newspaper. While some keywords point to specific advertisers, they also point to the general advertising language (inseraten, nvč).

Keyword analysis reveals that the periodicals were published when the standard Slovenian language was still being formed. Many papers are characterised by their specific writing of standard Slovenian words. Almost every periodical has a set of words that characterise their approach to Slovenian spelling. For example, Kmetijske in rokoldelske novice writes nograd for vinograd (vineyard) and berž for brž (as soon as). Slovenski narod writes denes for danes (today) and sklenica for steklenica (bottle). Edinost writes menenje for mnenje (opinion), zvršetek for konec (end), and žnjo/žnjimi for z njo/z njimi (with her, with them). Vertec writes otiti for oditi (leave), vzpomlad for spomladi (in spring), and rekši for je rekel/rekla (said). Even Ljubljanski zvon, which was at the forefront of gramamtical efforts at the time, contains words that are considered archaic in modern Slovenian, i.e. obsezati for obsegati (to cover), zanimljiv for zanimiv (interesting), smijati for smejati (to laugh).

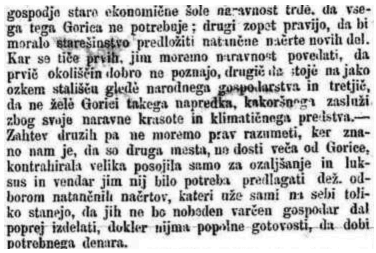
Finally, some periodicals have too many OCR errors to properly characterise them by content (Slovenski gospodar, partially Soča). OCR errors were calculated on keyword analysis results, which provide the 100 most characteristic keywords for a given subcorpus. Out of these, we manually labelled OCR errors and summed them. We considered a lack of carons an OCR error since the word without carons is counted as distinct from the word with a caron (drzaven vs državen) or can mean a different word altogether (čelo/celo). Percentages are reported for total OCR results. There were, in total, 1000 keyword results, which contained 266 errors. Note that periodicals were digitised using different OCR models, thus leading to periodical-specific errors.

Some OCR errors are recurring and reflect an underlying weakness of OCR models. The most common errors (24 %) are the letters n, s, or š transcribed as a. The errors are most common in Slovenski gopodar, the periodical most affected by OCR errors. The second most common error (21 %) is the lack of diacritics (stajerski, drzaven), while in the third place (9 %) are diacritics transcribed as numbers, specifically 6, 7 or 2 (dom6v, rek6, už6, u2e, pra7). Diacritics are often also transcribed as d (takdj). The preceding letter n often means the word begins with quotation marks (nkaj, nne, njaz). Conflation is very common, with č and e (oee, užč), i and l (ijubi, nefranklran), c and e (Marijea, evetice), and u and n (nčenki) conflated.

Diacritics and carons pose a particular problem in the transcription of sPeriodika. Here is an example from Ljubljanski zvon, the only periodical that regularly uses diacritics on vowels to denote word stress (Vertec uses them occasionally):

1. *Takó kričálo vse je gôri náme.* (original)
2. *Takd kričdlo vse je g6ri ndme.* (transcript)
3. *'ma krimu' vse ie gori "Am,* (tesseract)

The mistakes visually make sense. ó and á are transcribed as d (or occasionally 6), ô as 6, á also as ä, é as č. Nevertheless, issues with transcription limit the semantic analysis of significant keywords.



gospodje staro ekonomične šole naravnost trdė, 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 stojė na jako ozkem stališču gledė narodnega gospodarstva in tretjič, da ne želė Gorici takega napredka, kakoršnega zasluži zbog svoje naravne krasote in klimatičnega predstva.— Zahtev druzih pa ne moremo prav razumeti, ker znano nam je, da so druga mesta, ne 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.

Figure 2: A section of the Soča newspaper, with a poor scan quality.

Vertec has specific OCR errors. While not exclusive to this periodical, Vertec suffers strongly from character conflation. Characters and character sets m, u, and ru are often wrongly transcribed. The errors stem from character (set) similarity; thus, m is transcribed as ra, ni, or in. U is transcribed as ii, and ru as ni or m. V is often transcribed as r, ó as d or 6.

Slovenec and Edinost have a different problem. In Slovenec, 29 % of keywords refer to the newspaper's header. In Edinost, 68 % of keywords refer to Italian street names. These results do not tell much about the content besides a heavy influence of specific periodical sections. There are better techniques than keyword analysis to determine the journal's content in both cases.

For periodicals with frequent errors in the top keywords, we compared the frequencies of wrongly ORC-ed words to their original form. The erroneous "sloveaski" appears 1,855 times in the corpus, while the correct version "slovenski" appears 45,759. The top keywords cannot be analysed semantically, as all occurrences of the wrong word should be first converted to the correct form. However, the error is significantly more frequent in Slovenski gospodar compared to any other periodical. The discrepancy in frequencies means the error word characterises this particular journal and could be used as a part of post-processing. In other words, such erroneous words could be subsequently corrected in the selected publication. As Strange et al. demonstrate (Strange et al., 2014), OCR correction can be crucial for certain text analysis techniques, such as keyword analysis (and less for others).

Alternatively, the error rate could indicate candidate journals for re-scanning. Certain scans are already of poor quality or were among the first periodicals OCR-ed. Contemporary OCR solutions could provide a much better result than the existing version. However, scanning the entire corpus again is tedious and unnecessary. Thus, having a list of candidate periodicals for re-scanning is a good way forward. Based on our results, Slovenski gospodar and Soča (2) would benefit from both re-scanning and a modern OCR transcript, while Ljubljanski zvon would benefit only from an OCR (scans are already good).

However, contemporary state-of-the-art large language models (LLMs) can already achieve incredible transcription accuracy. Here is a GPT4-o transcription of Figure 2: *"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."*

The capabilities of LLMs and large multimodal models (LMMs) can overcome poor scan quality almost out-of-the-box. They outperform modern OCR solutions in post-OCR correction (Thomas et al., 2024) and in direct OCR (Liu et al., 2024), even for complex compositions such as old Chinese newspaper clippings (Chow, 2024). This opens up exciting venues for historical research (Garcia & Weilbach, 2023), especially in addressing corpora quality (OCR), but also for content summaries, event detection, trend analysis, and semantic search.

5 CONCLUSION

Keyword analysis reveals several aspects of the periodicals. Some papers are characterised by their general content, such as agriculture (Kmetijske in rokodelske novice) or pedagogy (Učiteljski tovariš). Some are characterised by the recurring feuilletons they publish (Dom in svet, Slovenec, Vertec, Soča). Others still are characterised by their advertising space (Slovenski Narod, Edinost). Slovenski gospodar, unfortunately, contains too many OCR errors for keyword analysis to reveal meaningful insights. Consistent OCR errors in the periodicals could be addressed in post-processing.

We substantiated the results with the most frequent nouns to alleviate the issues with keyword analysis. Many newspapers of the time focused on Slovenian nation-building, either through reports on (inter)national relations, discussions of politics, or debates on the role of the language (Kmetijske in rokodelske novice, Slovenski gospodar, Slovenski narod, Slovenec, Edinost and Soča). Among these, city vs countryside relations are a prominent topic (Kmetijske in rokodelske novice, Slovenski gospodar). Učiteljski tovariš and Vertec are domain-specific, rarely discussing topics outside their proclaimed focus. Slovenski gospodar and Dom in svet reveal the highest religious orientation. However, many periodicals of the time discussed the role of religion in nation-building.

The computational overview provides several opportunities for further analysis. For example, one could comparatively analyse the first two Slovenian daily papers, the liberal Slovenski narod and the conservative Slovenec. A similar comparative analysis could be applied to Edinost and Soča, the two periodicals of Slovenes in Italy, analysing their overlap and divergences (especially considering their merger intentions). A much more demanding research could consider analysing differences in advertisements, given that they feature prominently even in keyword analysis. The task is complex because it is extremely difficult to set the boundaries of individual advertisements. The problem could be approached by treating periodicals as images (van Galen, 2023) and using neighbour search to find similar advertisements. LLMs can be used for all of the above tasks, which shows how this technology will revolutionise historical research in the future, especially when dealing with lower-quality corpora.

ACKNOWLEDGMENTS

I would like to sincerely thank dr. Nikola Ljubešić and Filip Dobranić for their invaluable contribution to this work. The work described in this paper was funded by the Slovenian Research Agency research programme P6-0436: Digital Humanities: resources, tools and methods (2022–2027) and the DARIAH-SI research infrastructure.

REFERENCES

Amon, S. (2008). Vloga slovenskega časopisja v združevanju in ločevanju slovenske javnosti od 1797-1945. *Javnost*, *15*, S9-S24.

Anonymous, L. (1885). Slovenski časopisi leta 1885. *Ljubljanski zvon*, *5*, 631-635.

Chow, E. H. C. (2024). *An experiment with Gemini Pro LLM for Chinese OCR and metadata extraction.* Retrieved April 5, 2024, from https://digitalorientalist.com/2024/04/05/an-experiment-with-gemini-pro-llm-for-chinese-ocr-and-metadata-extraction/

Darovec, D. (2023). *Pregled zgodovine Istre*. Koper: Založba Annales.

Dobranić, F., Evkoski, B., & Ljubešić, N. (2023). *Corpus of Slovenian periodicals (1771-1914) sPeriodika 1.0.* http://hdl.handle.net/11356/1881 (Slovenian language resource repository CLARIN.SI)

Dobranić, F., Evkoski, B., & Ljubešić, N. (2024, May). A lightweight approach to a giga-corpus of historical periodicals: The story of a Slovenian historical newspaper collection. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 695–703). Torino, Italy: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.61

Dović, M. (2006). Literatura in mediji v Jurčičevem času. *Slavistična revija*, *54*(4), 543–557.

Dović, M. (2023). Anatomy of the "deathly silence": Slovenian newspapers in Carniola and the pre-March censorship. *Neohelicon*, *50*(2), 543–560.

Ehrmann, M., Bunout, E., & Düring, M. (2019, September). Historical newspaper user interfaces: A review. In *85th IFLA General Conference and Assembly (IFLA).* Athens, Greece: Zenodo.

Ehrmann, M., Düring, M., Neudecker, C., & Doucet, A. (2023). Computational approaches to digitised historical newspapers (Dagstuhl seminar 22292). *Dagstuhl Reports*, *12*(7), 112–179. https://drops.dagstuhl.de/entities/document/10.4230/DagRep.12.7.112 doi: 10.4230/DagRep.12.7.112

Eiselt, I. (2015). Newspapers in the National and University Library in Slovenia–Access model. *Review of the National Center for Digitization*, *26*, 77–85.

Garcia, G. G., & Weilbach, C. (2023). If the sources could talk: Evaluating large language models for research assistance in history. In A. Šeļa, F. Jannidis, & I. Romanowska (Eds.), *Proceedings of the Computational Humanities Research Conference 2023* (p. 616-638). Paris, France.

Hengchen, S., Ros, R., Marjanen, J., & Tolonen, M. (2021, 11). A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital*

Scholarship in the Humanities, *36*(Supplement 2), ii109-ii126. doi: 10.1093/llc/fqab032

Ilich, M. (1999). Nekaj o modi v slovenskem časopisju na prelomu stoletja (1895-1915). *Zgodovina za vse*, *6*, 98-108.

Jezernik, B. (2022). Katoliška duhovščina na prelomu devetnajstega in dvajsetega stoletja in proces modernizacije na Slovenskem. *Traditiones*, *51*(1), 103–145.

Kermavner, D. (1962). Drugi slovenski socialnodemokratski listi. *Kronika*, *10*, 80-89.

Kettunen, K., & Pääkkönen, T. (2016, May). Measuring lexical quality of a historical Finnish newspaper collection – Analysis of garbled OCR data with basic language technology tools and means. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 956–961). Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1152

Kilgarriff, A. (2009). Simple maths for keywords. In *Proc. corpus linguistics* (Vol. 6).

Liu, Y., Li, Z., Yang, B., Li, C., Yin, X., lin Liu, C., ... Bai, X. (2024). *On the hidden mystery of OCR in large multimodal models.* https://arxiv.org/abs/2305.07895

Marjanen, J., Kurunmäki, J., Pivovarova, L., & Zosa, E. (2020). The expansion of isms, 1820-1917: Data-driven analysis of political language in digitized newspaper collections. *Journal of Data Mining & Digital Humanities*.

Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L., & Tolonen, M. (2019). A national public sphere? Analyzing the language, location, and form of newspapers in Finland, 1771–1917. *Journal of European Periodical Studies*, *4*(1).

Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., & Tolonen, M. (2021). Topic modelling discourse dynamics in historical newspapers. In S. Reinsone, I. Skadiņa, A. Baklāne, & J. Daugavietis (Eds.), *Digital humanities in the Nordic countries 2020* (pp. 63–77). Germany: CEUR-WS.org. http://dig-hum-nord.eu/conferences/dhn2020/

Marušić, B. (2005). *Pregled politične zgodovine Slovencev na Goriškem: 1848-1899*. Nova Gorica: Goriški muzej.

Marušič, B. (2007). Izbor vesti o Istri v slovenskem časopisju do leta 1880. *Annales*, *17*.

Mayer, A. I. L., Gutierrez-Vasques, X., Saiso, E. P., & Salmi, H. (2022). 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*(4).

Mihelič, S. (1948). Kmetijska družba in ustanovitev "Novic". *Slavistična revija*, *1*(1/2). http://www.dlib.si/?URN=URN:NBN:SI:DOC-HU751MKO

Park, J., & Cordell, R. (2023, December). A quantitative discourse analysis of Asian workers in the US historical newspapers. In M. Hämäläinen et al. (Eds.), *Proceedings of the Joint 3rd international conference on natural language processing for digital humanities and 8th International workshop on computational linguistics for Uralic languages* (pp. 7–15). Tokyo, Japan: Association for Computational Linguistics. https://aclanthology.org/2023.nlp4dh-1.2

Pedrazzini, N., & McGillivray, B. (2022, November). Machines in the media: semantic change in the lexicon of mechanization in 19th-century British newspapers. In M. Hämäläinen, K. Alnajjar, N. Partanen, & J. Rueter (Eds.), *Proceedings of the 2nd international workshop on natural language processing for digital humanities* (pp. 85–95). Taipei, Taiwan: Association for Computational Linguistics. https://aclanthology.org/2022.nlp4dh-1.12

Pivovarova, L., Zosa, E., & Marjanen, J. (2019, September). Word clustering for historical newspapers analysis. In C. Vertan, P. Osenova, & D. Iliev (Eds.), *Proceedings of the workshop on language technology for digital historical archives* (pp. 3–10). Varna, Bulgaria: INCOMA Ltd. https://aclanthology.org/W19-9002 doi: 10.26615/978-954-452-059-5_002

Schoots, J. (2023, May). Analyzing political formation through historical isiXhosa text analysis: Using frequency analysis to examine emerging African nationalism in South Africa. In R. Mabuya, D. Mthobela, M. Setaka, & M. Van Zaanen (Eds.), *Proceedings of the fourth workshop on resources for african indigenous languages (rail 2023)* (pp. 65–75). Dubrovnik, Croatia: Association for Computational Linguistics. https://aclanthology.org/2023.rail-1.8 doi: 10.18653/v1/2023.rail-1.8

Stergar, N. (1977). Narodnostno vprašanje v predmarčnih letnikih Bleiweisovih Novic. *Kronika (Ljubljana)*, *25*(3). http://www.dlib.si/?URN=URN:NBN:SI:DOC-WWIM1UTI

Steska, V., & Stelè, F. (2013). *Magolič, srečko (1860–1943).* Slovenska akademija znanosti in umetnosti, Znanstvenoraziskovalni center SAZU. http://www.slovenska-biografija.si/oseba/sbi339057/#slovenski-biografski-leksikon

Strange, C., McNamara, D., Wodak, J., & Wood, I. (2014). Mining for the meanings of a murder: the impact of OCR quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*.

Thomas, A., Gaizauskas, R., & Lu, H. (2024). Leveraging LLMs for post-OCR correction of historical newspapers. In *Proceedings of the third workshop on language technologies for historical and ancient languages (lt4hala)@ lrec-coling-2024* (pp. 116–121).

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

Verheul, J., Salmi, H., Riedl, M., Nivala, A., Viola, L., Keck, J., & Bell, E. (2022). Using word vector models to trace conceptual change over time and space in historical newspapers, 1840–1914. *Digital Humanities Quarterly*, *16*(2).

Zorn, T. (1987). Odmevnost jezikovnega vprašanja v listu Slovenski pravnik v letih 1871-1918. *Kronika*, *35*, 146-155.

Štepec, M. (1987). Zločin v slovenskem časopisju v 80. letih 19. stoletja. *Kronika*, *35*, 30-38.

KORPUSNO-JEZIKOSLOVNA ANALIZA KORPUSA SPERIODIKA

Prispevek predstavi računsko analizo sPeriodika, zgodovinskega korpusa slovenskih periodičnih publikacij od leta 1771 do 1914. Avtorica se osredotoči na deset pomembnih časopisov iz korpusa, pri čemer uporabi analizo ključnih besed, štetje pogostosti besed in konkordančno analizo za karakterizacijo vsebine in zgodovinskega razvoja slovenskega jezika. Študija opisuje značilnosti časopisov z računalniškimi metodami ter ugotovitve povezuje z obdobjem intenzivnega oblikovanja naroda po marčni revoluciji leta 1848. Poleg tega obravnava izzive, ki jih povzroča slaba kakovost optičnega prepoznavanja znakov (OCR) v zgodovinskih dokumentih. Rezultati so trojni: 1) kvantitativni opis izbranih časopisov, 2) vpogled v zgodovinski razvoj slovenskega jezika, 3) analiza narave napak OCR v korpusu. Analiza ključnih besed razkriva specifične tematske usmeritve časopisov, kot so kmetijstvo, pedagogika, podlistki in oglaševanje. Prav tako poudarja vlogo časopisov pri oblikovanju naroda. Študija prispeva k področju digitalne humanistike s prikazom, kako lahko računalniška orodja odkrijejo zgodovinske vpoglede iz digitaliziranih besedilnih podatkov, kljub omejitvam tehnologije OCR.

Keywords: zgodovinski časopisi, analiza ključnih besed, OCR napake, korpusno jezikoslovje

MAK NA KONAC: A MULTI-REFERENCE SPEECH-TO-TEXT BENCHMARK FOR CROATIAN AND SERBIAN

Tanja SAMARDŽIĆ,[1] Peter RUPNIK,[2] Mirjana STAROVIĆ,[3] Nikola LJUBEŠIĆ[2]

[1] Language and Space Lab, University of Zurich
[2] Jožef Stefan Institute, Ljubljana
[3] Leksikom, Belgrade

The evaluation of computational models for speech-to-text conversion has become especially needed in the context of the latest technological advances, which have led to the real usability of these models and strong market competition. This paper presents a new data set designed to address the challenging problem of objective model comparison. Instead of a strict objective evaluation in relation to one given solution, our proposal is a flexible evaluation on a variable test data set. The new data set consists of transcribed spontaneous speech samples from three sources (one Croatian and two Serbian) with a total duration of about 15 hours. Our initial comparison of six competitive speech-to-text systems shows stable patterns across the three sources: zero-shot deployment of a large multilingual model gives better performance than single-language training or fine-tuning on small data sets.

Keywords: speech-to-text, automatic speech recognition, multi-reference evaluation, benchmark, Croatian, Serbian

1 INTRODUCTION

The evaluation of computational models for speech-to-text conversion has become an important question in the context of modern models trained with transfer learning. The performance of these models has finally reached such a level that automatic transcription has become relatively easily accessible for many languages, including Croatian and Serbian. Everybody would like to take advantage of this new opportunity: media companies would like to convert their archives to text to allow efficient search, various companies would like to have meeting minutes compiled automatically from converted speech, doctors would like to capture and later study conversations with patients and so on. With such a great demand comes strong competition of offered solutions and the main question is: which solution to choose? An objective evaluation of model performance turns out to be surprisingly complicated.

The fact that almost every segment of speech can be correctly transcribed in different ways is often overlooked or neglected in the evaluation of speech-to-text conversion, especially in the case of orthographic transcription in highly standardised languages, such as Croatian and Serbian. *Piši kao što govoriš* 'write exactly the way you speak' is a famous motto in these languages, but when we try to implement it in creating a reference transcription, we come across many caveats. For instance, should we write *OK*, *okay*, *okei*, or *okej*? Each of these options is correct in some way. In theory, we can pick up one option, try to be consistent and train a model to output this one option, but the current practice of using pre-trained models via transfer learning makes this impossible. The problem is not only that we have no control over pre-training data, but also that the large quantities of data needed for pre-training necessarily lead to inconsistency. The large volume of data cannot be produced with a strict design but needs to be collected from existing sources, which are most likely inconsistent.

The aim of our paper is to introduce a new multi-reference corpus for testing Croatian and Serbian speech-to-text models. The new data set consists of transcribed speech samples with a total duration of about 15 hours. We show how this data set allows a more objective and more insightful comparison of model performance.

2 BACKGROUND AND MOTIVATION

Before motivating our proposal, we introduce the most important terms that are necessary for a better understanding of the evaluation problem.

Converting speech into text takes several steps. The sound wave is first divided into very short segments called *frames*, from which we extract the most relevant physical properties of the sound, called *acoustic features*. These features give a numerical representation of a given frame so that each frame becomes a vector in a multidimensional space. In the next step, we train a classifier that assigns the corresponding phoneme to each frame. In this sense, each

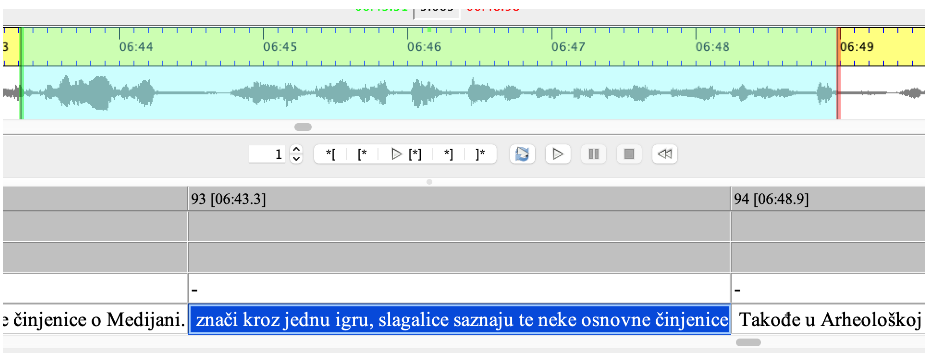


Figure 1: A segment of aligned transcription in EXMARaLDA

phoneme is one class predicted by the classifier based on the feature values of a given frame. Usually, several consecutive frames are associated with the same phoneme. This mapping, called an *acoustic model*, is learned from a large number of aligned sound-text examples as in Figure 1. The associated phonemes are then converted into characters, that is, text.

Due to the huge variability of the sound signal, the acoustic model is not sufficient to unambiguously associate the corresponding phoneme (and character). That is why a *language model* is added to the acoustic model. The task of the language model is to "correct" the output of the acoustic model by replacing the string of characters that does not match any word with the most likely word in a given context.

Techniques for training acoustic and language models are changing rapidly as the technology evolves (Jelinek, 2009). For a long time, speech-to-text conversion systems consisted of a series of programs, where each program would be responsible for one step in the process. Kaldi (Povey et al., 2011) is a very popular open-source system of this type, still used in practice, although considered outdated. Among many applications, one Kaldi recipe was created for Serbian a while ago (Popović et al., 2015).

Major advances in the development of neural network training technology since 2011 have allowed all steps to be combined into one large neural network. Although the same components are retained conceptually, they have become more abstract and flexible in a single network (often called *end-to-end*).

The performance leaps that we see today were made possible with the introduction of transfer learning in 2019. Using this technology, it is possible to pre-train models on large amounts of audio material in different languages. Also, pre-training is possible without aligned text, as in the case of the XLS-R model (Babu et al., 2021). Still, the best results are obtained by training on (at least partially) aligned data, which is the approach taken in the Whisper model (Radford et al., 2023). These large pre-trained models can be fine-tuned to maximise the performance on the target language. While models such as XLS-R have to be fine-tuned (otherwise they cannot output text), models such as Whisper can be used without fine-tuning (*zero-shot* deployment).

At the time of Kaldi, it was estimated that about 2,000 hours of transcribed audio material was needed to train a usable model. By comparison, the XLS-R is pre-trained on 500,000 hours of (untranscribed) audio, while Whisper was initially pre-trained on 680,000 hours of partially or approximately transcribed audio (the data size grows with each release of the model). Both of these models are multilingual, including Croatian and Serbian.

To measure the performance of a speech-to-text model, the output of the model is compared to a reference segment of text. The reference is considered the only correct output so any deviation from the reference is counted as an error. The standard measure is the *word error rate* (WER) and its character-level version (CER). While we are mostly interested in WER, CER is an additional measure that provides more information. In particular, a relatively good CER score can signal that the acoustic model is performing well even when WER is not good.

WER expresses the number of deviations of the model output from the reference relative to the length of the reference segment, as shown in (1).

$$WER = \frac{I + S + D}{N} \cdot 100 \tag{1}$$

The number of the deviations, more precisely called *Levenshtein edit distance* (Levenshtein, 1966), is the sum of the number of inserted (I), substituted (S) and deleted (D) words. The length of the reference segment is measured in the total number of words (N).

Consider applying this formula to the following model output (M) with respect to three possible references (R1-3).[1]

M	znači	i	kroz	jednu	igru		sa	znaju	-	neke	činjenice
R1	znači	-	kroz	jednu	igru		-	saznaju	te	neke	činjenice
E1	-	I	-	-	-		I	S	D	-	-
R2	znači	-	kroz	1	igru		-	saznaju	te	neke	činjenice
E2	-	I	-	S	-		I	S	D	-	-
R3	znači	-	kroz	1	igru	kažem	-	saznaju	te	neke	činjenice
E3	-	I	-	S	-	D	I	S	D	-	-

Counting the Levenshtein edits (E1-3), we obtain three different scores:

$WER(R1) = \frac{4}{8} \cdot 100 = 50$, $WER(R2) = \frac{5}{8} \cdot 100 = 62.5$, $WER(R3) = \frac{6}{9} \cdot 100 = 66.7$

The crucial point here is that none of the edits is necessarily an error. It is possible that the particle *i* can be heard as separate from the end of the previous word. Separating *sa znaju* is not correct according to the orthographic rules, but disregarding the space gives a correct string. Omitting the elements of spoken language *te* and *kažem* might be desirable if we want the output to be closer to written language. So, which score should we attribute to the model in this example? What should we do if the speaker said *iglu* 'needle', but the intention to say *igru* 'game' is obvious?

3 MAK NA KONAC MULTI-REFERENCE TEST DATA

To enable a robust evaluation, we have created a new multi-reference test set in Croatian and Serbian. The corpus was created in a collaboration between researchers at the Jožef Stefan Institute in Ljubljana, the URPP Language and Space at the University of Zurich and the ReLDI Centre Belgrade. The project, named *Mak na konac*, was jointly funded by the Slovenian language infrastructure CLARIN.SI, through the CLASSLA knowledge centre, and the Language and Space program of the University of Zurich. ReLDI Centre Belgrade was in charge of the annotation and data quality control tasks. The team consisted of 9 members (five annotators, a coordinator and three researchers). The creation of the data set took six months (1 November 2023 - 30 April 2024), followed by testing several models.

[1] This is a simplified version of the example in Figure 1.
Literal translation: So, (also) through one game ([I] say) [they] learn (those) some facts.

Table 1: The distribution of the Mak na konac speech samples over the three sources.

	SR1 Peščanik		SR2 Južne vesti		HR1 Ponedjeljkom u 3PM	
	f	m	f	m	f	m
Speakers	12	16	18	15	8	18
Duration	02:15:14	2:44:58	02:36:15	02:27:52	0:58:11	4:12:10
Avg. / speaker	11:16	10:19	08:41	09:51	07:16	14:01
Age range	33-77	33-89	16-45	17-48	30-52	33-65

The new data set consists of speech material taken from three sources (a total of 15h, about 5h per source):

- SR1: Radio shows produced by *Peščanik* (Belgrade),
- SR2: Television show 15 minutes produced by *Južne vesti* (Niš),
- HR1: Radio show *Ponedjeljkom u 3PM* 'On Mondays at 3PM' produced by Radio Student Zagreb (Zagreb).

Initially, the plan was to include one more Croatian source, which would represent more southern varieties of speech (Split), but until now, we have not been able to secure consent for the use of the data. For all the other sources, we received the consent of the media companies, so the data will be freely available and published through the CLARIN.SI infrastructure after the evaluation is completed. Data preparation took place in several steps (Figure 2), which we describe in the rest of this section.

When selecting the sources, we aimed at representing as diverse speakers as possible. Although it was not possible to implement a strict research design, we have managed to obtain approximately the same number of male and female speakers,[2] of varied ages (from 16 to over 70) and professions. Each speaker is represented with approximately 10 to 15 minutes of continuous speech (potentially interrupted at times).

Table 1 shows the distribution of samples with a summary of the main metadata categories. The topics of the conversations in the data sources determine the kind of speakers who participate. The highest diversity in terms of occupation and education level could be achieved in the SR2 samples, where we have

[2] One speaker is a transgender person classified as male according to the grammatical gender used by the person referring to himself.

a wide range of speakers, from high school students to university professors, including athletes, artists, and entrepreneurs. Conversely, SR1 conversations feature almost exclusively highly educated experts, such as lawyers, sociologists, historians, writers, etc. The HR1 show's format leads to a less balanced gender distribution, with most speakers being artists, mostly musicians.

Once the sources were determined, we downloaded the selected recordings from the respective web sites and converted them into .wav files (mostly from .mp3 and .mp4), which were then used for further processing. The first annotation task was to segment the audio recording into utterances similar to the example in Figure 2. For this task, we used the EXMARaLDA software (Schmidt & Wörner, 2014), which offers the option of manually aligning speech and text. More precisely, the program allows to create time stamps in the audio recording, marking the end of one and the beginning of the next segment. Initially, we create uniform segments of the length of 7 seconds. The task of the annotator at this step is to manually move the segment boundaries while listening to the audio recording. The aim of the manual adjustment was to have natural boundaries between segments and to minimise the overlap between speakers (by creating single-speaker segments) as much as possible. The annotators were instructed to place the boundary (the red vertical line in Figure 2) where they hear a natural pause. We can see in Figure 2 that such a pause was evident on the left boundary (green line), but not on the right boundary (red line). The boundary that is placed in the region of strong vocalisation shows that there was a hesitation in speech interpreted by the annotator as a segment boundary. On the other side, the boundaries could not be placed in the regions of low vocalisation when these were caused by the pronunciation of plosive consonants. These cases show that the visualisation of vocalisation in the software could be helpful for determining the boundaries between the segments, but the annotator had to listen to the recording to make sure the boundaries are well placed. Note that some recordings would have long spans of strong vocalisation without hesitations and pauses. If such spans were exceeding 20 seconds, the annotators were instructed to find the most convenient boundary and create a time stamp so that no segment is longer than 20 seconds. This was the hardest part of the task leaving some segments with an abrupt end. Overall, manual segmentation was a relatively expensive step requiring around 5 person hours for 1 hour of audio.

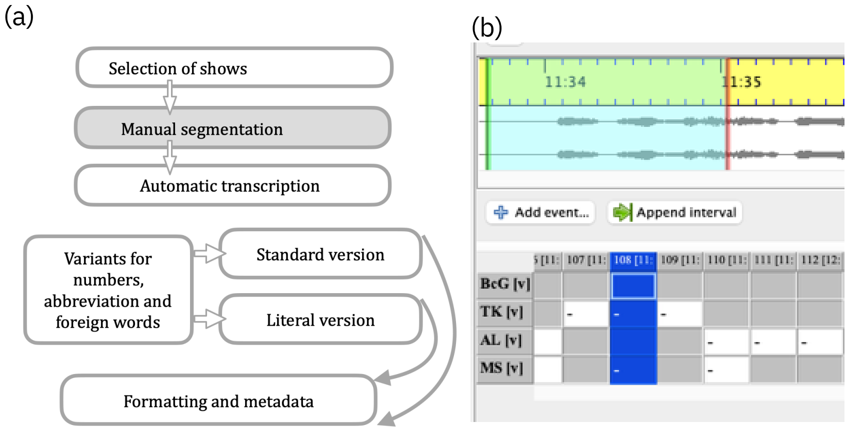


Figure 2: An overview of the major steps in the data set creation workflow (a) with an illustration of the manual segmentation step (b).

In principle, after creating a segment, the transcription is entered manually in the corresponding field. Instead, we resorted to semi-automatic insertion.

In the case of SR1, existing transcripts of the shows could be downloaded, but they were not aligned with the audio recordings at the segment level. Therefore, it was necessary to import the existing transcripts into EXMARaLDA and align them manually. During this process, it turned out that the transcripts were quite free and, often, not even close to the level of verbatim necessary to evaluate a model. These transcripts were considerably edited to follow more closely the speech segments.

In the case of SR2 and HR1, the transcripts did not exist,[3] so instead of manual transcription, we first entered an automatic one. We converted the audio recordings to text using the models available at the beginning of the project. The annotation task at this step was to insert only one character (hyphen) in the created segment instead of transcription. This sign shows who speaks in the given segment. The example in Figure 2 shows that the speakers overlap in

[3] SR2 is the same source that was used for creating the ASR training data set for Serbian JuzneVesti-SR v1.0 (Rupnik & Ljubešić, 2022). These shows were transcribed until 2018. The transcribed data were included in the previous release, while we work with the shows published after 2018 and not transcribed. To maximise the diversity of speakers, we included several shows published before 2018, but not transcribed.

the marked segment, while there is no overlap in the previous and the following segment.

To create the final samples, we selected the segments where there was no overlap so that the sum duration of all the samples of a single speaker is somewhere between 10 and 15 minutes. This step could have been performed automatically, but we opted for one more manual pass because it did not require a lot of time and it allowed us to balance the samples while selecting the segments. The SR2 and HR1 samples were then sent to automatic processing where the dashes were replaced by the output of the model.

The documents obtained in this way are further annotated in two steps. In the first step, we corrected the starting transcription to obtain a consistent standard version. Also, we added variants for numbers, abbreviations and foreign words. In the second step, we added more speech elements to the copies of the standard transcriptions. In this way, we obtained two transcriptions for each audio recording one standard and one literal, while variants of numbers, abbreviations and foreign words were entered in both transcriptions. The multiple references are thus a two-dimensional structure, where one dimension is the variation in the level of verbatim, while the other dimension is the variation in how some smaller elements of speech are written (e.g. *Kineski turisti u Srbiji troše minimalno <MD> 1000 // hiljadu </MD> <YY> eura // € // EUR </MD> dnevno*).[4]

3.1 Data format and sharing

In the final step, the data are formatted so that one table is created for each source, where each row contains one segment (about 3,000 segments per source). Each of the three main tables is accompanied by one auxiliary table that contains the speaker's metadata. These are the fields in the main file and the metadata:

[4] Translation: When in Serbia, Chinese tourists spend at least 1000 EUR per day.

Main data file:

1. Segment ID
2. Speaker ID
3. Path to the audio file
4. Standardised transcription
5. Literal transcription

Metadata file:

1 Speaker ID
2 Sample duration
3 Source ID
4 File ID
5 Name of the show
6 URL of the show
7 Name of the speaker
8 Gender
9 Approximate age
10 Occupation

The key that connects these tables is the speaker ID. With this information, we can measure the WER score on each segment and perform various analyses of model performance. We can establish whether demographic characteristics affect performance and we can also examine the impact of other properties of segments (specific vocabulary, constructions).

Since our data set is intended to be used for evaluation, we decided to share only the audio segments and keep the aligned text hidden until there is a new test set that can replace this one. In this way, we prevent model contamination[5] and create an evaluation setting that allows a realistic estimation of the performance. The shared audio segments can be downloaded from HuggingFace[6] as well as from the CLARIN.SI repository.[7] To evaluate a model, one needs to process the audio segments and upload the output to a given location. The CLASSLA team will evaluate the uploaded model output on request and return the results. This process can be automated if there is enough interest in the community. A small (10 instances) subset is available in a GitHub demo repository, [8] where one can inspect the data set encoding and run a simple evaluation of ASR against multiple references in a similar fashion as what we describe in the next section.

[5] Models are contaminated when the test set is included in the training data, which often happens with published test sets.
[6] https://huggingface.co/datasets/classla/mak_na_konac
[7] http://hdl.handle.net/11356/1833
[8] https://github.com/clarinsi/mak_na_konac

4 MODELS AND EVALUATION

For the first evaluation on the new test data, we select 6 systems that can potentially give good results on Croatian and Serbian. The systems can differ due to the architecture of the neural network built to estimate model parameters or due to the data that was used for training. Our selection represents three architectures each with two smaller variants (data or minor architecture differences).

Note that the models that we compare are not trained on the same data, which would make them not comparable in a strict sense of model comparison. As mentioned in the introduction, the transfer-learning paradigm makes the separation between the model and the data impossible, which is the main reason why a multi-reference evaluation is necessary. In addition to this, our comparison of models trained on different data still makes sense from the end-user point of view. It is intended to guide the choice between the models that are available as already (pre-)trained. We do not try to establish the advantages of any particular architecture, but ask what can publicly available models do on a new data set in Croatian and Serbian regardless of how these models are created.

We start with **Whisper Vanilla**.[9] This is the name we use to indicate that, in this setting, we apply the pre-trained multilingual Whisper model without fine-tuning. The version that we use (large-v3) is pre-trained on 1 million hours of weakly labelled data and 4 million hours of pseudo-labelled data, produced with its predecessor, Whisper-large-v2. It is capable of automatically determining the language of the input speech as well as translating input speech into a variety of languages. To see whether language-specific fine-tuning gives the expected effects, our next settings, named **Whisper Sagicc** and **Whisper Sagicc JV**, both available at (Sagić, 2023), are two variants of Whisper Vanilla fine-tuned on transcribed Serbian audio. The first variant is fine-tuned on Mozilla Common Voice 13 and Google Fleurs, while the ASR training data set for Serbian JuzneVesti-SR v1.0 (Rupnik & Ljubešić, 2022) is added to the training set for the second variant. The inclusion of the same kind of data in the training set of the second variant might lead to better scores on our SR2 subcorpus.

[9] https://huggingface.co/openai/whisper-large-v3

The next two systems are potentially interesting because they can be trained "from scratch" (without pre-training), which provides more control over the training data. These systems are the two main variants of the Conformer model (Gulati et al., 2020): **Transducer**[10] and **CTC**[11]. The main difference between the two variants is that Transducer takes previously generated letter as input at the next step, while CTC does not (it combines the acoustic and the language model in a more traditional way). In both of these settings, we test the model that was trained on Croatian parliamentary data set ParlaSpeech-HR (Ljubešić et al., 2022).

The last two systems belong to the `wav2vec` type, which means that they are pre-trained on audio data only, without text. In the **W2V2 Slavic**[12] setting, we test such a model pre-trained on Slavic audio in the VoxPopuli data set (Wang et al., 2021). In the **W2V2 XLS-R**[13] setting, pre-training is multilingual. In both cases, the models are fine-tuned on 300 hours of ParlaSpeech-HR (Ljubešić et al., 2022) with aligned audio and text.

At this time, two evaluation scenarios were studied:

1. For every instance, find the combination of variants that minimize the error metric to obtain what we call **best** results.

2. Do the opposite: for every instance, choose the variants in such a way that the reference text and the ASR transcription produce the highest error metric, denoted **worst**.

The reason for searching for the *worst* metric measurement is to stress the importance of multi-reference benchmarks, showing that even simplistic leaderboard-like orderings can be very different depending on which of the variants are taken into consideration. If more detailed feedback is ensured, the specific decisions made in single-truth benchmarks can be even more disastrous in understanding the (lack of) performance of specific systems.

Results for these scenarios were compared separately for every source (SR1, SR2, HR1), for every model, and for every metric (CER and WER). In addition

[10] https://huggingface.co/nvidia/stt_hr_conformer_transducer_large
[11] https://huggingface.co/nvidia/stt_hr_conformer_ctc_large
[12] https://huggingface.co/classla/wav2vec2-large-slavic-parlaspeech-hr
[13] https://huggingface.co/classla/wav2vec2-xls-r-parlaspeech-hr

to *best* and *worst* results, we also calculate the difference between the worst and the best score, which we report in the column **delta**. Results are reported in Table 2.

Whisper-based models reach the lowest WER and CER scores in our setup (vanilla takes the cake!). The two Whisper Sagicc models are comparable but with higher error rates on SR1 and SR2, while their performance is considerably worse on HR1 (worse than the two Conformer models as well). The inclusion of the JuzneVesti-SR data set does improve the results on the two Serbian subcorpora (SR1 and SR2), but only slightly more on the SR2 than on the SR1. The size of the models seems to be a clear contributing factor when comparing Whisper to the other models (Whisper is considerably larger than the other two types). On the other hand, the Conformer models tend to be better than the wav2vec ones, despite the latter being trained on smaller data sets. This points to the kind of the (pre-)training data as a contributing factor as smaller Conformer models trained from scratch on aligned speech-text data perform better than bigger wav2vec models pre-trained on audio-only.

Looking at the differences between the subcorpora, the performance of all the models that we tested tends to be the best on SR2, then on SR1, while the scores are considerably worse on HR1. Note that all the models except the three Whisper ones are fine-tuned and / or trained on Croatian, but they perform better on Serbian. The difficulty of the test data seems to play a more important role than the linguistic variety (HR1 seems the most difficult) but this would need to be tested in a more detailed analysis, together with other possible contributing factors such as sound quality, speaker clarity, or content complexity.

The importance of multi-reference evaluation is underlined by the fact that the rankings of the models would change in different settings. For example, the worst Whisper Vanilla score is worse than the best scores of some of the other models. The delta scores increase as the overall performance becomes better, which means that multi-reference evaluation becomes even more important when comparing highly competitive models. We note that this pattern does not hold across subcorpora. Although the scores on HR1 are generally lower than on the other two subcorpora, the delta values are higher. In this case, the delta values might be an indicator of the difficulty of the test data.

Table 2: Results for best and worst scenario. Models' names are explained in the text of the paper.

(a) Results for SR1

metric	CER			WER		
strategy	**best**	worst	delta	**best**	worst	delta
Whisper Vanilla	**5.35**	11.25	5.9	**14.62**	21.18	6.56
Whisper Sagicc	5.91	11.86	5.95	16.89	23.37	6.48
Whisper Sagicc JV	7.32	12.83	5.52	15.51	21.42	5.92
Transducer	8.15	13.78	5.63	20.08	26.18	6.11
CTC	7.74	13.41	5.67	20.88	26.96	6.08
W2V2 XLS-R	8.26	13.89	5.64	26.08	31.93	5.85
W2V2 Slavic	7.73	13.39	5.66	23.83	29.78	5.95

(b) Results for SR2

metric	CER			WER		
strategy	**best**	worst	delta	**best**	worst	delta
Whisper Vanilla	**4.76**	11.0	6.24	**11.39**	18.23	6.85
Whisper Sagicc	6.24	12.66	6.41	15.84	22.7	6.86
Whisper Sagicc JV	7.77	13.73	5.96	14.28	20.71	6.43
Transducer	8.17	14.12	5.95	19.8	26.06	6.26
CTC	7.74	13.81	6.07	20.25	26.7	6.45
W2V2 XLS-R	9.09	14.89	5.8	27.26	33.23	5.97
W2V2 Slavic	8.51	14.35	5.84	25.14	31.12	5.98

(c) Results for HR1

metric	CER			WER		
strategy	**best**	worst	delta	**best**	worst	delta
Whisper Vanilla	**6.78**	15.24	8.46	**16.18**	25.82	9.63
Whisper Sagicc	10.17	18.66	8.48	27.38	36.34	8.95
Whisper Sagicc JV	13.19	20.97	7.78	27.73	35.83	8.1
Transducer	11.29	19.13	7.85	24.97	33.06	8.1
CTC	11.06	18.97	7.91	27.02	35.09	8.08
W2V2 XLS-R	13.36	20.83	7.46	37.55	44.74	7.2
W2V2 Slavic	14.15	21.29	7.14	37.88	44.63	6.75

5 DISCUSSION

Our first evaluation outcomes show the importance of multi-reference test data for model comparison. The range of the variation between the best and the worst option shows that the rankings of the models could have been much different if a single reference was used. For instance, a single reference that results in the worst Whisper Vanilla score might result in the best Whisper Sagicc score. Without the possibility to neutralise the impact of arbitrary decisions in creating a single reference, one might arrive at a conclusion that single-language fine-tuning improves the scores, which would be wrong in this case. Allowing sufficient flexibility results in a more objective comparison and better insights into the interactions between the models.

Although our results suggest that single-language fine-tuning of large models does not give good results, these findings cannot be fully generalised given the limitations of the evaluated models. In the case of the Whisper Sagicc models, the training set for fine-tuning was extremely small (less than 100 hours). The wav2vec models were fine-tuned on a little more data in Croatian (300 hours), but this is still a small set by any standards. It is possible that more single-language data would give better results, but it remains unclear for now what data size would be beneficial.

Multi-reference benchmarks are mostly encountered in dialect data (Ali et al., 2015; Nigmatulina et al., 2020), but we show that they are necessary even if we are working with orthographic transcription in highly standardised languages. While varied transcriptions were already included in some previously published data sets (Žgank et al., 2014), our approach introduces systematic, controlled variation aimed specifically at neutralising arbitrary data biases when comparing speech-to-text models. We had to make some arbitrary decisions too, such as what elements of speech to mark (we do not mark laughter, for instance) and we could not capture all the fine nuances of possible writing, which, in reality, are infinite. Nevertheless, the possibility to choose from several references in a controlled way makes a big difference when it comes to understanding various aspects of model performance.

An important change that we introduce with this test set is the possibility to evaluate the models against desired values rather than attempting to obtain a universal measure of output quality. Instead of trying to rank all models on a single, universal scale of quality prescribed by one true solution defined by a single reference, we can determine a set of criteria that are important to us and evaluate the models according to these criteria. It may not matter to us whether the model mixes Serbian and Croatian, while it is important to us that it recognises numbers reliably and consistently. Also, we may prefer a model that always makes small mistakes over a model that processes some segments perfectly while making big mistakes in others. Up to now, we have only performed an aggregate evaluation, but many other analyses are possible in the future, including various biases and linguistic factors that might impact the model performance.

The observations that we made about the impact of various factors are currently limited because we have not performed any statistical tests and we have not covered all the categories that are needed for drawing sound generalisations. For instance, the remarks on the cross-lingual performance (e.g. Serbian models on Croatian corpus) would require making the experimental settings more comparable. We currently do not have the same models trained or fine-tuned on both Croatian and Serbian data.

Finally, some inconsistencies and mistakes in data annotation have persisted up to this point and will need to be resolved in several iterations. We believe that we will be able to spot most of these items in future fine-grained analyses and improve gradually the quality of the data set as it is used.

6 CONCLUSION

To know the performance of modern speech-to-text models, we need to evaluate them in a flexible setting using a multi-reference test set. In this paper, we have presented a new speech-to-text benchmark for Croatian and Serbian that enables such evaluation. The new data set consists of 15h of manually transcribed and aligned spontaneous speech, with 87 diverse speakers from different regions of Croatia and Serbia. Speech transcriptions are orthographic but varied according to two dimensions: the level of verbatim and whether the numbers and abbreviations are spelled out. Combining these two categories, we obtain up to 8 true transcriptions for a single segment of speech.

We have used this data set to perform an initial comparison of six competitive speech-to-text systems. This first evaluation revealed that zero-shot deployment of a large multilingual model (Whisper large v3) gives better performance than single-language training or fine-tuning when small data sets are used for fine-tuning. In future research, we plan to extend the data set to more sources and use demographic data and linguistic analyses to study how speaker and language variation impact the performance of speech-to-text models.

ACKNOWLEDGMENTS

We would like to thank our data providers for their help in gathering the audio necessary for the construction of Mak na konac data set: Radio Student Zagreb for allowing us the use of their programme 'Ponedeljkom u 3 PM', as well as the teams of Pešćanik (https://pescanik.net/), and Južne Vesti (https://www.juznevesti.com/). This work was partially funded by the programme P6-0411 "Language Resources and Technologies for Slovene", the CLARIN.SI infrastructure, and the project J7-4642 "MEZZANINE - Development of Spoken Language Resources and Speech Technologies for the Slovenian Language", all financed by the Slovenian Research and Innovation Agency (ARIS).

REFERENCES

Ali, A., Magdy, W., Bell, P., & Renais, S. (2015). Multi-reference wer for evaluating asr for languages with no orthographic rules. In *2015 ieee workshop on automatic speech recognition and understanding (asru)* (p. 576-580). doi: 10.1109/ASRU.2015.7404847

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., ... Auli, M. (2021). XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, *abs/2111.09296*. https://arxiv.org/abs/2111.09296

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., ... Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. interspeech 2020* (pp. 5036–5040). doi: 10.21437/Interspeech.2020-3015

Jelinek, F. (2009, December). ACL lifetime achievement award: The dawn of statistical ASR and MT. *Computational Linguistics*, *35*(4), 483–494. https://aclanthology.org/J09-4004 doi: 10.1162/coli.2009.35.4.35401

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*, 707.

Ljubešić, N., Koržinek, D., Rupnik, P., Jazbec, I.-P., Batanović, V., Bajčetić, L., & Evkoski,

B. (2022). *ASR training dataset for croatian ParlaSpeech-HR v1.0.* http://hdl.handle.net/11356/1494 (Slovenian language resource repository CLARIN.SI)

Nigmatulina, I., Kew, T., & Samardzic, T. (2020, December). ASR for non-standardised languages with dialectal variation: the case of Swiss German. In M. Zampieri, P. Nakov, N. Ljubešić, J. Tiedemann, & Y. Scherrer (Eds.), *Proceedings of the 7th workshop on nlp for similar languages, varieties and dialects* (pp. 15–24). Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL). https://aclanthology.org/2020.vardial-1.2

Popović, B., Ostrogonac, S., Pakoci, E., Jakovljević, N., & Delić, V. (2015). Deep neural network based continuous speech recognition for serbian using the kaldi toolkit. In A. Ronzhin, R. Potapova, & N. Fakotakis (Eds.), *Speech and computer* (pp. 186–192). Cham: Springer International Publishing.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011, December). The kaldi speech recognition toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding.* IEEE Signal Processing Society. (IEEE Catalog No.: CFP11SRW-USB)

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).

Rupnik, P., & Ljubešić, N. (2022). *ASR training dataset for serbian JuzneVesti-SR v1.0.* http://hdl.handle.net/11356/1679 (Slovenian language resource repository CLARIN.SI)

Sagić, A. (2023). *Whisper-large-v3-sr-combined.* Retrieved 2024-05-28, from https://huggingface.co/Sagicc/whisper-large-v3-sr-combined

Schmidt, T., & Wörner, K. (2014, 05). 402EXMARaLDA. In *The Oxford Handbook of Corpus Phonology.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571932.013.030 doi: 10.1093/oxfordhb/9780199571932.013.030

Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., ... Dupoux, E. (2021). Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *CoRR*, *abs/2101.00390*. https://arxiv.org/abs/2101.00390

Žgank, A., Vitez, A. Z., & Verdonik, D. (2014, May). The Slovene BNSI broadcast news database and reference speech corpus GOS: Towards the uniform guidelines for future work. In N. Calzolari et al. (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 2644–2647). Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/710_Paper.pdf

MAK NA KONAC: VEČREFERENČNI PRIMERJALNI PREIZKUS ZA RAZPOZNAVALNIKE GOVORA ZA HRVAŠČINO IN SRBŠČINO

Evalvacija razpoznavalnikov govora je postala še posebej potrebna v okviru nedavnih tehnoloških skokov, ki so povzročili široko uporabo teh modelov in močno konkurenco na trgu. V tem članku je predstavljena nova podatkovna množica, namensko zasnovana za reševanje zahtevnega problema objektivne primerjave modelov. Namesto togega primerjanja razpoznanega govora in enega pravilnega prepisa predlagamo prožno vrednotenje na več enakovrednih možnih prepisih. Novo podatkovno množico sestavljajo ročno urejene transkripcije vzorcev spontanega govora iz treh virov (enega hrvaškega in dveh srbskih), v skupni dolžini približno 15 ur. Naša začetna primerjava šestih primerljivih sistemov za razpoznavo govora kaže stabilne vzorce v vseh treh virih: t.i. 'zero-shot' uporaba velikega večjezičnega modela daje boljše rezultate kot modeli, ki so bili predhodno učeni ali doučeni v posameznih jezikih.

Keywords: avtomatska razpoznava govora, večreferenčna evalvacija, primerjalni preizkus, hrvaščina, srbščina

STROJNA PREVEDBA INTERNETNIH NASLOVOV NOVIČARSKIH PRISPEVKOV V NASLOV NA WAYBACK ARCHIVE

Janez ŠTEBE
Univerza v Ljubljani, Arhiv družboslovnih podatkov

V članku predstavimo dozdajšnje aktivnosti sodelovanja podatkovnih repozitorijev in znanstvenih revij v Sloveniji pri uvajanju citiranja in dostopa do raziskovalnih podatkov, ki spremljajo objave v znanstveni literaturi. Kot nadaljevanje aktivnosti je Arhiv družboslovnih podatkov (ADP) vodil pilot sodelovanja z nekaterimi družboslovnimi revijami. Prikazan je primer sodelovanja z revijo Javnost, ki je prestavljal poseben izziv za arhiviranje, saj so bili podatki novičarski prispevki, obravnavani v enem izmed člankov v reviji. Problem za redistribucijo so avtorske in sorodne pravice, obenem pa ni zagotovila, da bodo novičarski portali ohranjali objavljene vsebine. Preprosta koda v programskem jeziku Python je avtomatizirala prenos internetnih naslovov na *Wayback Machine*, v področnem podatkovnem arhivu ADP pa je bil shranjen vzorčni seznam vseh v raziskavi obravnavanih internetnih naslovov, katerih povezani novičarski prispevki predstavljajo raziskovalne podatke danega članka. V razpravi na obravnavanem primeru ponovno izpostavimo prednosti shranjevanja in dostopa do podatkov, ki spremljajo znanstveno literaturo.

Ključne besede: reproduktibilnost, raziskovalni podatki, znanstvena računalniška koda, znanstveno založništvo

1 UVOD

Predstavljamo primer računalniške kode, ki smo jo razvili *ad hoc* za trajno shranjevanje in dostop ene izmed prevzetih enot raziskovalnih podatkov v Arhivu družboslovnih podatkov (ADP). Gre za podatke avtorjev Kuhar in sod. (2021), ki so bili podlaga za članek avtorjev Smrdelj in sod. (2021). Podatki vsebujejo seznam referenc na originalne medijske objave. V nadaljevanju opišemo, v čem je bil problem pri arhivski obravnavi omenjenih podatkov. Za razumevanje širše slike podamo izhodišča in možne načine reševanja problema, kot se pojavljajo v literaturi, utemeljimo izbrano rešitev in predstavimo podrobnosti izsekov kode ter načina njene uporabe. Rešitev bi bila lahko z malo prilagoditve uporabljena tudi v drugih podobnih primerih.

V razpravi potegnemo vzporednice s podanim problemom ohranjanja referenc na novičarske članke in pomenom shranjevanja računalniške kode za namen ponovne uporabe ali pa za zagotovitev reproduktibilnosti v znanstveni literaturi objavljenih rezultatov analiz.

1.1 Izhodišča

Znanstvene revije in znanstvena skupnost se zavedajo pomena dostopnosti podatkov in tudi računalniške kode v povezavi z znanstveno literaturo. V ospredju so možnosti ponovne rabe podatkov za druge namene kot prvotno, lahko samostojno ali pa v povezavi z drugimi podatki. Pomen dvojice podatkov in računalniške kode, namenjene analizi in izdelavi rezultatov, je tudi v zagotavljanju računske replikabilnosti (ang. computational replicability) (Peer, 2024). Dostopni paket podatkov, računalniške kode in drugih elementov se ugotavljanju računske replikabilnosti podvrže preizkusu, če deluje in daje enake rezultate kot so ti, objavljeni v članku v reviji.

Na različnih področjih znanosti, pa tudi znotraj družboslovja, je zavzetost za reproduktibilnost ponekod večja kot drugod. Ekonomisti so zavzetejši, saj je pogostejša uporaba kvantitativnih podatkov in kompleksnih računalniško podprtih analiz (Hoynes, 2023). Sociologija je glede metodoloških pristopov bolj multiparadigmatska. Vprašanje zagotavljanja reproduktibilnosti še ni v večji meri vstopilo v navodila ter priporočila revij in združenj. Zlasti pri kvalitativnem raziskovanju, ki predstavlja velik del socioloških pristopov, je pomen doseganja transparentnosti raziskovanja drugačen, temu primerno pa so drugačni tudi postopki za njeno preverjanje (Weeden, 2023).

V Sloveniji je s pobudo za ozaveščanje znanstvene skupnosti in revij o pomenu deljenja raziskovalnih podatkov v povezavi s članki začela delovna skupina znotraj RDA-vozlišča Slovenije. V letih 2019 in 2020, ko je bilo delovanje vozlišča financirano prek evropskega projekta, je potekal tudi pilot dela z znanstvenimi revijami. Skupina je na podlagi usmeritev (Hrynaszkiewicz in sod., 2020) *Interesne skupine za standardizacijo in uvajanje podatkovnih* *politik* (RDA Alliance, 2024) pripravila nacionalnemu okolju prilagojene Smernice (Štebe in sod., 2020a) in z njimi povezana Pojasnila (Štebe in sod., 2020a). Oba dokumenta sta bila namenjena podpori slovenskim znanstvenim revijam pri uvajanju uredniških pravil glede obveznosti in načina citiranja ter dostopnosti raziskovalnih podatkov. Pri uvajanju Smernic so v okviru pilotnega preizkusa sodelovale posamezne revije, pretežno s področij družboslovja in humanistike. Pokazalo se je, da so glavne ovire pomanjkanje strokovne in tehnične podpore na strani revij ter negativna pričakovanja glede prisotnosti kulture deljenja podatkov področnih znanstvenih skupnosti, ki predstavljajo bazen potencialnih avtorjev teh revij (Štebe in sod., 2020c).

Delo z revijami se je nadaljevalo tudi v naslednjem obdobju. ADP je bil vključen v delovno skupino za sodelovanje z revijami pri evropski družboslovni infrastrukturi Consortium of European Social Science Data Archives (CESSDA). Eno izmed poročil delovne skupine je na podlagi pregleda literature in analize intervjujev z izbranimi predstavniki revij po sodelujočih državah ugotavljalo, da imajo globalni založniki večinoma elaborirane osrednje podatkovne politike, da pa se pozna odsotnost naslavljanja področnih specifik družboslovja, kot je vprašanje, kako ravnati z občutljivimi osebnimi podatki in različnimi novimi vrstami podatkov, ki jih uporabljajo družboslovci. Intervjuji s predstavniki nacionalnih uredništev manjših revij pa so pokazali, da se zavedajo pomena in tovrstnih izzivov deljenja podatkov, da pa se sami ne čutijo dovolj usposobljeni za svetovanje avtorjem glede pristopov priprave podatkov za dostop. Tovrstno svetovanje in podporo avtorjem bi bili pripravljeni prepusti nacionalnim organizacijam izvajalcev podatkovnih storitev znotraj združenja CESSDA (Alvanides in sod., 2020). Glavnina aktivnosti v koordinaciji delovne skupine je bila namenjena nadaljevanju preizkušanja modelov sodelovanja podatkovnih arhivov z revijami v nacionalnih okoljih (Štebe in sod., 2022).

V sklopu teh aktivnosti se je ADP povezal z uredništvi nekaterih družboslovnih revij, med njimi z revijo Javnost/The Public. Vabilo k razmisleku o spodbujanju deljenja podatkov v povezavi s članki so pri tej reviji vzeli resno. Boris Mance, pomočnik glavnega urednika, je predlagal, da se kar takoj lotimo dela. Preden bi pripravili svoja navodila glede dostopa do podatkov na ravni revije, so želeli sami model preizkusiti v omejenem obsegu. Ravno so pripravljali posebno številko revije, ki jo je urejal kot gostujoči urednik (Suplement, Javnost, 2021).

Avtorje v posebni številki so pozvali, da naj svoje članke obogatijo s povezavo na uporabljene podatke, in sicer tako, da podatke objavijo v področnem repozitoriju in jih ustrezno citirajo v seznamu literature. ADP je avtorjem ponudil podporo pri pripravi podatkov za objavo v katalogu in zagotovil prednostno obravnavo predanih podatkov. Avtorji so k predaji podatkov pristopili zavzeto, kar kaže, da so strahovi pred nesodelovanjem avtorjev člankov ob ustreznem pristopu odveč (Bezjak, 2022). Pripomoglo je gotovo tudi dejstvo, da so bili članki v sklepni redakciji in da so že prestali recenzijo, tako da sama objava ni bila vprašljiva ne glede na sodelovanje pri pilotu.

Revija Javnost nastaja pod okriljem mednarodne založbe Routledge, ki je del konglomerata Taylor & Francis (T & F). Založnik T & F sicer ima krovna navodila glede deljenja podatkov v povezavi s članki (T & F, 2024a). Opredeljujejo, kaj so raziskovalni podatki ter kateri so razlogi in koristi za deljenje podatkov. V nadaljevanju pa podajajo zgledno izčrpna in razdelana navodila, ki podrobno opredeljujejo potrebne korake. Opišejo pomen priprave Načrta ravnanja z raziskovalnimi podatki in pridobivanja soglasij za deljenje podatkov, kje podatke objaviti, in podajo pojasnila glede načel FAIR. Poudarijo pomembnost deljenja raziskovalne kode. Predstavijo, zakaj je pomembno povezovanje literature in podatkov z vključevanjem permanentnih identifikatorjev (PID) v metapodatke ene in druge entitete. Podani so napotki o tem, da je treba podatke v članku citirati, in sicer svoje pa tudi če uporabljaš podatke od drugod, ter da mora struktura navajanja podatkov vsebovati PID in mesto, na katerem so podatki in spremljajoči metapodatki dostopni. Navodila so skladna z vse bolj uveljavljenimi priporočili za različna znanstvena področja (za jezikovne podatke primerjajte Lenardič in sod., 2020).

1.2 V čem je problem?

Čeprav revija Javnost/The Public spada pod T & F, ima – kot smo že ugotavljali drugje – svoja posebna Navodila za avtorje (Javnost, 2024), ki sama ne omenjajo pomembnosti ali celo obveznosti citiranja in deljenja podatkov. Edino, kar je na strani Navodil, je povezava na krovno založniško politiko (T & F, 2024b), ki sicer vsebuje razdelek o dostopnosti podatkov in njihovi predaji, vendar je politika nezavezujoča za uredniško avtonomne revije, kakršna je tudi naša.

Ko smo Mancetu in kolegom v uredništvu posebne številke predstavili pobudo (tudi glede na njihovo pripravljenost za sodelovanje), smo se zavedali, da lahko glede na posebnosti primerov člankov, ki so jih pripravljali za objavo, pričakujemo določene izzive. Iz interne korespondence uredniške skupine, ki smo jo lahko spremljali, je razvidno tehtanje vprašanja, kako s podatki medijskih objav:

> Bolj pa imava probleme glede člankov, pri katerih gre za vsebinske analize medijskih objav (kar smo se pogovarjali zadnjič pred faksom). Kaj storiti v takem primeru (X. opozarja, da najbrž nimamo pravice arhivirati medijskih objav in da tu ne gre za podatke, ki jih mi generiramo (kot pri intervjujih))? Obenem gre tudi za javno dostopne podatke, torej v takem primeru arhiviranje ni tako nujno? Ali pač? (Štebe, 2021a; osebna imena zakrita).

Izpostavljeni sta dve dilemi. Prva je povezana z vprašanjem pravic (avtorskih in sorodnih) glede arhiviranja in ponovnega objavljanja podatkov člankov, ki so sicer že bili objavljeni drugje, vendar pod različnimi licencami, lahko tudi z omejitvijo redistribucije ali pa brez kakršnih koli oznak glede razpolaganja. Če bi hoteli to razčistiti, bi morali, tako kot pri objavah na socialnih medijih, preveriti, kakšne so licence, in druge morebitne omejitve pravice uporabe (ToS, Terms of Service, Thompson, 2016; Hagen in sod., 2019; Breuer in sod., 2021). Ob omejitvi razpolaganja bi morali, če bi hoteli zagotoviti pravno nedvoumno situacijo in brez tveganja za kršenje pravic, z lastniki medijev doseči dogovor, ki bi dovolil tovrstno uporabo. Ali pa tvegati in računati na to, da zaradi premajhnega interesa nihče izmed lastnikov vsebin ne bi oporekal arhiviranju in dajanju podatkov v dostop za raziskovalne namene, v smislu uporabe v javnem interesu (SSRC, 2019). Lahko pa bi vsebine arhivirali in določili omejen dostop pod posebnimi pogoji, s katerimi bi moral uporabnik soglašati.

Druga dilema v navedku zgoraj je, ali vsebine glede na to, da so že bile javno objavljene, sploh predstavljajo podatke, ki bi jih bilo treba arhivirati in ponovno objaviti pod titulo raziskovalnih podatkov, saj so vsebine že tako ali tako dostopne vsakomur.

1.3 Predlagana rešitev problema

V nadaljevanju bomo prikazali in utemeljili rešitev, ki se je zdela najprimernejša glede na naravo podatkov in z vidika učinkovite obravnave v arhivu podatkov ADP.[1]

Predlagali smo, da se arhivira samo navedke medijskih objav skupaj s povezavami na originalne podatke. Avtorji so namreč že razpolagali z naborom vseh navedkov, ki so ustrezale iskalnim pogojem v določenem obdobju. Dodatno je bilo potrebno upoštevati trajne formate datotek in pripraviti metapodatke.

Argument, da so vsebine medijskih objav že objavljene in se ni potrebno ukvarjati z njihovim dostopom, ima tudi svoje slabost. Pristop z arhiviranjem povezav na objavljene vsebine se zgleduje po ponujenih rešitvah za arhiviranje objav na *Twitterju* (zdaj *X*). Shrani se množica enkratnih identifikacijskih števil vsake izmed objav, ki nato prek API-ja omogoča rekonstrukcijo vsebine in z njo povezanih podatkov (Kaczmirek in sod., 2014; Kinder - Kurlanda in sod., 2017). Tako kot pri vsebinah socialnih medijev, ki se jih *rehidrira* na podlagi identifikatorja, je tudi pri povezavah na originalne objave težava, da lahko vsebina ni več dostopna pod določenim naslovom. Objavo na socialnem mediju je lahko uporabnik izbrisal in ni več dostopna. Portal medijskih objav lahko zamenja naslov, stare objave lahko briše, sam medij lahko propade in vsebine niso več dostopne. Da bi rešili še to težavo, saj je cilj nekoliko trajnejši dostop, smo vse vsebine, ki so bile v aktualnem času dostopne, označili za avtomatizirano preselitev na spletni arhiv *Wayback Machine* (2024).

Wayback Machine (WM) je projekt neprofitne digitalne knjižnice *The Internet Archive*. Med prednostmi storitve je, da ima poslanstvo trajnega shranjevanja in da deluje v javnem interesu, kar je tudi pravna podlaga upravičevanja morebitnih izjem glede GDPR in vprašanj pravic intelektualne zaščite. Prav tako pa ponuja različne možnosti strojnega dostopa do zbirk, ki poleg internetnih strani vključujejo tudi zbirke TV-vsebin, e-knjig, programske kode idr.

[1] Glede na jezikovno naravo podatkov bi prišlo v poštev tudi, da bi dajalce podatkov usmerili na repozitorij CLARIN.SI za jezikovne vire in orodja, ki deluje znotraj evropske raziskovalne infrastrukture za jezikovne vire in tehnologije CLARIN ERIC, vendar bi s tem samo preložili ukvarjanje s problemom na nekoga drugega, ob tem, da smo avtorjem pri reviji Javnost zagotovili podporo pri pripravi podatkov.

Rezultat izbora medijskih objav predstavlja enkraten izsek določene tematike v določenem časovnem obdobju, tudi zemljepisno zamejeno. Edinstvenost množice objav in njihova raba v kontekstu raziskave pa jim pridajata status raziskovalnih podatkov. Podatki so lahko namenjeni preverjanju in reproduktibilnosti objavljenih ugotovitev ali pa se jih uporabi v nadaljnjih raziskavah z drugačnimi izhodišči in analitičnim pristopom kot izvirno, lahko pa se jih uporabi za primerjavo z drugimi podatki.

Preiskali smo tudi možnost uporabe slovenskega Spletnega arhiva NUK. Kolega Janko Klasinc, spletni arhivar na NUK, s katerim smo se pogovarjali, je kot možnost ponudil prenos vsebin na pobudo, kar bi pomenilo predajo urejenega seznama internetnih naslovov sodelavcem NUK za prenos na njihov Spletni arhiv z njihovim dodatnim naporom indeksacije seznama. Prednost NUK je, da deluje na zakonski podlagi, saj leta 2009 dopolnjeni Zakon o obveznem izvodu publikacij dovoljuje zajem in dostop do arhiviranih javno objavljenih spletnih vsebin.[2] Omejitev je, da je zajem selektiven, saj temelji na presoji pomena za slovensko kulturno dediščino. Glavna omejitev je, da ne omogoča avtomatizacije zajema na zahtevo prek API. Kolega Klasinc je prijazno za preizkus dejansko poindeksiral nabor enega dela referenc na spletne strani za NUK-ov Spletni arhiv. V naši medsebojni korespondenci je zapisal:

> Najprej sem nameraval predlagati, da počakamo na zaključek naše indeksacije in potem primerjamo naše zajeme z njihovimi obstoječimi [mišljeni zajemi na WM, op. a.]. Ampak če pri njih omogočajo brezplačen in enostaven zajem posameznih strani, ki še niso zajete, bi bila to morda boljša rešitev za vas (Štebe, 2021b).

2 OPIS PODATKOV IN METODE SKUPAJ Z RAČUNALNIŠKO KODO

Osnova za prikaz priprave podatkov za podatkovni arhiv so podatki, zbrani za članek o objavah o istospolnih osebah in tujcih med obravnavo zakonov, ki zadevajo obe marginalizirani skupini, v letih 2015 (Zakon o zakonski zvezi in družinskih razmerjih) in 2017 (Zakon o tujcih) (Smrdelj in sod., 2021). Po določenih kriterijih in s ključnimi besedami so generirali vzorec objav na štirih novičarskih spletnih portalih: *RTVSLO.si*, *24ur.com*, *SiOL.net* in *Nova24TV.si*.

[2] https://web.archive.org/web/20230531195146/https://arhiv.nuk.uni-lj.si/zakonska-podlaga

Vzorec je skupaj obsegal 393 enkratnih enot.

Količina enot je bila prevelika za ročno iskanje, če so objave že prenešene v spletni arhiv WM. Če še niso, bi bilo treba ročno vnesti predlog na WM za prenos objave. Odločili smo se, da pripravimo preprosto kodo v programu Python, ki bo postopek avtomatizirala.

V arhivski obravnavi smo dve prejeti besedilni datoteki z navedki zadetkov iz Worda najprej pretvorili v enostavno tekstovno datoteko s končnico *.txt ('interd17_p1_sl_v1_r1.txt' in 'interd17_p2_sl_v1_r1.txt' (Štebe, 2024)), ki se lahko bere prek računalniške kode (Tabela 1).

Tabela 1: Izsek treh enot iz dokumenta vzorca navedkov v enostavnem tekstovnem formatu (datoteka 'interd17_p1_sl_v1_r1.txt' (Štebe, 2024)).

```
1) MMC RTV SLO

1) Je napočil čas, ko bodo istospolni pari poročeni in ne
registrirani? (19. 1. 2015) https://www.rtvslo.si/slovenija/je-
napocil-cas-ko-bodo-istospolni-pari-poroceni-in-ne-
registrirani/356207

2) "Če hočeš ekonomski razvoj, je enakost ne glede na spolno
usmerjenost prava pot" (9. 2. 2015)
https://www.rtvslo.si/slovenija/ce-hoces-ekonomski-razvoj-je-
enakost-ne-glede-na-spolno-usmerjenost-prava-pot/357921

3) Janša: Nismo brez moči, zato tudi ne brez odgovornosti (14. 2.
2015) https://www.rtvslo.si/slovenija/jansa-nismo-brez-moci-zato-
tudi-ne-brez-odgovornosti/358382
```

Programska koda v Python (datoteka ‘read_ln_txt.py’ v Štebe, 2024) bere zgornjo datoteko vrstico po vrstico in ignorira vrstice, ki ne vsebujejo internetnega naslova. Te vrstice samo po vrsti prepiše v izhodno datoteko.

V vrstici, ki vsebuje internetni naslov, najprej poišče prosti zapis datuma objave in ga pretvori v računalniški datumski format. Koda izlušči internetni naslov originalne objave. Nato z internetnim naslovom naredi poizvedbo na API WM. Pri tem uporabi funkcije iz paketa v Pythonu 'waybackpy' (Mahanty, 2022). Zapisi na *The Internet Archive* API-ju se vrnejo v lokalno opredeljenem formatu CDX. Na voljo je vrsta drugih možnosti za dostopanje do vsebin enot in metapodatkov prek API-jev, vključno s pretvorbo v format JSON (IA, 2024), ki pa nas tokrat niso zanimale, saj smo želeli čim bolj preprosto in hitro rešitev za naš konkretni namen. Če je naslov že prenešen v spletni arhiv, ga program samo zabeleži v izhodnem dokumentu v obliki za citiranje, kot je najbližja datumu objave na WM. Če še ni prenešen, pa da koda ukaz za prenos naslova na WM in po uspešnem prenosu vsebine strani prav tako poleg izvirne zabeleži novo obliko naslova. Pred vsako novo poizvedbo program določi odmor, saj je API na WM občutljiv, če so poizvedbe iz istega naslova prepogoste, in se lahko preneha odzivati.

3 REZULTAT OBRAVNAVE

Vseh 393 izvirnih internetnih naslovov je bilo med preizkusom še aktivnih. Tisti, ki še niso bili do takrat, so bili na novo uspešno prenešeni v WM. Večina naslovov je bila prenešena že predhodno, med manjkajočimi pa so bili večinoma iz *Siol.net*. Manjkajoči naslovi bi bili potencialno izpostavljeni izgubi, če ne bi bilo aktivno spodbujenega prenosa. Obdelava ene in druge datoteke z naslovi je bila popolnoma avtomatizirana, zahvaljujoč tudi doslednosti avtorjev pri navajanju URL-naslovov.

Z naslovi na WB obogateni podatki navedkov skupaj z metapodatki, ki vsebujejo opis raziskave in metode po strukturi Data Documentation Initative (DDI), so bili po končani arhivski obravnavi dostopni na ADP (Kuhar in sod., 2021). Sami podatki v navedkih naslovov poleg izvirnega naslova vsebujejo še naslov na WM (Tabela 2).

Tabela 2: Izsek končne izhodne datoteke vzorca navedkov z dodanim naslovom na spletnem arhivu *Wayback Machine* (krepko označeno) (Kuhar in sod., 2021).

1) MMC RTV SLO

1) Je napočil čas, ko bodo istospolni pari poročeni in ne registrirani? (19. 1. 2015) https://www.rtvslo.si/slovenija/je-napocil-cas-ko-bodo-istospolni-pari-poroceni-in-ne-registrirani/356207 **<https://web.archive.org/web/20150122200530/http://www.rtvslo.si:80/slovenija/je-napocil-cas-ko-bodo-istospolni-pari-poroceni-in-ne-registrirani/356207>**
2) "Če hočeš ekonomski razvoj, je enakost ne glede na spolno usmerjenost prava pot" (9. 2. 2015) https://www.rtvslo.si/slovenija/ce-hoces-ekonomski-razvoj-je-enakost-ne-glede-na-spolno-usmerjenost-prava-pot/357921 **<https://web.archive.org/web/20150710101105/http://www.rtvslo.si/slovenija/ce-hoces-ekonomski-razvoj-je-enakost-ne-glede-na-spolno-usmerjenost-prava-pot/357921>**
3) Janša: Nismo brez moči, zato tudi ne brez odgovornosti (14. 2. 2015) https://www.rtvslo.si/slovenija/jansa-nismo-brez-moci-zato-tudi-ne-brez-odgovornosti/358382

<https://web.archive.org/web/20150217101345/http://www.rtvslo.si:80/slovenija/jansa-nismo-brez-moci-zato-tudi-ne-brez-odgovornosti/358382>

Sama storitev spletnega arhiva WB ima z vidika trajnega shranjevanja tudi svoje pomanjkljivosti. Besedilo se v spletnem arhivu prenese in prikaže neokrnjeno. Videovsebine pa se ne prenesejo, kot kaže podrobna analiza enega izmed primerov[3] zajema spletnih strani medijske objave na *24ur.com*. V tem primeru se moramo vprašati, kaj so tiste 'pomembne lastnosti' (*significant properties*), ki si jih prizadevamo trajno ohraniti. V našem primeru sta ohranjena samo besedilni povzetek posnetka oddaje in spremljajoča slika.

4 RAZPRAVA IN ZAKLJUČEK

Med prednostmi arhiviranja podatkov medijskih objav je v povezavi s člankom v reviji tudi možnost zabeležke veliko več referenc, kot pa jih je uporabljenih v članku. Ta namreč vsebuje samo reference na neposredne navedbe dela novičarskega besedila v samem članku, ki so namenjene ilustraciji ugotovitev, dobljenih z metodo kritičnih študijev diskurza (KŠD) (Smrdelj in sod., 2021). Vsakdo na s člankom povezanih arhiviranih podatkih preveri ustreznost navedkov v smislu replikacije. Razširjen nabor referenc, ki je zajet v arhiviranih podatkih, omogoča nadaljnje analize pa tudi oceno nepristranosti originalne analize.

Podatki so kot vzporedna objava o članku tudi nagrada sodelujočim avtorjem, ki so brez dvoma v pripravo podatkov za predajo v arhiv morali vložiti dodane napore. Podatke je po objavi v arhivu mogoče navajati v literaturi, kar lahko pomeni, da so avtorji lahko deležni novih citatov. Po načelih citiranja podatkov so tudi avtorji zadevnega članka (Smrdelj in sod., 2021) v literaturi navedli lastne uporabljene podatke (Slika 1).

[3] Primerjajte https://www.24ur.com/novice/slovenija/med-begunci-tudi-lastniki-tovarn-organizatorji-porok-knjizni-zalozniki-stand-up-komiki.html in https://web.archive.org/web/20170104043031/https://www.24ur.com/novice/slovenija/med-begunci-tudi-lastniki-tovarn-organizatorji-porok-knjizni-zalozniki-stand-up-komiki.html.

Slika 2: Navajanje v članku uporabljenih podatkov v seznamu literature.

Ljubljana: Mirovni inštitut.

Kuhar, Roman. 2015. "Konec je sveta, kakršnega poznamo: populistične strategije nasprotnikov Družinskega zakonika." *Časopis za kritiko znanosti* 43 (260): 118–132.

Kuhar, Roman. 2017. "Changing Gender Several Times a Day: The Anti-Gender Movement in Slovenia.« V *Anti-Gender Campaigns in Europe: Mobilizing Against Equality*, uredila R. Kuhar in D. Paternotte, 215–232. London in New York: Rowman & Littlefield International.

Kuhar, R., Pajnik, M., Učakar, T., Ješe Perković, A. Zalta, A. Mandelc, D., Banjac M., Antić Gaber, M., Smrdelj R., Šori, I. in Fabijan, E. (2021) Državljanstvo in diskriminacija: Intersekcijski pristop k raziskovanju družbene izključenosti, 2017 [Podatkovna datoteka]. Ljubljana: Univerza v Ljubljani, Arhiv družboslovnih podatkov. ADP - IDNo:INTERD17. https://doi.org/10.17898/ADP_INTERD17_V1

Luthar, Breda. 2017. "Begunci in 'Odmevi': epistemologija konvencij." *Dve domovini* (45): 153–168.

McCombs, Maxwell. 2014. *Setting the Agenda. The Mass Media and Public Opinion*. Cambridge: Polity Press.

MOSS. 2021. "Rezultati MOSS. Valutni podatki o obiskanosti spletnih mest." https://www.moss-

Ta članek je hkrati tudi prikaz, kako lahko tudi programsko kodo arhiviramo in navajamo v literaturi. Skripta v Python je dostopna na GitHub in arhivirana na Zenodo. Alternativna lokacija za arhiviranje kode je po analogiji z internetnim arhivom Software Heritage Archive (2024). Preprosta skripta v Pythonu se lahko dopolni in ponovno uporabi na novih gradivih drugje.

Druga analogija z našim pristopom je pobuda za urejanje kompleksnih navedb, ki so jo predstavili na zadnji konferenci RDA (Stall, 2024). Na veliko znanstvenih področjih se v članku uporablja veliko količino referenc na druge podatke. Pobuda na podoben način, kot smo pokazali pri tem članku, zbere vse reference skupaj s stalnimi identifikatorji v enem samem dokumentu, ki tako postane nov predmet shranjevanja in citiranja.

Preizkus sodelovanja avtorjev posebne številke revije Javnost z arhivom podatkov je dal pozitivne rezultate obogatenih vsebin člankov. Še pomembnejša pa je popotnica pri nadaljnji širitvi sodelovanja arhivov in revij pri vzpostavitvi podatkovnih politik ter njihovi izvedbi. Področni podatkovni arhiv, kot je ADP, je pripravljen svetovati avtorjem in ob posebnih izzivih, kot je bil ta, poiskati priročne rešitve, ki so skladne z zakoni ter učinkovito zagotovijo dostop do podatkov in virov za namene replikacije ali nove analize.

LITERATURA

Alvanides, S., Štebe, J., Morgan de Paula, E., Kleiner, B., Hegedus, P., Glavica, M., & Kranjec, I. (2021). Assessment of Journal Requirements and Needs. CESSDA Work Plan 2020 (3.0). Zenodo. https://doi.org/10.5281/zenodo.5554436

Breuer, J., Borschewski, K., Bishop, L., Vávra, M., Štebe, J., Strapcova, K., & Hegedus, P. (2021). Archiving Social Media Data: A guide for archivists and researchers

(1.2). Zenodo. https://doi.org/10.5281/zenodo.6517880

Hagen, S., Bishop, E., Koščík, M., Vavra, M., Štebe, J., Ryan, L., Payne, E., Løvlie, A., Straume, Ø., Rød, L. M., Høgetveit Myhren, M. (2019). Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of social media data. Deliverable 6.3 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Pridobljeno 31. maja 2024. https://seriss.eu/wp-content/uploads/2019/11/D6.3-Report-on-legal-and-ethical-framework-and-strategies...__FINAL.pdf

Hoynes, H. (2023). Reproducibility in Economics: Status and Update. Harvard Data Science Review, 5(3). https://doi.org/10.1162/99608f92.80a1b88b

Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R. and Goudie, S. (2020). Developing a Research Data Policy Framework for All Journals and Publishers. Data Science Journal, 19(1). http://doi.org/10.5334/dsj-2020-005

IA, The Internet Archive (2024). Internet Archive Developer Portal. Pridobljeno 31. julija 2024. https://archive.org/developers/

Javnost – The Public (2024). Instructions for authors. Updated 28th February 2024. https://www.tandfonline.com/action/authorSubmission?show=instructions&journalCode=rjav20

Kaczmirek, L., Mayr, P., Vatrapu, R., Bleier, A., Blumenberg, M., Gummer, T., Hussain, A., Kinder-Kurlanda, K., Manshaei, K., Thamm, M., Weller, K., Wenz, A., & Wolf, C. (2014). Social media monitoring of the campaigns for the 2013 German Bundestag Elections on Facebook and Twitter (GESIS – Working Papers No. 31). https://www.gesis.org/en/services/sharing-knowledge/publications/archive/gesis-working-papers

Kinder - Kurlanda, K., Weller, K., Zenk - Möltgen, W., Pfeffer, J., & Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. Big Data & Society, 4(2), 1–14. https://doi.org/10.1177/2053951717736336

Kuhar, R., Pajnik, M., Učakar, T., Ješe Perković, A., Zalta, A., Mandelc, D.,Šori, I. (2021). Državljanstvo in diskriminacija: Intersekcijski pristop k raziskovanju družbene izključenosti, 2017 [Podatkovna datoteka]. Ljubljana: Univerza v Ljubljani, Arhiv družboslovnih podatkov. https://doi.org/10.17898/ADP_INTERD17_V1

Lenardič, J., Erjavec, T., & Fišer, D. (2020). Citiranje jezikovnih podatkov v slovenskih znanstvenih objavah v obdobju 2013–2019. Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave, 8(1), 1–34. https://doi.org/10.4312/slo2.0.2020.1.1-34

Mahanty, A. (2022). waybackpy 3.0.6 [code Python package]. https://pypi.org/project/waybackpy/

Peer, L. (2024). Why and How We Share Reproducible Research at Yale University's Institution for Social and Policy Studies. Harvard Data Science Review, 6(1). https://doi.org/10.1162/99608f92.dca148ba

RDA Alliance. (2024). Data policy standardisation and implementation IG. Pridobljeno 31. maja 2024, https://www.rd-alliance.org/groups/data-policy-standardisation-and-implementation-ig/

Smrdelj, R., Učakar, T., & Kuhar, R. (2021). Marginalizacija intersekcijske perspektive: istospolna partnerstva in tujci na slovenskih novičarskih spletnih portalih. Javnost – The Public, 28(sup1), S122–S140. https://doi.org/10.1080/13183222.2021.2012945

Software Heritage archive. Pridobljeno 31. maja 2024, https://archive.softwareheritage.org/

Bezjak, S. (2022). Izzivi publiciranja in citiranja podatkov v družboslovju. 31. maja 2022. https://www.adp.fdv.uni-lj.si/blog/2022/05/31/25-let-adp/

SSRC and QDR (2019). Addressing Copyright (managing-qualitative-data.org). Pridobljeno 31. maja 2024, https://managing-qualitative-data.org/modules/3/c/

Stall, S. (2024). Complex Citations: Working Towards Recommendations. Pridobljeno 31. maja 2024, https://www.rd-alliance.org/plenaries/complex-citations-working-towards-recommendations/

Suplement, Javnost (2021). Suplement, Javnost – The Public, Vol. 28 – 2021, Suplement, Pridobljeno 31. maja 2024, https://javnost-thepublic.org/issue/2021/5/

Štebe, J. (2021a). Zadeva: Arhiviranje podatkov v ADP – Suplement Javnosti. Osebna korespondenca avtorja s člani uredniškega odbora.

Štebe, J. (2021b). Zadeva: Arhiviranje spletnih člankov, ki so podlaga znanstvenemu članku. Osebna korespondenca avtorja z Jankom Klasincem.

Štebe, J. (2024). Read web address from the text file and write add Wayback Archive address (v1.0.2). Zenodo. https://doi.org/10.5281/zenodo.10992412

Štebe, J., Bezjak, S., Alvanides, S., Recker, J., Glavica, M., Kranjec, I., Laaksonen, H., Kondyli, D., Klironomos, N., Linardis, A., & Kleiner, B. (2022, June 9). The CESSDA Data Archives joint efforts to support journals in data sharing and reproducibility [Panel]. Zenodo. https://doi.org/10.5281/zenodo.6683796

Štebe, J., Bezjak, S., & Dolinar, M. (2020a). Smernice za oblikovanje politik znanstvenih založb glede navajanja raziskovalnih podatkov v znanstvenih publikacijah in zagotavljanja dostopa do primarnih podatkov, uporabljenih v člankih (2.4). Zenodo. https://doi.org/10.5281/zenodo.3757288

Štebe, J., Bezjak, S., & Dolinar, M. (2020b). Pojasnila za razumevanje politik znanstvenih založb glede zagotavljanja dostopa do primarnih podatkov, uporabljenih v člankih (1.0). Zenodo. https://doi.org/10.5281/zenodo.3754783

Štebe, J., Bezjak, S., Alvanides, S., Recker, J., Glavica, M., Kranjec, I., Laaksonen, H., Kondyli, D., Klironomos, N., Linardis, A., & Kleiner, B. (2022, June 9). The CESSDA Data Archives joint efforts to support journals in data sharing and reproducibility [Panel]. Zenodo. https://doi.org/10.5281/zenodo.6683796

Štebe, J., Dolinar, M., Bezjak, S., & Inkret, A. (2020c). Implementing the RDA Research Data Policy Framework in Slovenian Scientific Journals. Data Science Journal, 19(1), 49. https://doi.org/10.5334/dsj-2020-049

T & F (2024a). Taylor & Francis Editorial Policies. Pridobljeno 31. maja 2024, https://authorservices.taylorandfrancis.com/editorial-policies/

T & F (2024b). Sharing and citing data. Pridobljeno 31. maja 2024, https://authorservices.taylorandfrancis.com/data-sharing/

Thomson, S. D. (2016). Preserving Social Media. Digital Preservation Coalition. DOI: http://dx.doi.org/10.7207/twr16-01

Wayback Machine (2024). Pridobljeno 31. maja 2024, https://archive.org/web/

Weeden, K. A. (2023). Crisis? What Crisis? Sociology’s Slow Progress Toward Scientific Transparency. Harvard Data Science Review, 5(4). https://doi.org/10.1162/99608f92.151c41e3

MACHINE TRANSFER OF INTERNET ADDRESSES OF THE ONLINE NEWS ON THE WAYBACK ARCHIVE ADDRESSES

In the article, we present the ongoing activities of collaboration between data repositories and scientific journals in Slovenia regarding the introduction of citation and access to research data that accompany publications in scientific literature. As a continuation of these activities, the Social Science Data Archive (ADP) led a pilot collaboration with several social science journals. An example of collaboration with the Slovene journal *Javnost* (*The Public*) is shown, which posed a particular challenge for archiving, as the data consisted of news articles discussed in one of the journal's papers. The challenge for redistribution lies in copyright and related rights, while there is also no guarantee that news portals will preserve the published content. A simple Python script automated the transfer of internet addresses to the *Wayback Machine*, and a sample list of all internet addresses considered in the research, whose related news articles represent the research data of the given paper, was stored in the ADP's domain-specific data archive. In the discussion of the presented case, we once again highlight the advantages of storing and accessing data that accompany scientific literature.

Keywords: reproducibility, research data, scientific software code, scientific publishing

OSEBNI PODATKI V UMETNOSTI: NJIHOVA ZAKONITA OBDELAVA IN VLOGA ETIKE V NOVOMEDIJSKI KULTURI

Aleš VAUPOTIČ,[1] Narvika BOVCON[2]

[1]Univerza v Novi Gorici

[2]Univerza v Ljubljani

Besedilo obravnava vprašanja zakonite in etične obdelave osebnih podatkov v novomedijski kulturi in umetnosti ter dileme ob vključevanju takih kulturnih artefaktov v razstavne projekte in muzejske zbirke. Uvodoma predstavi Splošno uredbo o varstvu podatkov (2018), ki sta jo sprejela Evropski parlament in Svet Evropske unije, ter trenutne zakonodajne postopke Evropske unije za področje umetne inteligence. Problematiko varovanja osebnih podatkov osvetli skozi razmislek o etiki informacij, kot jo razvija Luciano Floridi, ter pojasni razmerje med trdo in mehko etiko, slednja se izkaže za posebej uporabno za razumevanje vloge novih informacijskih tehnologij v današnji družbi. Vlogo osebnih podatkov, obdelanih tudi z algoritmi umetne inteligence, in dileme, povezane z njimi, ilustrira na primerih novomedijskih umetniških del Vuka Ćosića in Sreča Dragana.

Ključne besede: Splošna uredba o varstvu podatkov, Vuk Ćosić, Srečo Dragan, mehka etika, novomedijska umetnost

1 UVOD

Novomedijska umetniška dela podobno kot ostale digitalne storitve zbirajo in uporabljajo podatke, ki so povezani z uporabniki.[1] Vprašanja o zakonitosti in etičnosti obdelave podatkov povezujejo novomedijsko umetnost s splošnimi vprašanji informacijske etike in etike hitro razvijajočega se področja, ki ga imenujemo umetna inteligenca.[2] Katere osebne podatke uporabnikov

---

[1]Besedilo uporablja spolno nevtralno množino, npr. uporabniki, za moške in ženske ter vse ostale spolne identitete, v ednini pa uporablja ženski spol kot spolno nevtralni, kadar gre za primer fizične osebe (npr. umetnica). V navedke ne posegamo.

[2]Za pomen izraza umetna inteligenca, ki se ne nanaša na t. i. splošno umetno inteligenco, ampak na računalniške postopke, ki se danes dejansko uporabljajo, prim. *Opredelitev umetne inteligence: glavne zmogljivosti in znanstvene discipline* (Strokovna skupina na visoki ravni za umetno inteligenco pri evropski komisiji, 2019).

umetniških instalacij lahko zbiramo, pod kakšnimi pogoji, kako jih hranimo in obdelujemo, kako naj postopa umetnica in kako galerija ali muzej? Kaj je zasebnost v času informatizacije? Ali se umetna inteligenca uči na osebnih podatkih? Splošne smernice za odgovarjanje na ta povsem praktična vprašanja bosta podala pregled aktualne zakonodaje v zvezi z varovanjem osebnih podatkov in premislek širših etičnih in filozofskih okvirov, v katerih je zakonodaja utemeljena, kar bo omogočalo ustrezno razumevanje problematike ter z zakonodajo skladno rabo osebnih podatkov tako pri načrtovanju kot tudi ohranjanju in poznejših rekonstrukcijah novomedijskih umetniških del.

V drugem poglavju besedilo oriše evropsko uredbo za področje osebnih podatkov, v tretjem poglavju osvetli razmerje med etiko in pravom v kontekstu digitalizacije in inovacij s področja informatike, v četrtem poglavju pojasni koncept zasebnosti kot informacijske celovitosti osebe. V študiji primerov v petem poglavju razčleni novomedijske umetniške projekte z vidika obdelave osebnih podatkov.

2 SPLOŠNA UREDBA O VARSTVU PODATKOV

Splošna uredba o varstvu podatkov (SUVP, General Data Protection Regulation, GDPR) se v Evropski uniji uporablja od 25. 5. 2018. Njen namen je varstvo posameznikov pri obdelavi osebnih podatkov.[3] Temelji mdr. na Evropski konvenciji o varstvu človekovih pravic, ki v 8. členu pravi: »Vsakdo ima pravico do spoštovanja njegovega zasebnega in družinskega življenja, doma in dopisovanja.«

Kdaj je podatek osebni podatek? Osebni podatki so povezani s posameznico, ko jo je mogoče prek njih določiti. Primeri so: »ime, identifikacijska številka, podatki o lokaciji, spletni identifikator, ali z navedbo enega ali več dejavnikov, ki so značilni za fizično, fiziološko, genetsko, duševno, gospodarsko, kulturno ali družbeno identiteto tega posameznika« (1. alineja 4. člena SUVP). Torej,

[3]SUVP se uporablja neposredno. Zakon o varstvu osebnih podatkov (ZVOP-2) v Sloveniji dodatno ureja nekatera vsebinska in postopkovna vprašanja, ne sme pa spreminjati odločb Splošne uredbe,
https://www.ip-rs.si/zakonodaja/reforma-evropskega-zakonodajnega-okvira-za-varstvo-osebnih-podatkov/najpogostejsa-vprasanja-in-odgovori.

imena oseb so vedno osebni podatki, standardni identifikatorji, telefonske številke, uporabniška imena na družabnih omrežjih, (elektronski) naslovi. Fotografija osebe – oz. katerikoli biometričen podatek – je osebni podatek, ker, oz. če, omogoča identifikacijo.[4] Spletni piškotki so lahko osebni podatki.[5] Kombinacije podatkov, torej podatki v povezavi z drugje dostopnimi podatki, lahko omogočijo identifikacije in s tem postanejo osebni podatki. V primeru osebnih podatkov v obliki psevdonimov[6] je potrebno upoštevati, kako zahteven, dolgotrajen ali kakorkoli dosegljiv je postopek, da se identificira konkretna oseba – anonimizirani podatki zato niso osebni podatki. Odločilno je torej vprašanje, ali je mogoče določiti, za katero fizično osebo gre, ob upoštevanju dosegljivih finančnih in tehničnih sredstev v trenutku obdelave ter drugih omejitev.

Uredba velja za podatke oseb, ki so v Evropski uniji, ne samo za podatke državljanov članic in ne glede na to, ali je sedež podjetja, ki upravlja ali obdeluje osebne podatke, v Evropski uniji. Gre npr. za osebne podatke strank ali obiskovalcev spletišč. Luciano Floridi v *Etiki umetne inteligence* (2023) zapiše, da je posebna odlika SUVP njen pristop,

> ko izkorišča »povezavo« osebne identitete in osebnih informacij, da bi obšla »ločitev« prava in ozemeljskosti. To doseže s tem, da utemelji varstvo osebnih podatkov v prvem (na koga so »pritrjeni«, kar je zdaj odločilno) in ne v drugem (kje se obdelujejo, kar ni več pomembno) (str. 8).

Strežniki ali partnerji in pravne osebe, ki so zunaj Evropske unije, potemtakem ne omogočajo, da bi se obšlo evropska pravila. Gre za pomembno vprašanje, ki se bo z rabo različnih oblik generativne umetne inteligence še zaostrovalo. Floridi opozarja na dvojnost dilem raziskovalne etike in potrošniške etike.

> Npr. podjetje bi lahko izvozilo svoje raziskave in potem zasnovalo, razvilo in naučilo algoritme (npr. za prepoznavanje obrazov) na lokalnih osebnih podatkih v državi zunaj EU, z drugačnim, šibkejšim, neizvajanim etičnim in pravnim okvirjem za varstvo osebnih podatkov. V skladu s SUVP bi bilo to neetično in nelegalno v EU. Vendar pa bi se algoritmi, ko bi bili naučeni, nato lahko uvozili v EU in uporabili brez

[4] https://www.ip-rs.si/mnenja-gdpr/6048a64130f43

[5] Prim. 225. člen Zakona o elektronskih komunikacijah (ZEKom-2) in https://gdpr.eu/cookies.

[6] Psevdonimizirane podatke je mogoče z uporabo dodatnih informacij pripisati posameznici (Uvodna izjava 26 in 5. alineja 4. člena SUVP).

> sankcij, ali da bi to sploh bilo dojeto kot sporno. /... Potrošnja neetičnih raziskovalnih rezultatov je v primerjavi z raziskovalno etiko/ bolj nejasna, manj očitno problematična in jo je zato težje nadzorovati in omejiti (Floridi, 2023, str. 73).[7]

Upravljavec osebnih podatkov je pravna ali fizična oseba (npr. zastopnica društva ali ustvarjalka), ki določa namene in sredstva obdelave. »Upravljavec je odgovoren za skladnost /.../ in je to skladnost tudi zmožen dokazati ('odgovornost')« (2. odst. 5. člena SUVP). Obdelovalec dejansko obdeluje podatke, lahko je to upravljavec ali pa izvajalec v imenu upravljavca.[8] Obdelava podatkov vključuje pravzaprav vse, kar se dela z njimi, avtomatično ali ne: »zbiranje, /.../ urejanje, strukturiranje, /.../ spreminjanje, /.../ vpogled, /.../ razširjanje ali drugačno omogočanje dostopa, /.../ izbris ali uničenje« (2. alineja 4. člena SUVP). Splošna uredba ne velja za »fizične osebe med potekom popolnoma osebne ali domače dejavnosti« (2. člen SUVP), kar pomeni, da tudi fizična oseba lahko postane upravljavka osebnih podatkov, če posreduje tuje osebne podatke, in je lahko za kršitve pri tem početju tudi kaznovana.[9]

Namen SUVP, kot je razvidno iz njenega celotnega naslova »o varstvu posameznikov pri obdelavi osebnih podatkov in o prostem pretoku takih podatkov«, je poleg varstva tudi prosti pretok osebnih podatkov. Kaj je torej dovoljeno in pod kakšnimi pogoji? Slediti je treba sedmim načelom (5. člen SUVP). (1) »Zakonitost, pravičnost in preglednost,« potrebno je objaviti obvestilo o obdelavi osebnih podatkov, ki je razumljivo za ciljno skupino uporabnikov, in si prizadevati za uresničevanje pravic posameznikov, na katere se nanašajo osebni podatki. (2) »Omejitev namena,« posameznici je treba predstaviti jasno določene in zakonite namene, ki nato določajo meje

---

[7]Prim. tudi celotna poglavja 5.5 Ethics Dumping, 5.2 Ethics Shopping in 5.3 Ethics Bluewashing.

[8]Npr.: »Arnes nudi storitve /obdelave osebnih podatkov/ članicam omrežja ARNES, preko njih pa njihovim končnim uporabnikom (zaposlenim, šolajočim ipd.). Poleg tega lahko Arnes v dogovoru z ustreznimi resornimi ministrstvi nudi storitve tudi nekaterim posebnim kategorijam posameznikov (samostojni raziskovalci in kulturni delavci, invalidi). Članice omrežja Arnes /.../ same upravljajo e-identitete svojih uporabnikov in pri tem obdelujejo njihove osebne podatke. Z izdajanjem teh identitet (AAI-računov) članica omogoča svojim uporabnikom uporabo Arnesovih storitev.« https://www.arnes.si/zavod-arnes/katalog-informacij/varovanje-zasebnosti-uporabnikov-storitev-omrezja-arnes.

[9]Npr. avstrijski nadzorni organ je kaznoval fizično osebo na podlagi SUVP, https://gdprhub.eu/index.php?title=DSB_(Austria)_-_2021-0.518.795.

dovoljenega zbiranja in obdelovanja osebnih podatkov. (3) Obdeluje se »najmanjši obseg podatkov«. (4) »Točnost,« podatki se morajo posodabljati, netočne je treba brisati. (5) »Omejitev shranjevanja« na le toliko časa, kolikor je nujno za namen obdelave.[10] (6) »Celovitost in zaupnost,« zagotoviti je treba varnost, npr. s šifriranjem. (7) »Odgovornost« (predvsem) upravljavca.

Poleg preglednosti glede namenov zbiranja je pomembna tudi skrb za varnost osebnih podatkov. Uredba sledi načelu »vgrajenega in privzetega varstva podatkov« (Uvodna izjava 78 SUVP), ki zahteva tehnične in druge postopke. Če pride do vdora ali kraje ipd. (nešifriranih, tj. uporabnih) podatkov, tudi strojne opreme s podatki, je treba takoj oz. najkasneje v 72 urah obvestiti posameznike (in pri nas Informacijskega pooblaščenca Republike Slovenije). Organizacijski vidik zagotavljanja varnosti pomeni mdr. določitev odgovornih oseb v organizaciji in omejitev dostopa.

Kdaj je obdelava osebnih podatkov dovoljena? Obdelava osebnih podatkov je zakonita, upoštevajoč vsaj enega od šestih pogojev (6. člen SUVP): (1) privolitev, (2) če je to potrebno za pripravo in izvajanje pogodbe, (3) zakonska obveznost, ki velja za upravljavca, (4) reševanje življenj, (5) »opravljanje naloge v javnem interesu ali pri izvajanju javne oblasti, dodeljene upravljavcu« (npr. privatno komunalno podjetje); (6) zakoniti interes (najbolj zahteven, ker upravljavec prevzame nase uresničevanje pravic posameznikov).

Privolitev posameznice, da se njeni osebni podatki obdelujejo, vključuje ločeno jasno prošnjo za vsak namen posebej, privolitev je treba dokumentirati, omogočiti je treba preklic privolitve; podlage za obdelavo seveda ni mogoče naknadno spremeniti. Pred 13. letom starosti se zahteva tudi dovoljenje staršev (v Sloveniji pred 15. letom, prim. 8. člen Zakona o varstvu osebnih podatkov, ZVOP-2).

[10]»/O/sebni podatki se lahko shranjujejo za daljše obdobje, če bodo obdelani zgolj za namene arhiviranja v javnem interesu, za znanstveno- ali zgodovinskoraziskovalne namene ali statistične namene v skladu s členom 89(1), pri čemer je treba izvajati ustrezne tehnične in organizacijske ukrepe iz te uredbe, da se zaščitijo pravice in svoboščine posameznika, na katerega se nanašajo osebni podatki.« V 89. členu sta omenjeni psevdonimizacija in obdelava podatkov, ki prepreči identifikacije posameznikov.

Vprašanje privolitve je vsebinsko povezano z zaupanjem na področju digitalne obdelave podatkov in umetne inteligence, kjer je poleg različnih kršitev in nezakonitih ter neetičnih ravnanj poseben problem razložljivost delovanja uporabljenih algoritmov in njihovih rezultatov (prim. *Etične smernice za zaupanja vredno umetno inteligenco* Strokovne skupine na visoki ravni za umetno inteligenco, ki jo je ustanovila Evropska komisija, 2019). V zvezi s privolitvijo je mogoče omeniti problematični pojav širitve prvotnega pomena (function creep) (Koops, 2021), ki pomeni obliko zavajanja (Floridi, 2023, pogl. 5.3 Ethics Bluewashing), ali pa pritiske in pogojevanja, ki onemogočijo prostovoljnost izbire, in uporabo po nepotrebnem zapletenega oz. za skupino naslovnikov prezahtevnega jezika. Molk ali vnaprej izpolnjeni obrazci ter taktika utrujanja (click fatigue) (Hornuf in Mangold, 2022) niso oblike veljavne privolitve. Obstajata veljavna in izrecna privolitev, ki je potrebna za obdelavo t. i. posebnih vrst podatkov (o rasi, politični usmeritvi ipd., 9. člen SUVP), ob prenosu podatkov v tretje države ali mednarodne organizacije, pri čemer niso sprejeti ustrezni zaščitni ukrepi (49. člen SUVP), in ko odločanje o posameznici temelji izključno na avtomatizirani obdelavi, vključno s profiliranjem. Izrecna privolitev naj bi bila pisna, lahko je elektronska, svetuje se dvostopenjsko preverjanje privolitve (npr. prek potrditvenega SMS). Preklic privolitve mora biti enako enostaven kot sama privolitev, o načinu preklica mora biti posameznica obveščena pred privolitvijo.

SUVP obravnava uresničevanje pravic posameznice, na katero se nanašajo osebni podatki, in to obsega obvezo seznanjanja oseb, katerih podatki se obdelujejo, tako glede samih podatkov in načinov obdelave kot tudi glede pravic posameznice v skladu z načelom preglednosti: dostop do osebnih podatkov posameznice, ki jih obdeluje upravljavec, popravek, izbris (pravica do pozabe), omejevanje obdelave, prenosljivost (npr. posredovanje digitaliziranih podatkov v strojno berljivi datoteki na uporabničino zahtevo), pravica do ugovora (ob avtomatizirani obdelavi in ustvarjanju profilov). Upravljavec mora zagotoviti izvajanje teh pravic najkasneje v roku enega meseca od zahteve in, načeloma, brezplačno (3. odst. 12. člena SUVP).

3 PRAVO, ETIKA IN DIGITALIZACIJA MEDIJEV

3.1 Akti Evropske unije o umetni inteligenci, digitalnih storitvah in digitalnih trgih

Evropski parlament je 13. 3. 2024 podprl Akt o umetni inteligenci. Gre seveda za področje, ki se zelo hitro razvija in je težko obvladljivo. Pristop temelji na določitvi stopnje tveganja pri končni uporabi umetne inteligence: če je stopnja visoka, bo omejitev več, če je stopnja nizka, bo regulacije manj ali je ne bo. Akt o umetni inteligenci uvaja štiri kategorije glede na stopnjo tveganja:

1. nesprejemljiva stopnja tveganja – takšna raba algoritmov umetne inteligence je prepovedana (rangiranje posameznikov v družbi, govoreče igrače, ki napeljujejo na nevarno obnašanje ipd.).
2. Visoko stopnjo tveganja pomeni npr. raba umetne inteligence v sodbah sodišč (Angwin in sod., 2016) – tovrstni sistemi bodo registrirani pri Evropski uniji, potreben bo certifikat, ob bistveni spremembi se preverjanje ponovi.
3. Stopnja omejenega tveganja se osredotoča na vprašanje preglednosti delovanja, npr. če se uporabnica na spletu pogovarja s sistemom umetne inteligence, je treba to predhodno izrecno označiti in sporočiti.
4. Akt dovoli uporabo umetne inteligence, kadar predstavlja minimalno tveganje, primer poleg računalniških iger so filtri za neželeno spletno pošto.[11]

Pristop se torej podobno kot v zvezi z osebnimi podatki osredotoča na učinke na uporabnike in njihovo informacijsko integriteto ter uvaja pravila, izhajajoč iz ocene tveganj v vplivih računalniških oz. hibridnih sistemov (tj. storitev, kjer v obdelavi sodelujejo tudi ljudje). Kot ni mogoče živeti brez obdelave osebnih podatkov, tako tudi ostalih (digitalnih) podatkov ne bo mogoče kratko malo ubraniti pred umetno inteligenco.

Omenjeni akt se povezuje z Aktom o digitalnih storitvah, ki sledi načelu, kar je nezakonito v realnem svetu, je nezakonito tudi na spletu (lažne novice,

[11] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

otrokom namenjene reklame …), dotika se v prvi vrsti družbenih medijev, platform in spletnih tržnic (pri nas npr. Bolhe), uporablja se od 17. 2. 2024.[12]

Akt o digitalnih trgih pa ščiti podjetja pred zlorabo monopolov globalnih platform. Uporablja se

> le za velika globalna podjetja. Ta so v uredbi v skladu z objektivnimi merili opredeljena kot vratarji (ang. gatekeepers). Gre za podjetja, ki nadzorujejo vsaj eno jedrno platformno storitev, kot so iskalnik, družbeno omrežje, storitve za pošiljanje sporočil, operacijski sistem ali spletna tržnica, ter imajo veliko uporabnikov v več državah Unije. Vratarji bodo morali ostalim podjetjem omogočati interoperabilnost s svojimi storitvami, dostop do podatkov.[13]

Trenutni seznam vključuje podjetja Alphabet, Amazon, Apple, ByteDance, Meta, Microsoft, ki morajo od 7. 3. 2024 dalje v celoti ustrezati zahtevam. Evropska unija je že 25. 3. sporočila, da je v zvezi z morebitnim kršenjem Akta uvedla preiskavo Mete, Appla in Alphabeta (ki ima v lasti Google).[14]

3.2 Mehka etika kot podpora za izbiro med možnostmi v okvirih zakonitega delovanja

Poseben problem – poleg nesorazmerja moči – predstavljata na področju digitalne komunikacije preglednost in razumljivost delovanja algoritmov, ki so s strojno obdelavo podatkov na še pred kratkim nesluten način olajšali prenosljivost in zbiranje podatkov. (Gre tudi za tehničen problem, ki pa v tem besedilu ne bo obravnavan.) Zakonodaja v tem kontekstu pravzaprav pričakovano zamuja, kar pa ni opravičilo. Nove oblike obdelave podatkov praviloma vključujejo možnost obdelave osebnih podatkov – npr. z merjenjem fizičnega videza in obnašanja posameznikov je osebo kaj hitro mogoče določiti. Kako se v takem okolju ravnati, kdaj smo v mejah zakona, kdaj delujemo etično? Luciano Floridi ob obravnavi izzivov in priložnosti umetne inteligence pregleda in komentira rešitev, ki jo predstavlja SUVP, ta deluje v

[12]Prim. https://www.consilium.europa.eu/sl/press/press-releases/2021/11/25/what-is-illegal-offline-should-be-illegal-online-council-agrees-on-position-on-the-digital-services-act in https://www.gov.si/novice/2024-02-19-zakljucena-prva-obravnava-predloga-zakona-o-izvajanju-uredbe-eu-o-enotnem-trgu-digitalnih-storitev.

[13]https://www.gov.si/novice/2022-11-04-v-veljavo-stopil-akt-o-digitalnih-trgih

[14]Prim. https://www.bbc.com/news/technology-68655093 in podrobneje https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1689.

mešani resničnosti digitalnega in analognega, s spletom povezanega in nepovezanega.

Floridi pri razumevanju primera SUVP ločuje mehko in trdo etiko (2023, 6. pogl.). Delitev je privzeta iz prava, ki razlikuje trdo in mehko pravo, slednje predstavljajo resolucije in priporočila, ki niso zavezujoča (Shaffer in Pollack, 2010). Trda etika obstaja pred zakonodajo in jo nato tudi oblikuje, Floridi jo poveže z načelom Rose Parks, ki se je v svoji akciji državljanske nepokorščine odločila za neupoštevanje rasističnega zakona. V primeru trde etike ne pomeni, da, če bi s pravnega vidika oseba nekaj morala narediti, to sme narediti tudi z vidika etike, zakon je lahko neetičen. Upor kot aktivna trda etika vodi – vsaj v nekaterih primerih – v spremembo zakonodaje. Nasprotno t. i. mehka etika nastopi ob spoštovanju zakonov, tukaj velja načelo, če je nekaj predpisano z zakonom, je to tudi etično dovoljeno.

Slika 1: Prostor za mehko etiko.

Na Sliki 1 je diagramatično označen konceptualni prostor za mehko etiko. V skladu s Kantom etične zahteve privzemajo, da je takšno delovanje tudi izvedljivo, saj ni etično zahtevati nekaj, kar je nemogoče (spodnja črta). Skozi čas se obseg izvedljivega zaradi tehničnih inovacij širi. Mehka etika je torej omejena s tem, kar je moralno dobro (v Evropski uniji so to najmanj človekove pravice) in kar je zakonito, onkraj tega je protizakonito. Floridi doda, da je mehka etika neuporabna v nemoralnih družbah (ni možnosti zakonitega delovanja in hkrati npr. spoštovanja človekovih pravic), omeni Kitajsko, dodaja pa, da so tudi v Evropski uniji potrebni posegi trde etike, npr. v primeru irskega referenduma o splavu iz leta 2018, ki je prinesel ustavni amandma.

> Znotraj EU se lahko mehka etika upravičeno uporablja za pomoč agentom (vključno s posamezniki, skupinami, podjetji, vladami, organizacijami), da z moralnega vidika bolj in bolje izkoristijo priložnosti, ki jih ponujajo digitalne inovacije. Kajti tudi za EU velja, zakonodaja je potrebna, vendar nezadostna (Floridi, 2023, str. 84).

Primer mehke etike je samoregulacija. Strah pred tehnologijo pogosto vodi v potiskanje glave v pesek z negativnimi posledicami. Kar je na tem mestu bolj pomembno, je, da je tudi prostor za zakonito delovanje pravzaprav širok, nabor možnih izbir je velik in ustrezno razumevanje posledic, npr. posledic računalniške obdelave podatkov, ni preprosta naloga. Že sama določitev, kaj je osebni podatek, je prepuščena interpretaciji in odločitvi upravljavke (oz. podjetja), ki sledi prej očrtanim načelom iz SUVP in drugim etičnim premislekom, hkrati pa včasih sprejema odločitve ob uporabi povsem novih tehničnih postopkov in algoritmov. Priložnost, ki jo omenja Floridi, ni primer zaslepljene vere v inovacije, ampak upošteva, da trenutno glavni izziv ni več izumljanje novega, temveč upravljanje z digitalnimi in ostalimi informacijskimi viri, gre za tretjega izmed normativnih pristopov, ki dopolnjuje etiko in pravo. Dobro upravljanje namreč samo po sebi ni nujno vezano na dvojnost moralnega in nemoralnega ali zakonitega in nezakonitega (str. 79).

Floridi mehko etiko razume kot ogrodje, ki pomaga pri razumevanju etičnih dilem, ki se pojavljajo. Za razumevanje njene vloge je treba upoštevati pet komponent, ki se povezujejo po dveh načelih, načelu generiranja in načelu interpretiranja (pogl. 6.5). Trda etika je podlaga za nastanek zakonodaje, predvsem devetindevetdesetih členov SUVP, pa tudi za sto triinsedemdeset Uvodnih izjav. Zakon nato generira etične, pravne in družbene implikacije in priložnosti, ki neposredno sledijo členom, kot tudi vprašanja, ki se deloma ali v celoti razkrijejo ob SUVP (upoštevati je treba, da priložnosti niso nujno pozitivne z vidika družbene dobrodejnosti). Povezave od mehke etike do Uvodnih izjav in naprej do členov potekajo prek interpretiranja: Uvodne izjave pojasnjujejo razloge in določbe zakonov in same po sebi niso zavezujoče, hkrati pa so neposredno orodje za pojasnjevanje členov. Mehka etika prispeva k interpretiranju Uvodnih izjav in s tem samih členov. Kot že rečeno, skladna mora biti s trdo etiko, ki je podlaga zakonodaje. Prispeva tudi k načinom, kako se je mogoče konstruktivno spopasti z izzivi etičnih in družbenih implikacij zakonodaje s področja digitalne kulture.

Na tem mestu ni mogoče podrobneje razviti konkretnih etičnih načel, ki omogočajo orientacijo na področju današnje poplave podatkov in izzivov njihovih obdelav. Omenimo, da Floridi ob primerjavi bioetike z etiko umetne inteligence dodaja štirim bioetičnim principom – dobrodejnost, preprečevanje škode, spoštovanje človekove avtonomije, pravičnost – še dodatnega petega, razložljivost, ki hkrati pomeni nov, še ne razrešen izziv, saj so izhodni podatki algoritmov včasih celo za ustvarjalce in uporabnike algoritmov nerazložljivi, tudi glede pomembnih vidikov, npr. avtomatičnega določanja (str. 64).[15] Aktivno upoštevanje in razvijanje digitalne etike ponuja dve prednosti. Ponuja podporo za strateško odločanje o priložnostih – previdnost mora biti uravnotežena z dolžnostjo, da se ne izognemo temu, kar bi bilo treba narediti: npr. uporabiti obstoječe podatkovne zbirke in, v primeru umetne inteligence, zaloge »pametne« aktivnosti (agency) nečloveških in hibridnih agentov. Po drugi strani je mehka etika rešitev za upravljanje s tveganji – etični premislek na začetku načrtovanja projektov pomeni prednosti kasneje, oz. preprečevanje napačnih potez. S tega vidika sta seveda smiselna – še posebej v tem trenutku, ko se sistemi obdelave podatkov in njihove uporabe v algoritmih pametnih umetnih agentov oblikujejo za bližnjo in bolj oddaljeno prihodnost – principa iz Uvodne izjave 78 SUVP o privzetem in vgrajenem varstvu podatkov. Varovanje osebnih podatkov mora biti izhodišče pri zasnovi in ne naknadni dodatek k novim informacijskim sistemom.

4 ZAKAJ JE TREBA VAROVATI OSEBNE PODATKE?

Na videz neobičajno vprašanje se vendarle zastavlja upravičeno, saj načrtovanje mehanizmov za varovanje osebnih podatkov ob tako rekoč vseh obdelavah podatkov predstavlja precejšnje breme za delo na številnih področjih. Zakaj je težava, če nekdo shrani podatke o meni? Saj nič ne izgubim? Ali če nekaj vendarle izgubim, kakšen je pravzaprav negativni učinek?

---

[15]Prim. tudi *Etične smernice za zaupanja vredno umetno inteligenco* Strokovne skupine na visoki ravni za umetno inteligenco pri Evropski komisiji.

Gre za kompleksno problematiko, ki se ji Luciano Floridi izčrpno posveti v monografiji *Etika informacije*, na tem mestu pa je mogoče le nakazati smer razmišljanja (2013, predvsem pogl. 3.4.5 Informacijska zasebnost: od vdora do ugrabitve). Privzemimo, da osebni podatki niso oblika lastnine – zdi se nam, da ob digitalni reproduktibilnosti podatkov pravzaprav ne izgubim nič otipljivega. Floridi predlaga, da je ljudi mogoče razumeti kot skupke informacij, kot informacijske organizme. S tega vidika je prilaščanje podatkov iz skupka, ki je posamezničina osebnost, odvzem dela te celote in poseg v integriteto posameznice. Vendar, kaj konkretno oseba izgubi, ko so njeni osebni podatki shranjeni in obdelani kot kopija? Težava je, da je oseba hkrati navzoča npr. v javnem prostoru, hkrati pa prek nadzornih kamer tudi v prostoru, ki ga opazuje in nadzoruje nekdo drug. V tem drugem prostoru pa je oseba brez možnosti delovanja, saj nadzorovalka sama ne vstopa v opazovani oddaljeni prostor – nadzorovana oseba je, na metaforičen način, ugrabljena. Floridi pravi:

> zasebnost postane obramba osebne identitete in edinstvenosti. Nekonsistentnost med zasebnimi in javnimi prostori se /če je oseba razumljena kot informacijska celota/ ne pojavlja več: opazovana želi ohraniti svojo celovitost kot informacijska entiteta, tudi ko je na povsem javnem mestu. Navsezadnje je ugrabitev kaznivo dejanje ne glede na to, kje je storjeno, v javnosti ali ne. Kar nekdo kupi, obleče ali počne v javnosti, spada v sfero, ki ni nikogaršnja posebna last, vendar spremljanje in beleženje tega odvzame tej javni sferi del informacij, ki sestavljajo opazovano /osebo .../, in jo naredi del /oddaljenega/ prostora, ki ga ima v lasti in ga nadzoruje le opazovalka, do katerega opazovana sama nima drugega dostopa, in to na način, ki je za opazovano lahko popolnoma neviden (opazovana se pogosto ne zaveda, da je del njenih informacij ugrabljen) (str. 50–51).

Ilustrativen je Floridijev primer vsiljenega, vendar pasivnega prehoda v tujo zasebnost, ki je del sistematične obravnave te teme.

> Ko v naši bližini slišimo nekoga glasno govoriti po mobilnem telefonu, morda v omejenem prostoru vlaka, nas to pogosto moti. Nočemo poslušati o njenih zadevah, vendar si ne moremo pomagati. Paradoksalno je, da vemo, da ta oseba krši načelo naše zasebnosti /.../ In naša zasebnost zagotovo ni kršena, ker ona vstopa v naš informacijski prostor: navsezadnje smo mi tisti, ki poslušamo. Gre za to, da nas ugrabi v svoj informacijski prostor in nas sili, da smo proti svoji volji teleprisotni v njenem prostoru. Naša zasebnost je prizadeta, ker to je primer *vsiljene vzvratne* prisotnosti (str. 51).

Tukaj posega v osebno celovitost posameznice dodajanje in ne odvzem informacij.

5 OSEBNI PODATKI NA RAZSTAVAH IN V PODATKOVNIH ZBIRKAH, KI SO SESTAVNI DEL NOVOMEDIJSKIH UMETNIN

Številna novomedijska umetniška dela uporabljajo zajem fotografij in video posnetkov obiskovalcev razstave kot gradivo umetniškega sporočila, konceptualno utemeljen sestavni del interakcije in kot gradnik v tehnološkem postopku, ki je podvržen različnim algoritmom. Nekateri interaktivni umetniški projekti od obiskovalke zahtevajo vnos osebnih podatkov v sistem za vzpostavitev uporabniškega računa (funkcionalni razlog) oz. identitete v sistemu (konceptualni, narativni razlogi), od imena in priimka, EMŠO, kontaktnih podatkov, do ponekod celo demografskih določil in politično-ideoloških opredelitev ter biometričnih podatkov. Postavlja se vprašanje, kako naj muzejska in umetnostna stroka ravna s tovrstnimi projekti, kako naj jih razstavlja in vključi v muzejske zbirke. In, kako naj sama umetnost kot družbena praksa obravnava osebne podatke?

V okviru projekta *Trajnostna digitalna hramba slovenske novomedijske umetnosti* je bil problem osebnih podatkov obravnavan ob dveh pilotnih študijah rekonstrukcije umetniških del (prim. Vaupotič in sod., 2023). Obe novomedijski umetniški deli sta v časovnem razponu od prve izvedbe do rekonstrukcije leta 2022 zamenjali tehnološke platforme, tj. softver in hardver, ki so medtem zastarele, niso bile več v uporabi, vendar sta ohranili svoj temeljni umetniško-komunikacijski koncept, pri čemer pa je bil potreben ponoven premislek vseh vidikov projekta, mdr. tudi vidikov zasebnosti, ki jih od leta 2018 ureja SUVP.

5.1 Vuk Ćosić: *Nacija – Kultura* (2000, 2022)

Vuk Ćosić je razvil prvo izvedbo projekta *Nacija – Kultura* leta 2000 na takrat vodilnem spletnem imeniku in iskalniku v Sloveniji *Mat'Kurja*.[16] Šlo je za vmesnik, ki je sestavljal iskalne nize uporabnikov v obliko soneta. S tem je,

[16] Predstavitev verzije projekta iz leta 2000 v Mrežnem muzeju Moderne galerije, https://mrezni-muzej.mg-lj.si/si/mreznimuzej/3/22/?artworkid=1789.

lahko bi rekli, prisluškoval uporabnikom pri njihovem zasebnem iskanju po spletu in predvajal anonimne iskalne nize na video projekciji v galerijskem prostoru. Kot član ekipe okoli *Mat'Kurje* je imel dostop do toka iskanj. Postavlja se vprašanje, ali je sistem umetnine takrat kakorkoli hranil identiteto uporabnikov, ali so bili ob obdelavi podatki anonimizirani? Vsekakor leta 2000 v praksi ni bilo mogoče identificirati posameznikov, ki vnašajo iskalne nize v iskalno okence portala, sploh ne iz prikazanih podatkov na razstavi v Narodnem muzeju Slovenije (zato ne z vidika takrat veljavnega ZVOP in tudi ne z vidika trenutno veljavne SUVP ne gre za osebne podatke), v nasprotju z današnjim stanjem, ko je profiliranje uporabnikov uveljavljena praksa in pretočnost podatkov (npr. med sistemi) precej večja. Do danes se je spletni imenik *Mat'Kurja* tehnološko toliko spremenil, da na njem izvorni projekt ne deluje več, predvsem pa *Mat'Kurja* ni več v razširjeni uporabi, zaradi česar na tej platformi projekt ni več smiseln s konceptualnega vidika. Dalje, lahko si predstavljamo, da je danes splošno mnenje v družbi bolj kritično do tovrstnih prikazov iskalnih nizov, za katere se pričakuje, da niso tudi javno predvajani. Pravzaprav o izvornem projektu ne obstaja veliko javno dosegljivih sledi, zgolj peščica zajemov zaslona s soneti in posredna pričevanja v besedilih, ki projekt opisujejo.

Slika 2: Vuk Ćosić, *Nacija – Kultura*, 2022, MSUM, Ljubljana, 13. 10. 2022–8. 2. 2023, foto: Dejan Habicht.

Rekonstruirana *Nacija – Kultura*, na Sliki 2, je leta 2022 našla novo aktualno komunikacijsko platformo, in sicer Twitter (še pred Muskovim prevzemom platforme in njenim preimenovanjem v X).[17] Vuk Ćosić in Marko Plahuta, programer, sta sonete iz slovenskih čivkov sestavila na sledeč način: zajela sta slovenske čivke in naučila nevronsko mrežo (GPT-2) pisati v jeziku slovenskih čivkov, v naslednjem koraku sta jo doučila sloga slovenske poezije na slovenskih sonetih in drugih (prosto dostopnih) digitaliziranih pesmih. Po učenju mreže so posamezni čivki anonimizirani v velikem jezikovnem modelu. Nato sta zajela dnevne čivke izbranih slovenskih politikov in medijev ter jih uporabila kot iztočnice, pozive in uvodne verze, iz katerih je nevronska mreža do konca napisala (skoraj smiseln) sonet. Na več točkah v postopku sta Ćosić in Plahuta presojala, izboljševala in izbirala rezultate z nevronsko mrežo ustvarjenih besedil, vse do uredniškega dela pred knjižno objavo sonetov

[17]http://www.mg-lj.si/si/razstave/3567/razstava-nacija-kultura

(Ćosić, 2022b). Tokrat je vse gradivo, na katerem se je nevronska mreža učila, tj. čivki in kanonizirana slovenska poezija, javno objavljeno. Kriterij za odločanje, ali gre za protizakonito obdelavo že objavljenih osebnih podatkov ali ne, je namen – zakoniti namen je moral biti temelj prvotne objave in nato tudi vseh naslednjih. Ne sme se obdelovati javno objavljenih osebnih podatkov v druge namene.[18] (Poraja se seveda dodatno vprašanje avtorskih pravic na gradivu, iz katerega se nevronska mreža uči, pa tudi ponovnega posredovanja vsebin Twitterja, ki pa tukaj ne bo obravnavano.)

V obeh verzijah *Nacije – Kulture* je Ćosić postavil v ospredje etični vidik komunikacije, ki je specifična za določeno, izrazito tehnično posredovano komunikacijsko platformo. V primeru zgodnjega spleta smo imeli opravka z zasebnim, torej skritim iskanjem opolzkih in podobnih vsebin, za katere si želimo, da ostanejo diskretno zakrite, v primeru zrelega Twitterja pa imamo opravka z razbohotenim sovražnim govorom, ki je hkrati pomemben del javnega političnega diskurza. Ćosić obe z novimi informacijskimi tehnologijami prežeti komunikaciji pokaže v galerijskem prostoru in v obliki kultiviranega pesniškega izraza, tj. v obliki soneta (ki mu v Sloveniji tradicionalno pripisujemo celo narodotvornost) (Vaupotič, 2019, str. 218–219) – v tej primerjavi postane očitno, da taka načina izražanja nista kulturna, Ćosić naslov bere kot »Nacija minus Kultura«. Tehnološka platforma sooblikuje način komunikacije, ki se na njej odvija. Etičnost izjav govorcev (iskalnih nizov, čivkov), ki so del informacijske celovitosti govorcev, se presoja v odnosu do zasebnega ali javnega izjavljanja. Ćosićev projekt kritično presoja tudi etične vidike, vgrajene v delovanje algoritmov, ki poganjajo in spremljajo platforme, to so vidiki, za katere so odgovorni načrtovalci in upravljavci platform, zajemajo pa v prvi vrsti zasebnost komunikacije, profiliranje uporabnikov na osnovi njihovih osebnih podatkov, uporabljanje osebnih podatkov in izjav za druge namene od prvotnih, kot je mdr. učenje nevronskih mrež, pomembno vprašanje pa je tudi kršenje avtorskih pravic uporabnikov platform itn.

[18] https://www.ip-rs.si/mnenja-gdpr/objava-javno-objavljenih-podatkov

5.2 Srečo Dragan: *Matrica koincidenca zmenkarski klub* (2005, 2022)

Druga pilotna študija je v središče postavila rekonstrukcijo novomedijskega umetniškega projekta *Matrica koincidenca zmenkarski klub* (2005, 2022) avtorja Sreča Dragana (prim. Bovcon, 2009, str. 166–170).[19] V izvorni in tudi v rekonstruirani verziji projekt tematizira deljenje osebnih podatkov v času vse večje informacijsko-komunikacijske povezanosti, kar avtor pojasni v intervjuju leta 2007:

> Vendar pa vsa ta sofisticirana tehnologija ni bila sama sebi namen. V projektu sem raziskoval odnose med mobilno pogovorno-verbalno in gestualno komunikacijo. Obiskovalci, ki so prišli na razstavo, so se sprehodili pred digitalno video kamero in iz primerjave z zapisom hoje drugih udeležencev bili izbrani v pare, ki so se dobili v kavarni, se morda prvič videli in lahko celo zaživeli skupno življenje. To pomeni, da se medijska umetnost zaveda časa digitalne družbe ter ponuja nove oblike druženja. Seveda je bil udeleženec, ko ga je kamera spoznala na mestu srečanja, opozorjen, da mu prednosti povezav lahko vzamejo tudi del njihove zasebnosti.[20]

Zaradi obdelovanja osebnih podatkov je obiskovalki ob vstopu v galerijo najprej ponujen obrazec, v katerega vpiše svojo telefonsko številko, ki je potrebna za delovanje projekta, in podpiše, da se strinja z udeležbo v tehno performansu, ki poleg telefonske številke obdeluje še udeleženkino fotografijo (portret) in video posnetek njene, ali njegove ..., hoje. Udeleženki je uvodoma predstavljen potek tehno performansa v zaporedju treh laboratorijev. V Laboratoriju za biomehanske meritve udeleženka stopi na potisno ploščo s senzorji pritiska, algoritem v realnem času vizualizira (in predvaja na video projekciji) pritisk njenih stopal med plesom na plošči; pri tem se trajno ne shrani noben podatek, hkrati pa je treba dodati, da je avtorjev koncept prvega laboratorija, ki ustreza konceptu *matrice*, prav to, da je ples družbeno pogojeno in s tem neosebno gibanje, torej koreografija ne uteleša pristnih osebnih podatkov.

Na tej ravni je instalacija podobna videu zaprtega krogotoka iz preddigitalnega obdobja, signal se obdeluje, vendar kot odsev v zrcalu tudi takoj briše.

[19] http://www.mg-lj.si/si/dogodki/3664/posvet

[20] Intervju *Obdobje čistih medijev je že za nami*, časnik *Delo*, 13. december 2007.

Naslednji navedek je iz slavne monografije Gena Youngblooda *Expanded Cinema*.

> Samonapajanje, samoupodabljanje, nadzor nad okolico v televizijski instalaciji zaprtega krogotoka so za nekatere umetnike sredstvo, da se lahko ukvarjajo s pojavom sporazumevanja in zaznave na zares empiričen način, podoben znanstvenemu eksperimentiranju. Ta pristop do medija je morda edina čista televizijska umetnost, saj je za njeno estetiko osrednja prav teleportacija informacij v zakodiranem elektronskem signalu (Youngblood, 1970, str. 337).

Vilém Flusser je video razumel kot zrcalo s spominom (Flusser, 2009, str. 182–184), spomin tukaj omogoča algoritmični prevod gibanja v animacijo črt, ki predstavljajo meritve plesa, vendar pa materialni zapis ni trajen, ne omogoča ponovne identifikacije posameznice, zato ne gre za obdelavo osebnih podatkov v pomenu SUVP.

V Laboratoriju za računalniški vid, drugi postaji instalacije, se zajameta portret udeleženke in video posnetek njene hoje na tekalni stezi, v bazo projekta se k telefonski številki udeleženke shranijo njen portret, video posnetek hoje in izračunan vektor hoje, ki ga algoritem v nadaljevanju projekta primerja z vektorji hoje drugih udeležencev v bazi (Slika 3). Drugi laboratorij, ki ustreza *koincidenci*, temelji na Draganovem prepričanju, da je hoja za posameznika karakteristična in edinstvena, izraža torej osebne podatke, po katerih osebo lahko prepoznamo, kar se sklada tudi z uporabo računalniškega vida v nadzornih sistemih za prepoznavanje posameznikov po hoji. Na tem mestu velja dodati, da so mnogi zgodnji novomedijski umetniški projekti, ki raziskujejo tehnologije v njihovem nastajanju, razviti predvsem na konceptualni, vizualni in performativni ravni, medtem ko tehnični vidik interakcije deluje slabo in nezanesljivo, kar pravzaprav pomeni, da so npr. njihove biometrične meritve bolj deklarativne kot pa zares uporabne za prepoznavanje oseb. V rekonstruirani verziji projekta *Matrica koincidenca zmenkarski klub* smo uporabili nevronsko mrežo za analizo hoje, ki je z metodo podpornih vektorjev dosegla 80% natančnost prepoznave hoje; zbirko testnih video posnetkov hoje na tekalni stezi smo zgradili s pomočjo študentov, ki so bili pripravljeni sodelovati v projektu.[21]

Slika 3: Srečo Dragan, *Matrica koincidenca zmenkarski klub*, 2022, MSUM, Ljubljana, 19. 1. 2023, foto: Dejan Habicht.

V tretjem, tj. Laboratoriju za mobilno komunikacijo se na prvem monitorju predvajata portreta dveh udeležencev, ki ju je sistem izbral iz baze na podlagi najbolj podobne hoje, na drugem monitorju pa se predvajata video posnetka njune hoje. Udeleženca dobita na svoja mobilna telefona sporočilo, da se lahko ob določeni uri srečata v določeni kavarni, to sporočilo pošlje izbrani ponudnik storitve za pošiljanje SMS sporočil iz baze, udeleženca pa ne vidita telefonske številke drug od drugega.

[21]Rekonstrukcijo laboratorijev za računalniški vid in mobilno telekomunikacijo iz projekta *Matrica koincidenca zmenkarski klub* je v okviru svoje diplomske naloge izdelal Miha Godec, vizualizacijo povezav med dvema video posnetkoma hoje je sprogramirala Mila Marinković na Fakulteti za računalništvo in informatiko Univerze v Ljubljani pod mentorstvom Narvike Bovcon in Boruta Batagelja.

V projektu se osebni podatki posredujejo na sledeč način: telefonska številka se ne razkrije, portreti in video posnetki hoje udeležencev se predvajajo v galerijskem prostoru na monitorju, analiza posamezničine hoje, izražena z vektorjem, se ne razkrije. Pri rekonstrukciji projekta smo sledili sedmim načelom SUVP, obiskovalko ob vstopu v galerijo seznanimo z namenom in načini obdelave osebnih podatkov v tehno performansu ter pridobimo njeno privolitev (dopolnili smo obrazec s privolitvijo za vsak posamezen namen obdelave osebnih podatkov), upoštevali smo tudi pravice oseb v zvezi z obdelavo njihovih osebnih podatkov, tako glede seznanjanja kot glede preglednosti. Vsi našteti osebni podatki se hranijo v bazi projekta omejen čas, ki sovpada s trajanjem razstave oz. ga je mogoče v bazi nastaviti. Osebne podatke je mogoče iz baze kadarkoli izbrisati na zahtevo udeleženke ali pa ji posredovati digitalno kopijo. Za take postopke je odgovoren upravljavec projekta, tj. umetnica, ki načrtuje projekt skladno s SUVP, in kustosinja, ki med trajanjem razstave skrbi za ustrezno izvajanje projekta, kar vključuje tudi obdelovanje in varovanje podatkov v okviru projekta.

V izvorni verziji projekta leta 2005 je prostor kavarne, v kateri so potekala srečanja udeležencev (takrat je bila to kavarna Minimal v Ljubljani), povezoval z galerijskim prostorom (Galerijo Spomeniškovarstvenega centra) obojesmerni video prenos na video projekcijah, ki je potekal prek spletnih kamer; tudi v kavarni so bili gostje ob vstopu obveščeni, da poteka realnočasovni video prenos v galerijo, ki pa se ni shranjeval. (Gre torej za situacijo teleprezence pri snemanju javnega prostora, kot jo obravnava Floridi, ki pa je simetrično dvosmerna.) V rekonstruirani verziji leta 2022 se za video prenos dogajanja nismo odločili, saj je po dveh letih epidemijske telekonferenčne prisotnosti imel povsem drugačne konotacije kot leta 2005, ko je šele postajal dosegljiv za širše kroge uporabnikov.

Ob rekonstrukciji novomedijskega umetniškega dela *Matrica koincidenca zmenkarski klub* za muzejsko zbirko[22] smo pripravili tudi verzijo tehno performansa za otroke, mlajše od 15 let, ki deluje ločeno od baze projekta in

[22]Projekt je bil izveden v okvirih Laboratorija za nove medije v Moderni galeriji, ki je kot raziskovalna organizacija partnerica v raziskovalnem projektu in hkrati kot državni muzej skrbnica nacionalne zbirke medijske umetnosti. Tako Ćosićev kot Draganov projekt sta bila odkupljena za državno zbirko in predstavljena v galerijskih prostorih.

se izvaja kot enkraten obisk šolskega razreda. Upravljavka instalacije v muzeju ob obisku šole izbere drugo nastavitev instalacije na računalniku. V tem primeru telefonske številke ne zahtevamo, projekt tudi ne vsebuje Laboratorija za mobilno komunikacijo. Fotografije in video posnetki hoje udeležencev se začasno shranijo v ločeno bazo, kar omogoča iskanje in prikaz parov otrok, ki hodijo na podoben način, omogoča torej izkušnjo tehno performansa, baza pa se po zaključku obiska izbriše. Šola mora pred obiskom razstave od staršev pridobiti dovoljenje za udeležbo njihovih otrok.

Vendar pa zajem biometričnih osebnih podatkov predstavlja le eno smer delovanja vmesnika, tj. smer, kjer informacije prehajajo od človeka proti digitalnemu dvojniku. Komunikacija poteka tudi v obratni smeri, od digitalnih vsebin k refiguraciji človekove osebnosti, ki se zgodi v aktu izjave. Srečo Dragan je avtor, ki v svojih novomedijskih instalacijah pogosto tematizira prav osebno izjavo udeleženke, ki opisuje svoje doživljanje v projektu ponujene konstelacije podob, barv, besed, okusov, konceptov, gibov, zvokov.[23] Izjave udeležencev shrani v projektnih podatkovnih zbirkah kot del in rezultat projekta, kot zapis socializacije udeleženke in dokument realizacije projekta v družbeni strukturi. Konstrukcija posameznice na informacijski ravni je pravzaprav medij projekta, soočenje npr. z likovno osebno izkušnjo poseže v identiteto uporabnice, ki se s tem kot informacijski organizem spremeni, nadgradi, opolnomoči. Osebni podatki tako vstopajo v zasnovo Draganovih projektov na dva načina: so točka, v kateri udeleženka sebe predstavlja in konstituira.

[23]Na festivalu Pixxelpoint leta 2007 je Dragan predstavil trilogijo projektov *Metaforične razširitve*, *Metaforične preslikave*, *Metamorfoza lingvistika*, http://www.pixxelpoint.org/2007/icns/katalog.pdf. Na pregledni razstavi *Prostor, vržen iz tira* v MSUM Ljubljana (2016–2017) je bilo mogoče videti še druge projekte te vrste: *Sonifikacija podobe*, *Pametna miza*, *Interaktivna tabla*, *E-knjižni nomad*, https://www.mg-lj.si/si/razstave/1905/sreco-dragan-prostor-vrzen-iz-tira. Prim. tudi Srečo Dragan, Narvika Bovcon, Borut Batagelj, Kristjan Kostanjšek, Luka Žontar, *Časovni tunel 2*, predstavljen na Mednarodnem festivalu novomedijske kulture Speculum Artium, Trbovlje, 20.–23. 9. 2023, https://speculumartium.si/wp21/wp-content/uploads/2024/01/Katalog_Speculum_Artium_2023_web.pdf.

6 SKLEP

V sodobni informacijski družbi imamo nenehno opravka s podatki – tudi sami smo informacijski organizmi v interakciji z drugimi naravnimi, umetnimi in hibridnimi informacijskimi agenti. V teh izmenjavah v infosferi želimo ohraniti svojo informacijsko celovitost, ki je z vidika etike informacij pravzaprav sinonim za našo osebnost. Pri varovanju zasebnosti in osebnih podatkov nam pomaga zakonodaja, predvsem SUVP. Razvoj tehnologije je sicer hitrejši od zakonodaje, vendar pa vprašanje zasebnosti v novih informacijskih medijih Evropska zakonodaja ureja na zgleden način.

Novomedijska umetnost je bila v preteklega pol stoletja neke vrste kreativni laboratorij, kjer so se načrtovali in eksperimentalno preverjali novi načini algoritemskega mišljenja in z informacijskimi tehnologijami podprte mrežne oblike sporazumevanja. Ker je zgodovina novomedijske umetnosti pravzaprav vzporedna zgodovini razvoja novih informacijsko-komunikacijskih tehnologij, jo lahko uporabimo na arheološki in primerjalni način kot podlago za razmislek o učinkih tehnologije in medijev, tudi tistih oblik, ki so se vključile v vsakodnevno rabo. Novomedijski umetniški projekti so prek svoje tehnološko-komunikacijske komponente zasidrani v času svojega nastanka, v kasnejših rekonstrukcijah pa delujejo dvojno: kot pričevanje o obravnavanih aktualnih (tehnoloških in družbenih) temah v času prve izvedbe in kot razmislek o razmerju do obravnavanih (tehnoloških in družbenih) tem v času druge izvedbe oz. rekonstrukcije, ko se je tehnologija že bistveno spremenila in (morda) zamenjala. Takrat moramo biti pozorni na vse plasti, ki sestavljajo novomedijsko umetniško delo, še posebej na tiste, ki so se spremenile zaradi njegove odprtosti, postopkovnosti in kontekstualnosti.

Ena izmed plasti novomedijskega umetniškega dela, ki je z razvojem informacijskih tehnologij v zadnjih dveh desetletjih doživela večje spremembe, je gotovo obdelava osebnih podatkov uporabnikov, saj se ti z današnjimi algoritmi z lahkoto analizirajo – tako da postajajo posamezniki določljivi – in nato dobičkonosno prodajajo, o neetičnosti tovrstnega početja postaja družba vedno bolj ozaveščena, zakonodaja pa področje vedno bolje regulira.

V besedilu smo obravnavali dve študiji primera rekonstrukcije novomedijskih umetniških projektov z vidika načinov obdelave osebnih podatkov: predstavili smo vlogo osebnih podatkov v prvi izvedbi in spremenjeno obdelavo osebnih podatkov v rekonstruirani verziji, tj. načrtovano v skladu z aktualno zakonodajo, na novi tehnološki platformi in v muzejskem kontekstu za različne vrste uporabnikov. S teoretskim pregledom področja in nato z razčlenjenim prikazom uporabe zakonodaje o varovanju osebnih podatkov in algoritmih umetne inteligence smo pokazali smer, kako najti odgovore na številna odprta vprašanja o digitalizaciji sveta, ki imajo pomembno mesto v sodobni družbi, umetnost pa pri tem ni izvzeta.

ZAHVALA

Besedilo je nastalo v okvirih interdisciplinarnega raziskovalnega projekta *Trajnostna digitalna hramba slovenske novomedijske umetnosti* (J7-3158), ki ga financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije in poteka na Univerzi v Novi Gorici, Fakulteti za računalništvo in informatiko Univerze v Ljubljani, Inštitutu za intelektualno lastnino v Ljubljani in v Moderni galeriji v Ljubljani.

LITERATURA

ARNES (Ur.) (2019). Varovanje zasebnosti uporabnikov storitev omrežja ARNES. Pridobljeno 25. maja 2024, https://www.arnes.si/zavod-arnes/katalog-informacij/varovanje-zasebnosti-uporabnikov-storitev-omrezja-arnes

Angwin, J., Larson, J., Mattu, S. in Kirchner L. (2016). Machine Bias. *ProPublica*. Pridobljeno 25. maja 2024, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bovcon. N. (2009). *Umetnost v svetu pametnih strojev: Novomedijska umetnost, Sreča Dragana, Jake Železnikarja in Marka Peljhana.* Raziskovalni inštitut Akademije za likovno umetnost in oblikovanje v Ljubljani.

Ćosić, V. (2022a). Nacija – Kultura. *MG+MSUM, Mrežni muzej*. MG-2022-PRED-738, 11. januarja 2022. Pridobljeno 25. maja 2024, https://mrezni-muzej.mg-lj.si/si/mreznimuzej/3/22/?artworkid=1789

Ćosić, V. (2022b). *Nacija – Kultura: Venec vencev vencev 1*. LUD Šerpa.

Ćosić, V. (2022c). Razstava | Vuk Ćosić: Nacija – Kultura. 13. oktobra 2022–8. februarja 2023 *MG+MSUM.* Pridobljeno, 25. maja 2024, http://www.mg-lj.si/si/razstave/3567/razstava-nacija-kultura

Dragan, S. (2007). Obdobje čistih medijev je že za nami. *Delo*, 13. decembra 2007. Pridobljeno 25. maja 2024, https://old.delo.si/kultura/obdobje-cistih-medijev-je-ze-za-nami.html

Dragan, S., Bovcon, N., Batagelj, B., Kostanjšek, K., Žontar, L. (2023). Časovni tunel 2. *Mednarodni festival novomedijske kulture Speculum Artium, Trbovlje, 20.–23. septembra 2023*. Pridobljeno 25. maja 2024, https://speculumartium.si/wp21/wp-content/uploads/2024/01/Katalog_Speculum_Artium_2023_web.pdf

Dragan, S (2023). Matrica koincidenca zmenkarski klub: Posvet o ohranjanju novomedijske kulturne dediščine – primer rekonstrukcije tehno performansa Sreča Dragana *Matrica koincidenca zmenkarski klub* iz leta 2006. Muzej sodobne umetnosti Metelkova, 19. januarja 2023. *MG+MSUM.* Pridobljeno 25. maja 2024, http://www.mg-lj.si/si/dogodki/3664/posvet

Dragan, S. (2016). Razstava | Prostor, vržen iz tira. Muzej sodobne umetnosti Metelkova, 22. decembra 2016–5. marca 2017. *MG+MSUM.* Pridobljeno 25. maja 2024, https://www.mg-lj.si/si/razstave/1905/sreco-dragan-prostor-vrzen-iz-tira

Dragan, S. (2007). Metaforične razširitve, Metaforične preslikave, Metamorfoza lingvistika. V N. Bovcon in A. Vaupotič (Ur.), *Pixxelpoint, 9. mednarodni festival novomedijske umetnosti, 7.–15. decembra 2007, Nova Gorica, Gorizia*. Mestna galerija Nova Gorica. Pridobljeno 25. maja 2024, http://www.pixxelpoint.org/2007/icns/katalog.pdf

DSB (Austria) – 2021-0.518.795 (2021). *GDPR hub*. Pridobljeno, 25. maja 2024, https://gdprhub.eu/index.php?title=DSB_(Austria)_-_2021-0.518.795

European Commission (2024). Commission Opens Non-Compliance Investigations against Alphabet, Apple, and Meta under the Digital Markets Act. Pridobljeno, 25. maja 2024, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1689

European Commission. Shaping Europe's Digital Future. Pridobljeno 25. maja 2024, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Evropska konvencija o varstvu človekovih pravic. Pridobljeno, 25. maja 2024, https://www.gov.si/assets/ministrstva/MK/Zakonodaja-ki-ni-na-PISRS/Kulturna-raznolikost/00149ee27a/Evropska-konvencija-o-varstvu-clovekovih-pravic.pdf

Evropski svet. Svet Evropske unije (2021). Kar je nezakonito zunaj spleta, bi moralo biti nezakonito tudi na spletu: svet dosegel dogovor o stališču glede akta o digitalnih storitvah. Pridobljeno, 25. maja 2024. https://www.consilium.europa.eu/sl/press/press-releases/2021/11/25/what-is-illegal-offline-should-be-illegal-online-council-agrees-on-position-on-the-digital-services-act

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities.* Oxford University Press.

Floridi, L. (2013). *The Ethics of Information.* Oxford University Press.

Flusser, V. (2009). *Kommunikologie weiter denken: Die Bochumer Vorlesungen*. Fischer.

GDPR.EU (Ur.). *Complete Guide to GDPR Compliance*. Pridobljeno, 25. maja 2024, https://gdpr.eu

GDPR.EU (Ur.). Cookies, the GDPR, and the ePrivacy Directive. Pridobljeno, 25. maja 2024, https://gdpr.eu/cookies

Hornuf, L., Mangold, S. (2022). Digital Dark Nudges. V L. Hornuf (Ur.), *Diginomics Research Perspectives* (str. 89–104). Springer.

Informacijski pooblaščenec RS (2021). Objava javno objavljenih podatkov. Pridobljeno, 25. maja 2024, https://www.ip-rs.si/mnenja-gdpr/objava-javno-objavljenih-podatkov

Informacijski pooblaščenec RS (2020). Fotografije kot osebni podatek. Pridobljeno, 25. maja 2024, https://www.ip-rs.si/mnenja-gdpr/6048a64130f43

Informacijski pooblaščenec RS (2018). Evropski okvir za varstvo podatkov: Najpogostejša vprašanja in odgovori. Pridobljeno, 25. maja 2024, https://www.ip-rs.si/zakonodaja/reforma-evropskega-zakonodajnega-okvira-za-varstvo-osebnih-podatkov/najpogostejsa-vprasanja-in-odgovori

Koops, B.-J. (2021). The concept of function creep. *Law, Innovation and Technology* 13(1), 29–56.

Ministrstvo za digitalno preobrazbo (2024). Zaključena prva obravnava predloga Zakona o izvajanju uredbe EU o enotnem trgu digitalnih storitev. *Gov.si.* Pridobljeno 25. maja 2024, https://www.gov.si/novice/2024-02-19-zakljucena-prva-obravnava-predloga-zakona-o-izvajanju-uredbe-eu-o-enotnem-trgu-digitalnih-storitev

Shaffer, G. C. in Pollack, M. A. (2010). Hard Vs. Soft Law: Alternatives, Complements, and Antagonists in International Governance. *Minnesota Law Review* 49, 706–799. Pridobljeno 25. maja 2024, https://scholarship.law.umn.edu/mlr/491

Služba Vlade Republike Slovenije za digitalno preobrazbo (2022). V veljavo stopil Akt o digitalnih trgih. *Gov.si.* Pridobljeno 25. maja 2024, https://www.gov.si/novice/2022-11-04-v-veljavo-stopil-akt-o-digitalnih-trgih

Splošna uredba o varstvu podatkov (SUVP/GDPR). Pridobljeno, 25. maja 2024, http://data.europa.eu/eli/reg/2016/679/oj

Strokovna skupina na visoki ravni za umetno inteligenco pri Evropski komisiji, AI HLEG (2019). *Etične smernice za zaupanja vredno umetno inteligenco*. Evropska komisija. Pridobljeno 25. maja 2024, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60438

Strokovna skupina na visoki ravni za umetno inteligenco pri Evropski komisiji, AI HLEG (2019). *Opredelitev umetne inteligence: Glavne zmogljivosti in znanstvene discipline*. Evropska komisija. Pridobljeno 25. maja 2024, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60667

Tom Gerken & Zoe Kleinman (2024). Apple, Meta and Google to be investigated by the EU. *BBC News*, 25. marca 2024. Pridobljeno, 25. maja 2024. https://www.bbc.com/news/technology-68655093

Vaupotič, A. (2019). *Vprašanje realizma*. Založba Univerze v Novi Gorici.

Vaupotič, A., Polonyi, E., Bovcon, N. in Železnikar, J. (2023). Sustainable digital preservation of the new media art. V T. C. W. Wong & W. van der Plas (Ur.), *Proceedings Third Summit on New Media Art Archiving: On the occasion of the 28th International Symposium on Electronic Art, ISEA2023 Paris (May 19-20, Paris, France)*. Pridobljeno 25. maja 2024, https://isea-archives.org/docs/2023/3rd_SNMAA_Provisional_Proceedings.pdf

Youngblood, G. (1970). *Expanded Cinema*. Dutton.

Zakon o elektronskih komunikacijah (ZEKom-2). Pridobljeno, 25. maja 2024, https://pisrs.si/pregledPredpisa?id=ZAKO8611

Zakon o varstvu osebnih podatkov (ZVOP). Pridobljeno, 25. maja 2024, https://pisrs.si/pregledPredpisa?id=ZAKO1213

Zakon o varstvu osebnih podatkov (ZVOP-2). Pridobljeno, 25. maja 2024, https://pisrs.si/Pis.web/pregledPredpisa?id=ZAKO7959

PERSONAL DATA IN ART: ITS LEGAL PROCESSING AND THE ROLE OF ETHICS IN NEW MEDIA CULTURE

The text discusses the issues of legal and ethical processing of personal data in new media culture and art, as well as the dilemmas of including such cultural artifacts in exhibition projects and museum collections. It introduces the General Data Protection Regulation (2018), which was adopted by the European Parliament and the Council of the European Union, as well as the current legislative procedures of the European Union in the field of artificial intelligence. It sheds light on the issue of personal data protection through reflection on information ethics, as developed by Luciano Floridi, and explains the relationship between hard and soft ethics, the latter proving to be particularly useful for understanding the role of new information technologies in today's society. The contribution illustrates the role of personal data in art projects, also those involving the artificial intelligence algorithms, and the dilemmas associated with such data using the examples of new media artworks by Vuk Ćosić and Srečo Dragan.

Keywords: General Data Protection Regulation, Vuk Ćosić, Srečo Dragan, soft ethics, new media art

IZBOR IN UREJANJE GRADIV ZA UČNI KORPUS GOVORJENE SLOVENŠČINE ROG

Darinka VERDONIK,[1] Nikola LJUBEŠIĆ,[2] Peter RUPNIK,[2] Kaja DOBROVOLJC,[3] Jaka ČIBEJ[3]

[1] Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

[2] Institut Jožef Stefan

[3] Univerza v Ljubljani, Filozofska fakulteta

Učni korpusi vsebujejo premišljen nabor gradiv in zanesljive, praviloma ročno pripisane oznake na različnih jezikoslovnih ravneh. Služijo tako za učenje avtomatskih označevalnikov kot za temeljne jezikoslovne raziskave. Učni korpus za slovenščino SUK vsebuje samo pisna besedila, medtem ko je bilo stanje za govorjeni jezik zelo fragmentarno. Prispevek predstavlja izbor in urejanje gradiv za učni govorni korpus slovenščine ROG. Korpus obsega približno 75.000 besed oziroma preračunano okrog 8 do 9 ur govora. Gradiva za ROG so bila izbrana iz aktualne različice korpusa Gos 2.1 in ga delimo v tri podenote. V prispevku so podrobno opisane sestavne enote korpusa ROG skupaj s številčnimi podatki o gradivih. Korpus bo predvidoma do konca leta 2024 objavljen v repizitoriju CLARIN.SI pod licenco Creative Commons.

Ključne besede: govorni viri, govor, učni korpus

1 UVOD

Z razvojem tehnologije in umetne inteligence je poudarek na razumevanju in obdelavi govorjenega jezika vse večji. Eden ključnih korakov za razvoj tovrstnih tehnologij je učni korpus, tj. »premišljeno grajene besedilne množice z zanesljivimi (tipično ročno pripisanimi ali pregledanimi) dodatnimi informacijami, ki se uporabljajo pri nadzorovanem strojnem učenju postopkov za obdelavo naravnega jezika« (Arhar Holdt in sod., 2023, str. 121). Za slovenski jezik se učni korpus, ki vsebuje pisna besedila, razvija že več kot desetletje (Krek in sod., 2020). Najnovejšo izdajo pod imenom SUK je doživel v 2022 (Arhar Holdt in sod., 2022) in obsega 1 mio. pojavnic. Vključuje ročno pregledane jezikoslovne oznake na naslednjih ravneh: tokenizacija, stavčna segmentacija, lematizacija, oblikoskladnja MULTEXT-East, oblikoslovje ter skladnja Universal Dependencies, skladnja JOS-SYN, udeleženske vloge, imenske entitete in koreference.

Korpus SUK vključuje besedila iz referenčnega pisnega korpusa Gigafida, iz slovenskih novičarskih portalov, iz Wikipedijinih člankov, skratka, iz pisnih virov. Številne raziskave so potrdile pomembne razlike med pisno in govorjeno rabo jezika tako na leksikalni kot slovnični in drugih jezikoslovnih ravneh (Akinnaso, 1982; Henrichsen in Allwood, 2005; Adolphs in Carter, 2003; Biber, 2012; Dobrovoljc in Nivre, 2016). Jezikovni viri, ki vključujejo samo pisno rabo jezika, zato ne morejo biti zadostni za obravnavo govorjenega jezika (Siepmann, 2015; Verdonik in Sepesy Maučec, 2017), ampak so za uspešno procesiranje kot tudi za celostno razumevanje jezikovne rabe potrebni tudi viri, ki vključujejo avtentične primere govorjene rabe. S tem namenom ne samo za slovenščino (Verdonik in sod., 2024), ampak za mnoge jezike nastajajo t. i. govorni korpusi, ki vključujejo posnetke in zapise govora (npr. Komrsková in sod., 2023; Kuvač in Hržica, 2016; Schmidt, 2016; Love in sod., 2017). Nezapolnjena vrzel pa ostajajo govorni korpusi, ki bi imeli ročno pripisane oz. popravljene jezikoslovne oznake. Za slovenščino je bila edini tovrstni vir doslej drevesnica govorjene slovenščine Spoken Slovenian Treebank (SST) v obsegu 30.000 pojavnic, ki ima ročno pripisane leme, oblikoskladenjske oznake MULTEXT-East ter oblikoslovne in odvisnostne skladenjske oznake po sistemu Universal Dependencies (Dobrovoljc in Nivre, 2016).[1] V primerjavi s pisnimi viri tako beležimo precejšnjo vrzel.

S ciljem zapolniti to vrzel v projektu Temeljne raziskave za razvoj govornih virov in tehnologij – MEZZANINE poteka aktivnost izdelave učnega korpusa govorjene slovenščine, poimenovanega ROG (Ročno označeni govorni korpus), ki bo imel ročno pripisane oznake na naslednjih ravneh: tokenizacija, lematizacija, oblikoskladnja po sistemu MULTEXT-East v6,[2] skladnja Universal Dependencies, prozodične enote, netekočnosti in dialoška dejanja. Ta prispevek opisuje prvi korak izdelave tega korpusa, to je izbor gradiv in pripravo gradiv za označevanje. Označevanje je v času pisanja članka še potekalo in v tem prispevku ni obravnavano.

[1] Korpus Spoken Slovenian Treebank je bil v času pisanja tega prispevka objavljen z razširjenim gradivom, ki je opisano v tem prispevku.

[2] Oblikoskladenjske oznake po sistemu MULTEXT-East v6: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

2 IZBOR GRADIV

Za učni korpus je pomembno, da zajema raznovrstne vzorce in primere jezikovne rabe in s tem omogoči dovolj raznoliko množico učnih podatkov. Hkrati so ročno označeni učni podatki omejeni z razpoložljivimi sredstvi in časom ter posledično majhni, zato je potrebna premišljena sestava gradiv, ki jih vključimo v korpus. V govornih korpusih se vse od govorne komponente korpusa British National Corpus (Crowdy, 1993) praviloma upoštevata dve vrsti kriterijev za zajem gradiv: besedilnovrstni in demografski. Obe vrsti kriterijev smo upoštevali tudi ob sestavi učnega korpusa ROG, pri čemer pa smo se omejili z obstoječimi gradivi, saj je bil cilj osredotočiti se na označevanje.

Gradiva smo zajeli iz najnovejše izdaje referenčnega govornega korpusa Gos 2.1 (Verdonik in sod., 2024), ki že vključuje smiseln nabor ob času nastanka razpoložljivih govornih virov. Pri tem smo upoštevali že izdelani in skladenjsko ročno označeni nabor gradiv iz korpusa Gos 1.1 (Verdonik in sod., 2013). Zaradi potreb procesiranja in analiziranja na akustični ravni je bil cilj, da je vsaj polovica učnega korpusa javno dostopna tudi s kvalitetnimi avdio posnetki. Glede na navedene cilje korpus ROG vključuje tri podenote:

1. Izbor iz tistega dela Gos 2.1, ki izhaja iz korpusa Artur, kjer so posnetki kvalitetni in javno dostopni.

2. Izbor iz stare različice korpusa Gos 1.1 z manj kvalitetnimi in za dostop omejenimi posnetki, med katerimi pa mnogi vključujejo bolj interaktivne, bolj spontane in bolj raznolike nejavne govorne situacije kot korpus Artur. Ta del vključuje že obstoječo podenoto 30.000 pojavnic (Dobrovoljc in Nivre, 2016) in dodatno enoto na novo izbranih 10.000 pojavnic.

V nadaljevanju poglavja so vse tri podenote opisane po kriterijih za izbor gradiv in s številčnimi podatki.

2.1 Izbor iz korpusa Artur

Iz korpusa Artur je bil cilj izbrati gradiva v obsegu 40.000 pojavnic. Izbor gradiv iz korpusa Artur je potekal v dveh korakih: (1) ročni izbor posnetkov, (2) avtomatski izbor odsekov posnetkov. Kriteriji za ročni izbor posnetkov iz korpusa Artur:

1. V posnetku ni presluha in prisotnosti šuma.
2. Uravnotežen nabor različnih tipov glede na opis govornega dogodka v korpusu.
3. Uravnoteženost po spolu govorcev.
4. Uravnoteženost po regiji stalnega bivališča govorcev.

Avtomatski izbor segmentov v posnetkih je sledil kriterijem:

1. Izbrani odsek posnetka vključuje okvirno med 700 in 800 besed.
2. Pri posnetkih javnega govora se izbere odsek, v katerem govori predhodno ročno izbrani govorec.
3. Pri nejavnem govoru, kjer je isti pogovor posnet v dveh ločenih posnetkih, vsak za enega govorca, se izbereta oba posnetka pogovora in v obeh posnetkih isti odsek pogovora.
4. Začetek izbranega odseka posnetka je pri menjavi vloge.
5. Konec izbranega odseka je konec segmenta, ki si konča s piko.

Na podlagi tako definiranih kriterijev so bila izbrana gradiva tako v formatu TEI kot v formatu TRS,[3] prav tako so bili pripravljeni tudi avdio posnetki na način, da je bil celoten neizbrani del posnetka utišan. Posnetki so tako ostali enako dolgi kot izvorno, slišen pa je samo odsek, izbran za učni korpus, kar je pomembno za ohranitev obstoječe segmentacije na časovne enote glede na izvorni posnetek. Tabeli 1 in 2 predstavljata podatke o izbranih gradivih.

Tabela 1: Izbor iz korpusa Artur – posnetki.

Tip diskurza	*Govorni dogodek*	*Število posnetkov*	*Število besed*[4]
Javni		23	19.165
	Spletni dogodek	7	5792
	Okrogla miza	6	5039
	Intervju	6	5123
	Novinarska konferenca	3	2479
	Nagovor na dogodku	1	732

[3] Format TRS je izhodni format programa Transcriber, s katerim so bile narejene transkripcije za korpus Gos. Gre za XML-format, ki je enostaven za razčlenjevanje ter uvozljiv v pomembnejša orodja za označevanje in analizo govora (Praat, ELAN, EXMARaLDA).

[4] V izvornih datotekah korpusa Artur.

Nejavni		28	15.533
	Prosti dialog med dvema sogovornikoma	14	5771
	Prosti monološki govor	7	4799
	Razlaganje in opisovanje	7	4963
Parlamentarni	Seja državnega zbora	6	4241
SKUPAJ		57	38.939

Kot vidimo iz tabele 1, obsega izbor 57 približno enako dolgih posnetkov. Povprečen obseg enega pogovora je 683 besed. Razmerje med javnim, nejavnim in parlamentarnim je 50 % javni diskurz, 40 % nejavni diskurz in 10 % parlamentarni diskurz.

Tabela 2: Izbor iz korpusa Artur – govorci.

Značilnosti govorca	*Vrsta*	*Število govorcev*	*Število besed*
Spol	Moški	38	20.073
	Ženski	34	18.866
Starost	18 do 34	11	6785
	30 do 59	36	18.381
	nad 60	13	6561
	Nedoločeno	12	7212
Statistična regija*	Osrednjeslovenska	18	10.319
	Podravska	10	4566
	Savinjska	6	3149
	Pomurska	3	1584
	Goriška	5	2591
	Gorenjska	2	1459
	Jugovzhodna	2	870
	Koroška	2	676
	Primorsko-notranjska	3	995
	Posavska	4	1847
	Nedoločeno	17	10.883
SKUPAJ		72	38.939

* Oznaka pomeni statistično regijo stalnega bivališča.

Tabela 2 predstavlja podatke za celoten izbor. Odstotek nedoločenih podatkov je dokaj visok, ker za javni govor pogosto ni vseh podatkov o govorcih. Povprečno število besed na enega govorca je 540. Število oseb moškega in ženskega spola je približno enakomerno, zastopane so vse starostne skupine in skoraj vse statistične regije.

2.2 Izbor iz korpusa Gos 1.1

Izbor iz korpusa Gos 1.1 vsebuje že obstoječi korpus SST, razširjen z dodatnimi 10.000 pojavnicami prav tako iz korpusa Gos 1.1. Obe podenoti ostajata v ločenih datotekah in ju opisujemo v ločenih podpoglavjih.

2.2.1 ŽE OBSTOJEČI IZBOR V OBSEGU 30.000 BESED

V raziskavi (Dobrovoljc in Nivre, 2016) je predstavljen nabor gradiv iz korpusa Gos 1.1 v skupnem obsegu 30.000 pojavnic, tj. korpus Spoken Slovenian Treebank – SST. Gradiva so bila izbrana tako, da se je iz vsakega posnetka od 287 posnetkov v korpusu Gos 1.1 izbral proporcionalen del pojavnic. Vsak zajeti odsek posnetkov vključuje eno ali več zaporednih vlog govorcev. Ta nabor je bil uporabljen za dve ročni označevalni kampanji: označevanje večbesednih diskurznofunkcijskih stalnih besednih zvez (Dobrovoljc, 2018) in slovnično označevanje v okviru izdelave prve drevesnice govorjene slovenščine (Dobrovoljc in Nivre, 2016), v kateri so bili besedilom ročno pripisani podatki o lemah in oblikoskladenjskih oznakah po sistemu JOS/MULTEXT-East ter oblikoslovne in skladenjske oznake po medjezikovno usklajeni označevalni shemi Universal Dependencies. Drevesnica SST je bila prva govorna drevesnica, označena s shemo UD, v slovenskem prostoru pa je bila uporabljena za raziskave in razvoj prilagojenih slovničnih označevalnikov za govor (Dobrovoljc in Martinc, 2018; Verdonik in sod., 2024).

Tabela 3: SST izbor iz korpusa Gos 1.1 – posnetki.

Tip diskurza	*Kanal*	*Število posnetkov*	*Število besed*[5]
Javni informativni/ izobraževalni	Televizija	61	3068
	Radio	27	2310
	Osebni stik	33	3555

[5] V izvornih datotekah XML TEI korpusa Gos 1.1.

Javni razvedrilni	Televizija	14	2792
	Radio	27	3439
Nejavni nezasebni	Osebni stik	26	3007
	Telefon	17	1018
Zasebni	Osebni stik	49	5580
	Telefon	18	1562
SKUPAJ		272	26.331

Ker vključuje korpus SST del vsakega posnetka v izvornem korpusu Gos 1.1, je število datotek dokaj veliko, 272. Povprečen obseg enega (po)govora je 97 besed, tako da je korpus precej fragmentiran.

2.2.2 DODATEN IZBOR V OBSEGU 10.000 BESED

S ciljema, da se tudi iz korpusa Gos 1.1 zajamejo gradiva v obsegu 40.000 pojavnic in da se učni korpus razširi s posnetki pogovorov oz. medosebne interakcije, je bil korpus SST dopolnjem z dodatnimi 10.000 pojavnicami. Gradiva so bila izbrana tako, da vključujejo posnetke interaktivnih neformalnih govornih situacij, ki v korpusu Artur večinoma niso zajeta, v korpusu Gos 1.1 pa so. Izbor je potekal podobno kot pri korpusu Artur z ročnim izborom posnetkov in avtomatskim izborom odsekov iz posnetkov. Za izbor posnetkov so bili upoštevani kriteriji:

1. javni diskurz:
    a. samo posnetki neformalne interakcije
    b. enakomerna zastopanost posnetkov s televizije in radia
    c. vključijo se vrste govornih dogodkov, ki še niso bile zajete v gradivih iz korpusa Artur
2. nejavni nezasebni diskurz:
    a. vključijo se vrste govornih dogodkov, ki še niso bile zajete v gradivih iz korpusa Artur
    b. raznolikost tematik
    c. 85 % posnetkov v osebnem stiku in 15 % posnetkov po telefonu
3. nejavni zasebni diskurz:

a. vključijo se vrste govornih dogodkov, ki še niso bile zajete v gradivih iz korpusa Artur
b. raznolikost tematik
c. 85 % posnetkov v osebnem stiku in 15 % posnetkov po telefonu
d. zastopanost različnih regij

Hkrati smo v celotnem izboru sledili cilju, da bi bili ustrezno zastopani oba spola in različne starostne skupine.

Izbor odsekov iz posnetkov je bil izveden tako, da se je nadaljeval od tam naprej, kjer se je obstoječi izbor iz iste datoteke nehal. Začetna ocena je bila, da naj bi bili izbrani odseki v vseh datotekah dolgi od 350 do 360 besed, vendar smo za dosego cilja 10.000 pojavnic ta obseg nekoliko povečali.

Tabela 4: Dodaten izbor iz korpusa Gos 1.1 – posnetki.

Tip diskurza	*Kanal*	*Število posnetkov*	*Število besed*[6]
Javni razvedrilni	Televizija	2	817
	Radio	2	874
Nejavni nezasebni	Osebni stik	5	2347
	Telefon	1	425
Zasebni	Osebni stik	10	4301
	Telefon	2	855
SKUPAJ		22	9.619

Dodaten nabor obsega 22 posnetkov govora, kot vidimo iz tabele 4. Vsi posnetki so približno enako dolgi, povprečen obseg govora na enem posnetku pa je 437 besed.

Tabela 5: Dodaten izbor iz korpusa Gos 1.1 – govorci.

Značilnosti govorca	*Vrsta*	*Število govorcev*	*Število besed*
Spol	Moški	24	3624
	Ženski	37	5995
Starost	10 do 18	2	60

[6] V izvornih datotekah XMLTEI korpusa Gos 1.1.

	18 do 34	20	3988
	30 do 59	21	3019
	Nad 60	5	861
	Nedoločeno	13	1691
Statistična regija*	Osrednjeslovenska	14	2546
	Podravska	4	735
	Obalno-kraška	4	589
	Jugovzhodna Slovenija	4	792
	Goriška	3	297
	Pomurska	3	440
	Posavska	2	421
	Gorenjska	1	115
	Savinjska	1	199
	Mešano	15	2229
	Nedoločeno	10	1256
SKUPAJ		61	9619

* Oznaka pomeni vse statistične regije, v katerih je oseba bivala dalj časa.

Tabela 5 predstavlja podatke za celoten izbor. Odstotek nedoločenih podatkov je dokaj visok, ker za javni govor pogosto ni vseh podatkov o govorcih. Povprečno število besed na enega govorca je 157. Več oseb je ženskega spola. Zastopane so vse starostne skupine. Geografsko so pokrite vse večje slovenske regije. Velik delež govorcev ima označeno več kot eno regijo daljšega bivanja in so v tabeli uvrščeni pod kategorijo mešano.

Celoten učni korpus ROG skupaj obsega približno 75.000 besed. Preračunano v čas trajanja posnetkov znaša to skupaj med 8 in 9 ur govora.

3 UREJANJE GRADIV

3.1 Usklajevanje segmentacije

Ker korpus Gos 2.1 združuje posnetke iz različnih virov (Gos 1.1, Gos Videolectures in Artur), naletimo na nekaj razlik tako na ravni metapodatkov o posnetkih in govorcih (glej prispevek Verdonik in sod., 2022) kot na ravni zapisovanja in segmentiranja govora.

Za učni korpus je predstavljala poseben izziv razlika med načeli segmentacije na enote govora v gradivu, ki je vključeno v GOS 1.1, in tistim gradivom, ki je bilo v korpus dodano v različici 2.1. Transkripcije posnetkov so bile namreč v različici 1.1 razdeljene na izjave in segmente, ki so bile obravnavane kot osnovna enota govora, ki približno ustreza pojmu povedi v pisnem jeziku in je določena tako, da je prozodično, semantično in skladenjsko zaokrožena (Verdonik in sod., 2013). Del, ki izhaja iz zbirke Artur (Verdonik in Bizjak, 2023), pa je bil predvsem zaradi potreb razvoja razpoznavalnika govora segmentiran z večjim upoštevanjem prozodičnih kriterijev, tj. s strožjim ločevanjem izjav na segmente glede na premore. Premor, ki je bil dolg vsaj 0,2 sekunde, je bil obravnavan kot mejnik med segmentoma, semantična in skladenjska zaključenost pa je bila šele drugotnega pomena. Razlika v segmentaciji je razvidna iz primerov v tabeli 6. Ker je predstavljala težavo pri označevanju v uporabljenih orodjih, ki so prilagojena za označevanje pisnih besedil, smo se odločili za dodatno usklajevanje segmentacije. Preučitev orodij, prilagojenih označevanju govora, za izvajanje vseh ravni označevanja, tudi skladnje, z vzporednim označevanjem osnovnih enot, ki so lahko različne glede na prozodijo, skladnjo, pragmatiko ali tehnične zahteve in podobno, ostaja eden od izzivov za naprej.

Tabela 6: Razlike v segmentaciji v gradivih iz Gos 1.1 in Artur.

Različica korpusa	*Primer segmentirane transkripcije*
GOS 1.1	<seg>kot vedno naši eem sodni mlini delujejo zelo zelo zelo počasi</seg>
Artur	<seg>Drage prijateljice, dragi prijatelji</seg> <seg>govorjene slovenščine.</seg> <seg>razmišljal sem, kako naj začnem ta</seg> <seg>svoj nastop.</seg>

Segmenti, razdeljeni na podlagi premorov, so pogosto kratki in pri označevanju skladenjskih povezav le-te potekajo preko mej segmentov, kar povzroča težave na nivoju tehnične implementacije pri označevalnih platformah in programski opremi, kot je Q-CAT (Brank, 2021). Pred označevanjem smo tako segmente strojno preporazdelili glede na ločila, ki so bila postavljena med transkripcijo posnetkov. Uporabili smo končna ločila pika (.), vprašaj (?) in tropičje (...) in algoritem resegmentacije je vsakič, ko je naletel na eno od teh ločil, zaključil segment in začel novega. Pri tem so bile ohranjene vse informacije o prvotni delitvi na segmente (tj. identifikacijske kode prvotnih segmentov za vsako pojavnico), kar je omogočilo popolno sledljivost in skladnost z izvornimi podatki, obenem pa je poenotilo reprezentacijo segmentov med tistimi deli, ki so bili vzorčeni iz Gos 1.1, in tistimi iz korpusa Artur. Primer resegmentirane transkripcije prikazuje tabela 7.

Tabela 7: Primer strojno resegmentirane transkripcije iz korpusa Artur.

Izvorna segmentacija	*Resegmentirana transkripcija*
<seg>in tu se navezujem na</seg> <seg>misli eee direktorja</seg> <seg>ZRC Sazu,</seg> <seg>čez mejo</seg> <seg>in še čez eno mejo,</seg> <seg>v Prago.</seg>	<seg>in tu se navezujem na misli eee direktorja ZRC Sazu, čez mejo in še čez eno mejo, v Prago.</seg>

Strojna resegmentacija je število segmentov v vzorcu iz korpusa Artur zmanjšala s 5.587 na 1.968. V povprečju vsak nov segment vsebuje po tri stare segmente (najmanj enega in največ 70), polovica pa vsebuje manj kot dva (s standardnim odklonom 3,1 segmenta). Iz tabele 8 je razvidno, da gradivo iz korpusa Artur vsebuje nekoliko daljše segmente – v povprečju skoraj 25 pojavnic na segment, kar je več kot dvakrat več od gradiva iz Gos 1.1, kjer segment v povprečju zajema okrog 9 pojavnic. To je do določene mere pričakovana razlika, saj je bila v gradivu iz Gos 1.1 ob transkribiranju preferirana krajša dolžina segmenta, prav tako je v korpusu Artur manj interaktivnih in neformalnih govornih situacij, za katere so značilni krajši segmenti kot pri govornih dogodkih, kakršna so predavanja, okrogle mize ipd.

Tabela 8: Statistika resegmentacije vzorca iz korpusa Artur in primerjava z vzorcem iz Gos 1.1.

Statistika	*Vzorec GOS 1.1*	*Vzorec Artur*	*Vzorec Artur z resegmentacijo*
Število segmentov	4.912	5.587	1.968
Povprečno število pojavnic na segment	9,23	8,68	24,71

Mediana števila pojavnic na segment	6,00	7,00	18,00
Največje število pojavnic na segment	120	51	452

3.2 Sledljivost pojavnic

Izbrani segmenti za korpus ROG so bili izločeni iz korpusa Gos 2.1 (Verdonik in sod., 2023) v formatih TRS, TEI in WAV. Sledilo je sočasno označevanje na več nivojih: prva skupina raziskovalcev je izvajala ročno popravljanje oblikosladenjskih oznak in lem, druga skupina skladenjsko označevanje, tretja skupina označevanje netekočnosti in četrta označevanje prozodičnih enot. V uporabi so bili različni programi za označevanje: leme in oblikoskladenjske oznake so se pregledovale v razpredelnicah Google Sheets; odvisnostna skladnja se je označevala v orodju Q-Cat; za označevanje netekočnosti je bilo izbrano orodje EXMARaLDA,[7] ki je namenjeno podpori pri razvoju in označevanju govornih virov in omogoča označevanje na podlagi časovnih značk, ki ohranjajo povezavo s signalom; za označevanje prozodičnih enot je bilo uporabljeno orodje Praat, ki je prilagojeno analizam na akustični ravni.

Cilj je vse različne nivoje ročno pregledanih in dodanih oznak združiti v skupno XML-datoteko z več nivoji oznak. Da bi zagotovili združljivost gradiv, so bile v izbrana gradiva, uvožena v različna orodja in posledično orodjem prilagojene formate, dodane identifikacijske oznake pojavnic iz korpusa Gos2.1 (t. i. xml:id, na primer xml:id="Artur-P-G7036-P701111.tok603"). To omogoča naknadno preverjanje sprememb (denimo popravki ali naključne napake) in olajša prehajanje med različnimi formati.

4 ZAKLJUČEK

V prispevku smo predstavili izbor in urejanje gradiv za nov učni korpus govorjene slovenščine ROG. Korpus obsega skupaj okvirno 75.000 besed oz. po oceni približno 8 do 9 ur govora. Polovica korpusa izhaja iz korpusa Artur, kjer so prednost kvalitetni avdio posnetki, dostopni brez omejitev, polovica pa iz korpusa Gos 1.1, kjer je prednost večja avtentičnost in interaktivnost zlasti

[7] https://exmaralda.org/en/

nejavnega, delno pa tudi nekaterih posnetkov javnega govora. Celoten korpus bo ročno pregledan in označen na ravni lematizacije, oblikosladnje in skladnje, tisti del, ki izhaja iz korpusa Artur in ima na voljo prosto dostopne in kvalitetne avdio posnetke, pa tudi z ročnimi oznakami prozodičnih enot, netekočnosti in dialoških dejanj. Učni korpus ROG bo objavljen v repozitoriju CLARIN.SI, potem ko bodo dodani vsi nivoji oznak in združeni v eno datoteko, pod eno od licenc Creative Commons predvidoma do konca leta 2024.

Načrti za nadaljnje delo vključujejo razširitev korpusa s kvalitetnimi in odprto dostopnimi avdio posnetki pogovornega gradiva in razvijanje dodatnih nivojev oznak, ki omogočajo raziskave in razvoj na področju razumevanja govorne komunikacije.

ZAHVALA

Prispevek je nastal v okviru raziskovalnega projekta *Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik* (MEZZANINE, J7-4642), raziskovalnega projekta *Na drevesnici temelječ pristop k raziskavam govorjene slovenščine* (SPOT, Z6-4617) in raziskovalnega programa *Jezikovni viri in tehnologije za slovenski jezik* (P6-0411), ki jih financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS).

ITERATURA

Arhar Holdt, Š. in sod. (2022). *Training corpus SUK 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1747.

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Erjavec, T., Gantar, P., Krek, S., Munda, T., Robida, N., Terčon L. in Žitnik, S. (2023). Nadgradnja učnega korpusa ssj550k v SUK 1.0. V Š. Arhar Holdt, S. Krek (ur.), *Razvoj slovenščine v digitalnem okolju* (str. 119-156). Ljubljana: Založba Univerze. https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/522/852/9441.

Adolphs, S. in Carter, R. (2003). And she's like it's terrible, like: Spoken discourse, grammar and corpus analysis. *International Journal of English Studies*, 3(1), 45–66.

Akinnaso, F. N. (1982). On the differences between spoken and written language.

Language and Speech, 25(2), 97–125.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37.

Brank, J. (2021). Q-CAT Corpus Annotation Tool 1.2. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1442.

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing*, 8(4), 259–265. doi: 10.1093/llc/8.4.259

Dobrovoljc, K. in Martinc, M. (2018). Er ... well, it matters, right? On the role of data representations in spoken language dependency parsing. V *Proceedings of the workshop*. Second Workshop on Universal Dependencies (UDW 2018), November 1, 2018, Brussels. Strasbourg: Association for Computational Linguistics, 2018. Str. 37-46. https://aclanthology.info/papers/W18-6005/w18-6005.

Dobrovoljc, K. in Joakim, N. (2016). The Universal Dependencies Treebank of Spoken Slovenian. V *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), Portorož, Slovenia.

Dobrovoljc, K. (2018). Raba tipično govorjenih diskurznih označevalcev na spletu. *Slavistična revija: časopis za jezikoslovje in literarne vede* 66(4), str. 497-513. https://srl.si/ojs/srl/article/view/2018-4-1-6.

Henrichsen, P. J. in Allwood, J. (2005). Swedish and Danish, spoken and written language: A statistical comparison. *International Journal of Corpus Linguistics* 10(3), 367–399.

Komrsková, Z., Kopřivová, M., Lukeš, D., Poukarová, P. in Goláňová, H. (2017). New Spoken Corpora of Czech: ORTOFON and DIALEKT. *Journal of Linguistics/Jazykovedný casopis* 68(2), 219–228. https://doi.org/10.1515/jazcas-2017-0031.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J. in Brank, J. (2020). The ssj500k training corpus for Slovene language processing. V D. Fišer in T. Erjavec (Ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (str. 23–33). Ljubljana, Slovenija. Inštitut za novejšo zgodovino.

Kuvač Kraljević, J. in Hržica, G. (2016). Croatian Adult Spoken Language Corpus (HrAL). *FLUMINENSIA* 28(2), 87–102.

Love, R., Dembry, C., Hardie, A., Brezina, V. in McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations.

International Journal of Corpus Linguistics 22(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Schmidt, T. (2016). Construction and dissemination of a corpus of spoken interaction – tools and workflows in the FOLK project. *Journal for Language Technology and Computational Linguistics* 31(1), 127–154.

Siepmann, D. (2015). Dictionaries and spoken language: A corpus-based review of French dictionaries. *International Journal of Lexicography*, 28(2), 139–168.

Verdonik, D. in Bizjak A. (2023). *Pogovorni zapis in označevanje govora v govorni bazi Artur projekta RSDO*. Maribor: Univerzitetna založba.

Verdonik, D. in Sepesy Maučec, M. (2017). A speech corpus as a source of lexical information. *International Journal of Lexicography* 30(2), 143–166. DOI: 10.1093/ijl/ecw004.

Verdonik, D., Kosem, I., Zwitter Vitez, A., Krek, S. in Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation* 47(4), 1031–1048.

Verdonik, D., Bizjak, A., Žgank, A., Dobrišek, S. Metapodatki o posnetkih in govorcih v govornih virih: primer baze Artur. V D. Fišer in T. Erjavec (Ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (str. 206-2012), Ljubljana, Slovenija. Inštitut za novejšo zgodovino.

Verdonik, D., in sod. (2023). *Spoken corpus Gos 2.1 (transcriptions).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1863.

Verdonik, D., Dobrovoljc, K., Erjavec, T. in Ljubešić, N. (2024). Gos 2: A New Reference Corpus of Spoken Slovenian. V *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 7825–7830). Torino, Italia. ELRA and ICCL.

SELECTION AND PREPARATION OF DATA FOR THE ROG 1.0 TRAINING CORPUS OF SPOKEN SLOVENIAN

The article presents the selection and preparation of data for ROG, the training corpus of spoken Slovenian. The corpus comprises approximately 75,000 words, equivalent to around 8 to 9 hours of speech. Materials for the corpus were selected from the current version of the Gos 2.1 corpus and are divided into three subunits. The first subunit consists of data derived from the Artur corpus and encompasses 40,000 tokens. The advantage of these data is the availability of high-quality and freely accessible audio recordings. The other two subunits are selected from the Gos 1.1 corpus: the existing SST spoken language learning corpus, comprising 30,000 tokens, and an additional selection to fill the gap in interactive, conversational data, comprising 10,000 tokens. The data were exported in TEI, TRS, and WAV formats. In the WAV format, unselected parts of the recordings were muted to preserve the alignment with original segment time codes, while in the TEI and TRS formats, unselected parts were removed. The learning corpus will contain manually corrected and added annotations for lemmas, morphosyntax, syntax, prosodic units, disfluencies, and dialog acts, and is expected to be published by the end of 2024 in the CLARIN.SI repository under a Creative Commons license.

Keywords: spoken resources, speech, training corpus

GENERATIVE MODEL FOR LESS-RESOURCED LANGUAGE WITH 1 BILLION PARAMETERS

Domen VREŠ,[1] Martin BOŽIČ,[1] Aljaž POTOČNIK,[2] Tomaž MARTINČIČ,[2] Marko ROBNIK-ŠIKONJA[1]

[1]University of Ljubljana, Faculty of Computer and Information Science
[2]XLAB d.o.o.

Large language models (LLMs) are a basic infrastructure for modern natural language processing. Many commercial and open-source LLMs exist for English, e.g., ChatGPT, Llama, Falcon, and Mistral. As these models are trained on mostly English texts, their fluency and knowledge of low-resource languages and societies are superficial. We present the development of large generative language models for a less-resourced language. GaMS 1B - Generative Model for Slovene with 1 billion parameters was created by continuing pretraining of the existing English OPT model. We developed a new tokenizer adapted to Slovene, Croatian, and English languages and used embedding initialization methods FOCUS and WECHSEL to transfer the embeddings from the English OPT model. We evaluate our models on several classification datasets from the Slovene suite of benchmarks and generative sentence simplification task SENTA. We only used a few-shot in-context learning of our models, which are not yet instruction-tuned. For classification tasks, in this mode, the generative models lag behind the existing Slovene BERT-type models fine-tuned for specific tasks. On a sentence simplification task, the GaMS models achieve comparable or better performance than the GPT-3.5-Turbo model.

Keywords: large language models, generative models, knowledge transfer, OPT model, GaMS model, language adaptation

1 INTRODUCTION

Large language models (LLMs), in particular generative LLMs like GPT models (Brown et al., 2020; OpenAI et al., 2024), have dramatically transformed natural language processing (NLP), advancing the understanding and generation of human language. As a result of this rapid development, new open-source decoder-type transformer LLMs such as Llama, Falcon, Mistral, and many others are released on a monthly basis. These models are trained on high-resource languages (primarily English), leaving many less-resource languages, such as Slovene, behind. In this work, we present the development of GaMS 1B (Generative Model for Slovene), the first Slovene open-source generative model with 1 billion parameters. The aim is to transfer recent advancements in language technologies from English to Slovene and, therefore, improve the technological development of Slovene. We release the model under open-source license. The creation of the model is fairly general and offers useful lessons to other less-resourced languages.

The main problem in training LLMs for Slovene is the lack of data. For example, the Llama 3 model (AI@Meta, 2024) was trained on 15 trillion tokens, while the currently available Slovene corpora contain around 11 billion tokens, a thousand times fewer. This means that training an LLM from scratch for Slovene is unfeasible. Hence, we adapt the already trained English OPT model (Zhang et al., 2022) to Slovene. To increase the amount of available training data, we also include texts from Croatian, Bosnian, and Serbian languages, which can improve the models' performance due to the language similarity. Taking an English model as a starting point raises the problem of the model's vocabulary, as the existing one is not adapted to Slovene, resulting in an inefficient tokenization of Slovene texts (i.e. considerably more tokens are generated compared to efficient tokenization). To solve this problem, we train a new tokenizer and employ embedding initialization methods WECHSEL (Minixhofer et al., 2022) and FOCUS (Dobler & de Melo, 2023) to transfer the embeddings from the English model to ours with the Slovene-tailored vocabulary.

An efficient evaluation of LLMs poses an additional challenge for low-resource languages. We demonstrate that models can not be directly compared based on training/validation losses observed during generative pretraining. The main reason is different vocabularies, as distributions of their output tokens differ, impacting the cross-entropy loss computation. English models are often evaluated on benchmarks testing models' reasoning, language understanding, etc. Such benchmarks are rare in Slovene, and using machine translation on complex datasets is mostly infeasible due to contextual differences between the languages. Hence, additional effort is required to obtain and adapt such benchmarks to a new language. We evaluate our models on three benchmarks already created or adapted to Slovene: the Slovene adaptation of the SuperGLUE benchmark suite (Žagar et al., 2020), the Slovene natural language inference dataset SI-NLI (Klemen et al., 2022), and the sentence simplification dataset SENTA (Žagar et al., 2024).

The paper is organized into six sections. In Section 2, we present related work on the development of large language models and transferring their knowledge to low-resource languages. In Section 3, we present the data used for training of our GaMS model. We offer a detailed technical description of GaMS model, i.e. the training of a new tokenizer, embedding transfer methods, and training details, in Section 4. In Section 5, we evaluate the models. We provide conclusions and directions for further work in Section 6.

2 RELATED WORK

New LLMs (or model families) are released on a monthly basis, with the most notable representatives being LLaMa (Touvron, Lavril, et al., 2023; Touvron, Martin, et al., 2023; AI@Meta, 2024), Falcon (Penedo et al., 2023), Phi (Li et al., 2023), Mistral (Jiang et al., 2023), and Mixtral (Jiang et al., 2024). Most of these models were trained on mainly English texts, and those trained on more languages have seen a very small proportion of Slovene texts compared to more represented languages. Therefore, the performance of these models for Slovene can be improved with additional pretraining on Slovene texts.

To spread the benefits of LLMs to languages other than English, multilingual models were developed. BLOOM (Workshop et al., 2023), YAYI 2 (Luo et al., 2023), PolyLM (Wei et al., 2023) and XGLM (Lin et al., 2022) were all trained on over 15 languages. However, they do not achieve the performance of state-of-the-art English models due to a lower number of parameters or smaller training data size. Additionally, Slovene is not included in the supported languages or is included in such a minority that the models do not work well for Slovene.

Recently, some English models were adapted for specific languages. Most notable examples are GPT-SW3 (AI-Sweden, 2024) for Swedish, Chinese LLaMa (Cui et al., 2023) and Open-Chinese-LLaMA (OpenLMLab, 2023) for Chinese, and Gervasio (Santos et al., 2024) for Portuguese. However, these models were either trained from scratch (GPT-SW3), did not use embedding transfer methods after vocabulary expansion (Chinese LLaMA and Open-Chinese-LLaMa), or were just instruction tuned for the target language (Gervasio).

Slovene is not without LLMs, though. However, existing works focused on encoder-type models, such as CroSloEngual BERT (Ulčar & Robnik-Šikonja, 2020) and SloBERTa (Ulčar & Robnik-Šikonja, 2021), or encoder-decoder-type models, such as SloT5 (Ulčar & Robnik-Šikonja, 2023). The only working open-source decoder-type model for Slovene we are aware of is GPT-sl-base (Ulčar & Robnik-Šikonja, 2022), which has only 100 million parameters and was trained on only 5 billion unique tokens and is therefore not comparable to the proposed model.

3 PRETRAINING DATA

LLMs require huge training sets. We use existing Slovene corpora for additional pretraining of our model. Our training corpora covers different types of text, such as news articles (Trendi (Kosem et al., 2023) - up to and including September 2023), academic works (KAS (Žagar et al., 2022)), web crawls (mC4 (Raffel et al., 2020), MaCoCu (Bañón et al., 2023), CC100 (Wenzek et al., 2020)), and a mixture of them (Metafida (Erjavec, 2023)). These corpora collectively contain around 10 B tokens, while Hoffman scaling laws (Hoffmann et al., 2022) suggest 20 B tokens as a suitable quantity for 1 B model. Note that pretraining of the recent Llama 3 model (AI@Meta, 2024) used even more tokens than these scaling laws suggest resulting in still better model performance. For these two reasons, we also include Croatian, Bosnian, and Serbian texts to increase our training data. We hypothesize that using these languages should improve the model's performance due to their similarity to Slovene. This was also shown in previous works, such as CroSloEngual BERT (Ulčar & Robnik-Šikonja, 2020). Additionally, we use English Wikipedia (Wikimedia Foundation, 2022) and CC-News (Hamborg et al., 2017) to prevent the model's forgetting of English. The used corpora and their properties are shown in Table 1.

We performed an additional cleaning of the KAS corpus, containing some unwanted artifacts due to the scanning of PDF documents. We cleaned these artifacts using the following heuristics. We define a set of problematic characters (Non-ASCII characters except Slovene characters (č, ž, š) and characters

Table 1: Corpora used for additional pretraining of GaMS 1B model. CBS stands for a combination of Croatian, Bosnian, and Serbian languages. The "OPT tokenizer" column shows the number of resulting tokens when the texts are tokenized with the original OPT tokenizer, while the "Slovene tokenizer" shows the number of tokens when the texts are tokenized with our tokenizer, described in Section 4.1.

Corpus	*Language*	*# tokens (OPT tokenizer)*	*# tokens (Slovene tokenizer)*
Metafida	Slovene	6.59 B	3.35 B
KAS	Slovene	3.61 B	1.66 B
Trendi	Slovene	1.4 B	0.68 B
mC4	Slovene	5.5 B	2.88 B
MaCoCu	Slovene	4.68 B	2.34 B
CC100	Slovene	0.54 B	0.29 B
Rižnica	Croatian	0.21 B	0.11 B
HrNews	Croatian	4.16 B	2.14 B
MaCoCu	CBS	15.65 B	8.63 B
Wikipedia	English	4.7 B	5.61 B
CC-News	English	0.4 B	0.46 B
Total	/	**47.44 B**	**28.13 B**

of other alphabets, such as Chinese, Greek, Cyrillic, etc.). We consider an unwanted artifact a sequence of tokens (texts are tokenized using NLTK (Bird et al., 2009) tokenizer) with a combined length of at least 5 characters that contain only problematic characters. We remove these sequences. We did not clean other corpora, as they were already thoroughly cleaned.

We performed near deduplication on Slovene corpora using the Onion tool (Pomikálek, 2011). Similarly to Ulčar and Robnik-Šikonja (2020), we use 9-grams with a duplicate content threshold of 0.9. The statistics, shown in Table 1, are computed on cleaned and deduplicated corpora.

4 ADDITIONAL PRETRAINING PROCESS

Due to the lack of Slovene data, we cannot train the model from scratch and have to use an already pretrained English model as a starting point. Most generative decoder-based LLMs have more than 1 B parameters as their performance scales with the number of parameters. There were not many open-source 1

Table 2: An example of an 1 B Falcon model output. The output shows clear signs of repetition and a tendency to generate too-long texts. The generation was interrupted, when the maximal number of allowed tokens was reached.

Prompt: This is the initial run of the Falcon model. My first message to the world is:
Output: I've done my research. I'm not a fan of the Falcon and I think the Falcon is a bad idea. I've spent some time on this site looking at the Falcon and the Falcon is not a very good idea. I've read about it and I have read a lot about it and the Falcon is not a very good idea. It is a very expensive, very inefficient, and very bad idea. It is an expensive, inefficient, and terrible idea. The reason I am writing this post is that I want people to see what I'm talking about and to see why I think the Falcon is a bad idea. I am not a huge fan of the Falcon and I think it's a horrible idea. The Falcon is a terrible idea. The Falcon is a terrible idea. It's an incredibly bad idea. It

B models available at the time we started with our work; e.g., Microsoft's Phi-1.5 (Li et al., 2023) was not yet published under an open-source license. Two possible choices were Falcon (Penedo et al., 2023) and OPT (Zhang et al., 2022). While Falcon models generally perform better on LLM leaderboards, such as Beeching et al. (2023), the 1 B version was trained only on a 350 B tokens subset of the RefinedWeb dataset (Penedo et al., 2023). For comparison, the larger versions were trained on the whole RefinedWeb (1.19 T tokens) and some other corpora, resulting in a training dataset of around 1.5 T tokens. Even the authors of the 1 B Falcon model advise treating this model only as a research artifact. By manually testing the 1 B Falcon model on some prompts, we found out that the model tends to repeat itself (even with sampling), generates longer outputs than necessary, and outputs meaningless sentences on a regular basis. An example of such output is shown in Table 2. When testing the 1.3 B version of the OPT model in a similar way, it made a better impression, and we chose it as our starting model.

OPT follows the GPT-3 architecture (Brown et al., 2020). The 1.3 B model has 24 layers with 32 attention heads. Its hidden (embedding) dimension is 2048, it uses Pre-LayerNorm (Xiong et al., 2020), ReLU activation function, absolute learned positional embeddings, and the encoder sequence length (context length) is 2048. It offsets the positional embeddings by 2 (instead of starting with 0, it starts with 2), appends EOS token at the beginning of the sequence, and its vocabulary size is 50.272.

We additionally pretrain two versions of this model, one with the original OPT vocabulary and the other with Slovene vocabulary (see Section 4.1). We refer to the versions with the original OPT vocabulary as OPT_GaMS models and to versions with Slovene vocabulary as GaMS models for the rest of this paper.

4.1 Building Slovene vocabulary of the model

We train the tokenizer for the new vocabulary using the CC100, KAS, Metafida, and HrNews (Ljubešić et al., 2024) corpora. We initially trained six different tokenizers, primarily differing in size. Our aim for the tokenizer is to be efficient on both English and Slovene texts. For vocabulary evaluation, we utilize the OpenSubtitles (Lison & Tiedemann, 2016) dataset, which includes Slovene and English subtitles, totaling around 19 million aligned lines in these two languages.

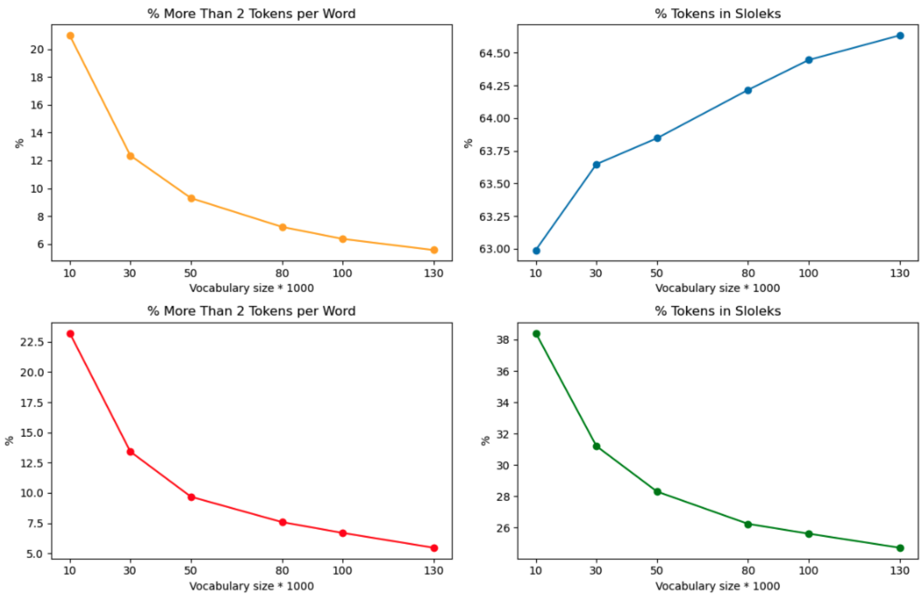
To train the tokenizer, we utilize the SentencePiece library (Kudo & Richardson, 2018) with the Byte Pair Encoding (BPE) (Sennrich et al., 2016) segmentation algorithm. We create a SentencePiece tokenizer model with a specified vocabulary size and include special tokens such as '<s>' (beginning of sequence), '</s>' (end of sequence), '<pad>' (padding token), and '<unk>' (unknown token).

We evaluate the tokenizer using three metrics. The first metric was described by Ali et al. (2023) and measures how many words are written with two or more tokens. A good tokenizer shall keep this number relatively low. The second metric assesses how many vocabulary tokens are part of the Slovene lexical database Sloleks (Dobrovoljc et al., 2019). We wish for a high value of this metric. Lastly, we create a distributional histogram displaying 10 different groups of columns, illustrating for each tokenizer the number of words written with 1, 2, ..., up to 10 or more tokens. We wish for the bulk of mass in the histogram to be on the left-hand side of the histogram. We show the results of the first two metrics, evaluated on Slovene and English subtitles datasets, in Figure 1.

The results on the Slovene and English OpenSubtitles datasets show that larger vocabularies yield better results[1]. However, the improvement in results slightly

[1]We observe that the percentage of tokens in Sloleks increases when evaluated on the Slovene dataset and decreases when evaluated on the English dataset. This trend is favorable in both

Figure 1: The evaluation of different vocabulary sizes tested on the Slovene Subtitles dataset (upper two graphs) and the English Subtitles dataset (lower two graphs).



diminishes when the vocabulary size increases from 80,000 to 100,000 tokens. As a larger vocabulary implies more model parameters, which consequently require more data for training, more required computational resources and longer training times, we have to settle for a suitable sweet spot. We opt for a vocabulary size of 80,000 tokens as our choice for the 1 B model.

4.2 Embedding transfer

Zhao et al. (2024) recently showed that vocabulary change (or expansion) can have a negative impact on the model's performance when the new embedding matrix is initialized randomly. They performed their experiments using Chinese LLaMA (Cui et al., 2023). As the Chinese language uses specific characters that are not well-represented in the vocabularies of English LLaMA models, the

cases. Initially, we have tokenized parts of words that may be similar across both languages. As the token size increases, more complete English words, which are not in the Sloleks dictionary, appear, while more complete Slovene words, which are in Sloleks, also appear.

vocabulary change (or expansion) seemed a necessary step in adapting the model for Chinese. However, if vocabulary change/expansion had a negative effect for Chinese models, it should have an even more negative impact for the Slovene model. Nevertheless, the benefit of a vocabulary change is a larger context window of the model. While the number of tokens the model can process (2048 in case of OPT) is not dependent on its vocabulary, the amount of text that can be tokenized using this number of tokens is. As seen in Table 1, tokenizing the same amount of Slovene text with OPT tokenizer results in twice as many tokens as tokenizing it with Slovene tokenizer. Hence, when using the Slovene tokenizer, the model can process Slovene texts that are double the size of those processed by the OPT tokenizer.

To keep the upsides of vocabulary change and mitigate its adverse effect on the model's performance, we tried to initialize the embedding matrix using WECHSEL (Minixhofer et al., 2022) and FOCUS (Dobler & de Melo, 2023) initialization methods. These methods initialize the embedding matrix for a new vocabulary based on the embedding matrix of the original vocabulary. Let T^s be the source tokenizer (OPT tokenizer in our case) with vocabulary V^s and corresponding embedding matrix E^s. We have a target tokenizer T^t (tokenizer from Section 4.1) with vocabulary V^t. Our goal is to initialize the embedding matrix E^t. WECHSEL and FOCUS do that by computing the similarities between tokens in a common embedding space W. We denote the representations of V^s and V^t in W with W^s and W^t. Both WECHSEL and FOCUS use FastText embeddings (Bojanowski et al., 2017) as W. We test both the original versions of these methods and our own versions, where we replace the FastText embeddings with CroSloEngual BERT embeddings (Ulčar & Robnik-Šikonja, 2020).

We denote models obtained by using WECHSEL/FOCUS as WECHSEL/FOCUS GaMS models. Additionally, OPT uses the same weights for embedding and output layer. Hence, it makes sense to transfer the output layer as well. We denote the models, where output layer is also transfered as WECHSEL/FOCUS Tied models.

4.2.1 THE WECHSEL EMBEDDINGS TRANSFER METHOD

WECHSEL (Minixhofer et al., 2022) obtains representations of vocabulary in source and target embeddings W^s and W^t by applying monolingual fastText word embeddings to V^s and V^t and aligning them using the Orthogonal Procrustes method (Schönemann, 1966; Artetxe et al., 2016) with bilingual dictionaries[2]. Based on this embeddings, it computes the cosine similarity $s_{x,y}$ between every pair $x \in V^t, y \in V^s$ using the following equation:

$$s_{x,y} = \frac{{w_x^t}^T w_y^s}{\|w_x^t\| \cdot \|w_y^s\|}, \tag{1}$$

where column vectors w_x^t and w_y^s denote the representations of x and y in W^t and W^s.

The target embeddings in E^t are initialized as a convex combination of embeddings in E^s. Let $\mathcal{J}_x \subset V^s$ denote the set of k nearest neighbors of $x \in V^t$ based on $s_{x,y}$ (k is the hyperparameter of the method). The embedding $e_x^t \in E^t$ is then computed using the softmax function:

$$e_x^t = \frac{\sum_{y \in \mathcal{J}_x} \exp(s_{x,y}/\tau) \cdot e_y^s}{\sum_{y' \in \mathcal{J}_x} \exp(s_{x,y'}/\tau)}, \tag{2}$$

where e_y^s denotes the embedding of $y \in V^s$ in E^s and τ denotes the temperature hyperparameter. We use $k = 10$ and $\tau = 0.1$ (these are default WECHSEL values) in our models.

4.2.2 THE FOCUS EMBEDDINGS TRANSFER METHOD

The FOCUS embeddings transfer method (Dobler & de Melo, 2023) initializes the target embeddings based on tokens that appear both in V^s and V^t (overlap). Let $O = V^s \cap V^t = \{o_1, o_2, ..., o_n\}$. The target embeddings of tokens in O are the same as their source embeddings:

$$\forall o \in O : e_o^t = e_o^s. \tag{3}$$

The set of non-overlapping (additional) target tokens is defined as $A = V^t \setminus O$. The embeddings e_a^t are computed based on similarities between tokens from A and O. Hence, FOCUS does not need W^s but needs only W^t, which is obtained by FastText. The difference between FOCUS and WECHSEL is that WECHSEL uses pretrained FastText, and FOCUS trains it on unlabeled data in the target language. Based on W^t, similarity $s_{a,o}$ is computed using Equation 1 for every

[2]WECHSEL code comes with already aligned embeddings, hence we did not need to align them.

pair $a \in A, o \in O$. For every $a \in A$, FOCUS defines the similarity score vector as:

$$c_a = [s_{a,o_1}, s_{a,o_2}, ..., s_{a,o_n}]. \tag{4}$$

Based on c_a, the vector of weights w_a is computed using sparsemax function (Martins & Astudillo, 2016):

$$w_a = \text{sparsemax}(c_a). \tag{5}$$

The target embedding $e_a^t \in E^t$ for an additional token $a \in A$ is then computed as:

$$e_a^t = \sum_{o \in O} w_{a,o} \cdot e_o^s. \tag{6}$$

We train the FastText model used with FOCUS on the same corpora as the tokenizer from Section 4.1. We train the FastText model for 3 epochs and include every token that occurs more than 10 times in the training dataset. The dimension of token vectors is set to 768.

4.2.3 USING CROSLOENGUAL BERT EMBEDDINGS

Croatian, Slovene, and English languages, which are part of our vocabulary, are also used in the CroSloEngual BERT model (CSE BERT). Hence, we try to upgrade WECHSEL and FOCUS by using the embedding matrix of CSE BERT as a common embedding space W. The reasoning is that CSE BERT embeddings of similar English, Slovene, and Croatian tokens shall be aligned since they are modeled by the same model. As CSE BERT has shown some promising results on SloBench classification tasks (Dragar, 2022), it should have good internal language knowledge. We expect that our approach will benefit the WECHSEL method more than FOCUS, as WECHSEL's bilingual alignment is not suitable for multi-lingual models such as ours. Even for mono-lingual models, we suspect that the linear alignment is the weakest point of WECHSEL, and our approach should address that. We refer to the models that are trained using CSE BERT embeddings as W as FOCUS/WECHSEL CSE models.

We use the following approach to embed the tokens from V^s and V^t using CSE BERT. Let $v \in V^s \cup V^t$ be the token we want to embed. First, we tokenize it with the CSE BERT's tokenizer. We denote this tokenization with t_v^{CSE}. Since CSE BERT vocabulary is not the same as V^s and V^t, v is tokenized using $k \geq 1$

tokens:

$$t_v^{CSE} = [t_{v,1}^{CSE}, ..., t_{v,k}^{CSE}]. \tag{7}$$

Let $e_{v,i}^{CSE}, 1 \leq i \leq k$ denote the product of token $t_{v,i}^{CSE}$ with embedding matrix E^{CSE} of CSE BERT (the CSE BERT embedding of token $t_{v,i}^{CSE}$). We define the common space embedding $w_v \in W$ for v as:

$$w_v = \frac{1}{k} \sum_{i=1}^{k} e_{v,i}^{CSE}. \tag{8}$$

4.3 Training the 1B models

We train our models on the Slovene HPC Vega computer (60 GPU nodes, each containing 4 NVIDIA A100 GPUs with 40 GB of RAM). We use the NVidia NeMo toolkit (version 1.22, container 23.10) for training, enabling efficient parallelization over multiple nodes on the model and data levels. As NeMo does not support positional embedding offset and ReLU activation, we forked the NeMo repository[3] and added the support for the OPT models.

We train our models on 16 nodes, using tensor parallel rank 4, enabling one instance of the model to be located on a single node, which is faster than having the model split over multiple nodes. We use a batch size of 1024, which equals around 2 million tokens (batch size in tokens is obtained by multiplying batch size with the context length of the model). Given our data, this results in 22,000 training steps for the OPT_GaMS model and 13,400 training steps for the GaMS models. We use fused Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. We use a cosine learning rate scheduler with minimal learning rate $\eta_{min} = 2 \cdot 10^{-5}$. The learning rate is first linearly increased from 0 to $\eta_{max} = 10 \cdot \eta_{min} = 2 \cdot 10^{-4}$ during warmup steps and then decayed using cosine function to η_{min}, being equal to η_{min} during the final constant steps. We use the following warmup and constant steps:

- OPT_GaMS: 1000 warmup steps, 1000 constant steps;
- GaMS: 2000 warmup steps, 500 constant steps.

When training the FOCUS/WECHSEL GaMS models, we freeze the inner parameters of the model for the first 1500 steps. During these steps, we train only the embedding and the output layer. This helps to avoid the catastrophic forgetting

[3] https://github.com/SloLama/NeMo

of the model, which can happen due to vocabulary change. We use 0.05% of our data as a validation set. Even though this percentage seems small, it still results in around 15 or 24 million (depending on tokenizer) validation tokens, which should be enough to detect potential overfitting. Additionally, we can not afford large validation sets due to low amount of training data.

As Muennighoff et al. (2023) showed, it might help to repeat the data when dealing with constrained data, we train the model for multiple epochs. We train the WECHSEL CSE GaMS model with both embedding and output layer transferred from the original OPT model (this is the best performing GaMS model on a single epoch according to validation loss) for 4 epochs. Additionally, we freeze the model's hidden layers (only the output and embedding layers are trained) for the entire first epoch. We train the whole model for the next 3 epochs. With a multi-epoch scenario, we set the LR scheduler's warmup steps to 10,000 and constant steps to 5,000.

Inspired by Li et al. (2023), we test training OPT_GaMS model (we choose OPT_GaMS instead of GaMS as GaMS seems to require more data due to a vocabulary change) only on "higher quality" data. We define higher quality data to be all data except web crawls; the selection includes news articles, literature, academic works, etc., and represents diverse, informative, and well-written texts. We use the following corpora: Metafida, KAS, Trendi, Rižnica, HrNews, Wikipedia, and CC-News. This results in around **21 B** tokens, encoded with OPT tokenizer. We train the model for 10,050 steps and set the LR scheduler's warmup and constant steps to 1,000 and 500, respectively. We refer to this model as OPT_GaMS Quality Data.

The training and validation cross-entropy losses observed during the training are shown in Figure 2. The plots were obtained using Weights & Biases platform[4]. While GaMS losses seem to be much larger than OPT_GaMS losses, the losses of these two model groups cannot be directly compared due to different vocabularies. Note that the loss is computed on different distributions (even though the training data is the same, it is tokenized into different tokens - even the ratios between languages are different as OPT tokenizer uses more tokens on average to tokenize Slovene words than Slovene tokenizers). To avoid unfair comparisons, we compare the losses of GaMS models. It is evident that FOCUS

[4] https://wandb.ai/site

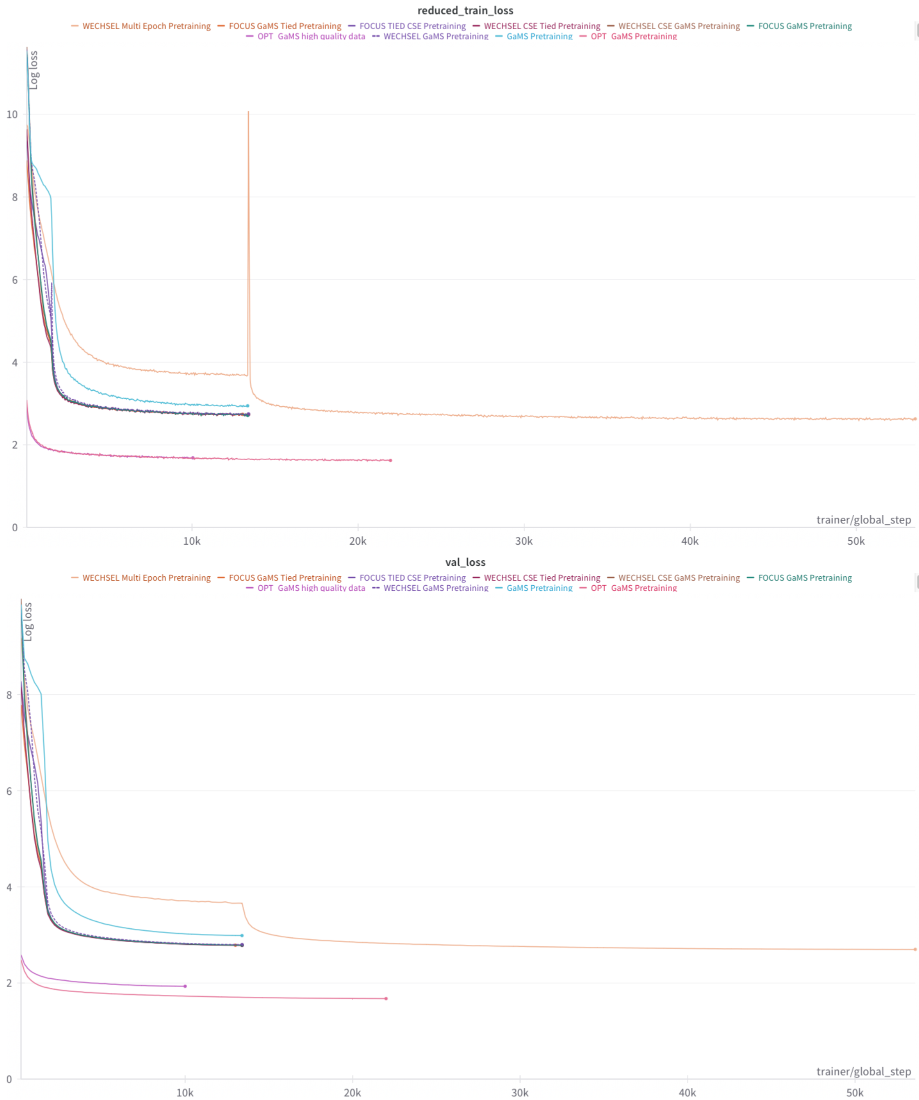
and WECHSEL improve the model performance compared to random initialization of the embedding matrix. While different transfer approaches behave differently in the early stages of the training, their losses all converge to a similar value (validation losses differ by less than **0.02**, showing no significant difference in the performance of these methods). Although Figure 2 does not show this clearly, using multiple epochs actually reduces the validation loss (the final validation loss of multi-epoch model is **2.699**, while the final validation loss of its single-epoch counterpart is **2.781**). Furtherher, training the OPT_GaMS model only on ”higher quality” data does not improve its performance.

5 EVALUATION

LLMs are commonly benchmarked for knowledge, reasoning, safety, natural language understanding, etc. The commonly used benchmarking suites for LLM evaluation in English are GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), BIGBench (Srivastava et al., 2023), Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021), etc. The benchmarks for Slovene are very limited, as due to the complexity of most LLM benchmarks, obtaining them via pure machine translation is not a viable solution. Some SuperGLUE tasks are unsuitable even for human translation due to contextual differences between languages (such as the Word in Context task) and have to be rewritten for Slovene. Besides classification tasks contained in the Slovene SuperGLUE (Žagar et al., 2020) benchmarking suite, we used two more datasets: a natural language inference classification dataset SI-NLI (Klemen et al., 2022), which is already part of SloBench (Dragar, 2022), and sentence simplification task SENTA (Žagar et al., 2024) that tests text generation abilities of LLMs.

In our evaluation scenario, all models are evaluated using in-context learning, with few-shot prompts (models are not fine-tuned on given tasks but shown a few solved examples in the prompt). The in-context examples are randomly sampled from the training set (each test instance is given different examples). None of the models, apart from OPT_GaMS INZ, are instruction-tuned. The OPT_GaMS INZ model is LoRA (Hu et al., 2022) tuned on the QA dataset that was provided to us by Inštitut za novejšo zgodovino (INZ). The dataset consists of approximately 7,000 questions and answers and is not suitable for general-purpose instruction tuning, as it contains only one task. However, this fine-

Figure 2: The training (top) and validation (bottom) cross-entropy losses observed during the training. Note that the losses of OPT_GaMS models can not be directly compared to the losses of GaMS models due to differences in the distributions.



tuning helps with the evaluation of question-answering tasks, as it helps the model to generate the answer in the correct form. All models are evaluated using greedy sampling during the generation phase, i.e. the most probable token, according to the model, is always selected as the next generated token.

5.1 Classification tasks

The number of few-shot examples and number of test set instances for each dataset from the Slovene SuperGLUE suite and the SI-NLI dataset are shown in Table 3. The number of few-shot examples in prompt (k) is determined based on the models' performances on the validation set. The number of test set instances is quite low for BoolQ (30) and RTE (29) because only human-translated examples are used for the evaluation.

Table 3: The number of test examples and the number of in-context examples in prompts (k) per data set in SupeGLUE tasks and SI-NLI.

Task	k	*# test examples*
BoolQ	3	30
CB	5	250
COPA	5	500
MultiRC	2	333
RTE	3	29
WSC	4	146
SI-NLI	5	998

To adapt the classification tasks to generative LLMs, we wrote our own framework for the evaluation of generative models, where we specify the expected form of an answer in the prompt. We observe that our 1 B models struggle to understand what output is required to complete the tasks. This is typical for models below 5 B parameters; for example, Li et al. (2023) observed similar behavior for their Phi model. This behavior is not present in larger generative models for English, especially the ones trained for instruction following. Hence, we measure the percentage of invalid predictions where a model did not generate the answer in a required form. We measure other metrics for each task only on valid predictions. The alternative would be to label the invalid predictions as wrong answers, but in this way, we cannot distinguish between invalid and wrong predictions. We also observe a high correlation between the majority label of few-shot examples and the models' output. Hence, we hypothesize that few-shot examples did not help the model to understand the tasks but only helped it with the form of the answer - more few-shot examples resulted in fewer invalid predictions.

The results for Slovene SuperGLUE tasks are shown in Tables 4 and 5. Overall, the performance of the models is quite similar and there is no model that would outperform others across all tasks. The models are outperformed by the representation model CroSloEngual BERT, which was fine-tuned on these tasks. As this model has seen significantly more training instances, the comparison is not fair but the score indicates what is achievable with relatively small LLMs. The most difficult task for the models, according to the percentages of invalid predictions, is MultiRC. In this task, the model is given a text, a question, and a list of answers. The model has to return the numbers of correct answers. We could make this task easier for the models by giving them each answer separately and asking them to classify them as correct or wrong. However, as the purpose of the task is to check whether the model can select the correct answers from multiple choices, we decided to present it in this more challenging form. OPT_GaMS INZ model produced a significantly lower percentage of invalid predictions on this task than other models, suggesting that instruction tuning should make the task less challenging.

Table 4: Test set results with 95 % confidence intervals for Slovene Super GLUE tasks BoolQ, CB, and COPA. Columns Acc. represent models' accuracy, and columns Inv. pred. represent the percentage of invalid predictions for each model. Confidence intervals are computed using standard error estimation for accuracy, and using quantile bootstrap for F_1-score. The results for CroSloEngual BERT are copied from SloBench.

	BoolQ		*CB*			*COPA*	
Model	*Acc.*	*Inv. pred.*	*Acc.*	F_1	*Inv. pred.*	*Acc.*	*Inv. pred.*
OPT_GaMS	0.57 [0.38, 0.75]	0 %	0.44 [0.38, 0.50]	0.32 [0.26, 0.39]	0 %	0.46 [0.42, 0.51]	0 %
GaMS	0.50 [0.31, 0.69]	0 %	0.43 [0.37, 0.50]	0.30 [0.25, 0.33]	1.20 %	0.49 [0.44, 0.54]	17.20 %
WECHSEL GaMS	0.67 [0.49, 0.85]	0 %	0.50 [0.44, 0.56]	0.39 [0.32, 0.47]	1.20 %	0.48 [0.44, 0.52]	0.20 %
FOCUS GaMS	0.67 [0.49, 0.85]	0 %	0.51 [0.45, 0.58]	0.38 [0.31, 0.46]	1.60 %	0.48 [0.43, 0.53]	27.80 %
WECHSEL CSE	0.57 [0.38, 0.75]	0 %	0.50 [0.44, 0.56]	0.34 [0.30, 0.38]	0.40 %	0.48 [0.44, 0.53]	2.80 %
WECHSEL CSE Tied	0.47 [0.28, 0.66	0 %	0.51 [0.45, 0.57]	0.38 [0.32, 0.46]	2.40 %	0.48 [0.44, 0.53]	0.40 %
FOCUS CSE Tied	0.50 [0.31, 0.69]	0 %	0.48 [0.42, 0.54]	0.36 [0.29, 0.44]	0.40 %	0.47 [0.43, 0.51]	3.40 %
FOCUS GaMS Tied	0.53 [0.34, 0.72]	0 %	0.48 [0.42, 0.54]	0.36 [0.29, 0.44]	1.20 %	0.48 [0.43, 0.53]	12.00 %
OPT_GaMS Quality Data	0.60 [0.41, 0.79]	0 %	0.44 [0.37, 0.50]	0.35 [0.28, 0.43]	0.80 %	0.48 [0.44, 0.52]	0 %
OPT_GaMS INZ	0.60 [0.41, 0.79]	0 %	0.44 [0.37, 0.50]	0.32 [0.26, 0.40]	0 %	0.45 [0.41, 0.49]	0 %
WECHSEL Multi-Epoch	0.60 [0.41, 0.79]	0 %	0.51 [0.45, 0.57]	0.38 [0.31, 0.46]	0.80 %	0.46 [0.42, 0.51]	1.20 %
CroSloEngual BERT	0.73	/	0.79	0.74	/	0.57	/

Table 5: Test set results with 95 % confidence intervals for Slovene Super GLUE tasks MultiRC, RTE and WSC. Columns Acc. represent models' accuracy, column EM represents the exact match between predictions and true labels, and columns Inv. pred. represent the percentage of invalid predictions for each model. Confidence intervals are computed using standard error estimation for accuracy and exact match, and using quantile bootstrap for F_1-score. The results for CroSloEngual BERT are copied from SloBench.

	MultiRC			*RTE*		*WSC*	
Model	*EM*	F_1	*Inv. pred.*	*Acc.*	*Inv. pred.*	*Acc.*	*Inv. pred.*
OPT_GaMS	0.15 [0.02, 0.28]	0.43 [0.32, 0.54]	90.09 %	0.41 [0.22, 0.60]	0 %	0.51 [0.43, 0.60]	0 %
GaMS	0.03 [-0.03, 0.09]	0.16 [0.12, 0.20]	89.49 %	0.43 [0.23, 0.62]	3.45 %	0.42 [0.34, 0.50]	0 %
WECHSEL GaMS	0.15 [0.03, 0.26]	0.37 [0.30, 0.43]	87.69 %	0.43 [0.23, 0.62]	3.45 %	0.47 [0.38, 0.55]	0 %
FOCUS GaMS	0.11 [0.00, 0.21]	0.36 [0.29, 0.44]	88.59 %	0.54 [0.34, 0.73]	3.45 %	0.50 [0.42, 0.58]	0 %
WECHSEL CSE	0.06 [-0.01, 0.12]	0.26 [0.20, 0.31]	84.08 %	0.43 [0.23, 0.62]	3.45 %	0.45 [0.36, 0.53]	0 %
WECHSEL CSE Tied	0.12 [0.03, 0.21]	0.21 [0.17, 0.25]	84.38 %	0.46 [0.27, 0.66]	3.45 %	0.55 [0.47, 0.63]	0 %
FOCUS CSE Tied	0.09 [0.01, 0.17]	0.26 [0.21, 0.31]	83.48 %	0.43 [0.23, 0.62]	3.45 %	0.55 [0.47, 0.64]	0 %
FOCUS GaMS Tied	0.05 [0.02, 0.08]	0.22 [0.20, 0.24]	36.64 %	0.46 [0.27, 0.66]	3.45 %	0.49 [0.41, 0.58]	0 %
OPT_GaMS Quality Data	0.12 [0.04, 0.19]	0.32 [0.25, 0.39]	79.28 %	0.38 [0.19, 0.57]	0 %	0.47 [0.39, 0.55]	0 %
OPT_GaMS INZ	0.07 [0.04, 0.09]	0.34 [0.32, 0.37]	2.10 %	0.38 [0.19, 0.57]	0 %	0.45 [0.36, 0.53]	0 %
WECHSEL Multi-Epoch	0.13 [0.05, 0.21]	0.28 [0.22, 0.33]	79.28 %	0.50 [0.30, 0.70]	3.45 %	0.54 [0.46, 0.62]	0 %
CroSloEngual BERT	0.09	0.52	/	0.66	/	0.61	/

The results for the SI-NLI dataset are shown in Table 6. The performance of our models is quite similar, and the confidence intervals overlap. All models return invalid predictions for approximately half of the test instances (the best-performing model with respect to that metric is WECHSEL CSE, with 44.69 % of invalid predictions). The reason for these similarities is that all models perform poorly due to lack of task understanding. Hence, the models should be instruction-tuned first in order to spot any significant differences between them. The models are significantly outperformed by GPT and BERT models; again the comparison is not fair as BERT models were fine-tuned on this data set and GPT-3.5-Turbo is significantly larger.

5.2 Sentence simplification

The models introduced in this paper are generative. Therefore, it makes sense to evaluate them on language generation tasks. We choose sentence simplification task SENTA (Žagar et al., 2024). The model is given a sentence and asked to simplify it. Here, we observe that our models perform better than in classification tasks but there are still some problems with the task understanding, as the models sometimes return "Poenostavi naslednji stavek." (eng. "Simplify the given sentence.") as an answer in case of few-shot prompts. They return

Table 6: Test set results with 95 % confidence intervals for the SI-NLI dataset. Columns Inv. pred. represent the percentage of invalid predictions for each model. Confidence intervals are computed using standard error estimation for accuracy and using quantile bootstrap for F_1-score. The results for GPT-3.5-Turbo, SloBERTa, and CroSloEngual BERT are copied from SloBench.

Model	*Accuracy*	*Entailment F_1*	*Neutral F_1*	*Contradiction F_1*	*Inv. pred.*
OPT_GaMS	0.32 [0.27, 0.36]	0.38 [0.32, 0.45]	0.17 [0.10, 0.24]	0.34 [0.27, 0.40]	51.40 %
GaMS	0.29 [0.25, 0.33]	0.31 [0.24, 0.38]	0.32 [0.25, 0.38]	0.25 [0.19, 0.32]	50.00 %
WECHSEL GaMS	0.33 [0.29, 0.37]	0.39 [0.33, 0.45]	0.33 [0.27, 0.39]	0.26 [0.20, 0.32]	44.69 %
FOCUS GaMS	0.34 [0.30, 0.38]	0.40 [0.34, 0.46]	0.37 [0.31, 0.44]	0.20 [0.13, 0.26]	49.40 %
WECHSEL CSE	0.32 [0.28, 0.36]	0.38 [0.32, 0.43]	0.37 [0.30, 0.43]	0.17 [0.11, 0.24]	47.80 %
WECHSEL CSE Tied	0.35 [0.31, 0.39]	0.40 [0.34, 0.46]	0.41 [0.35, 0.47]	0.20 [0.14, 0.27]	48.30 %
FOCUS CSE Tied	0.34 [0.30, 0.38]	0.38 [0.32, 0.44]	0.38 [0.32, 0.44]	0.23 [0.17, 0.30]	47.19 %
FOCUS GaMS Tied	0.32 [0.28, 0.36]	0.37 [0.31, 0.43]	0.37 [0.31, 0.44]	0.20 [0.14, 0.26]	47.19 %
OPT_GaMS Quality Data	0.31 [0.27, 0.35]	0.38 [0.32, 0.44]	0.28 [0.22, 0.35]	0.28 [0.22, 0.35]	47.39 %
OPT_GaMS INZ	0.30 [0.26, 0.35]	0.36 [0.29, 0.42]	0.25 [0.18, 0.32]	0.29 [0.23, 0.36]	53.31 %
WECHSEL Multi-Epoch	0.30 [0.26, 0.34]	0.37 [0.31, 0.43]	0.37 [0.31, 0.43]	0.17 [0.11, 0.24]	51.10 %
GPT-3.5-Turbo	0.86	0.85	0.82	0.90	/
SloBERTa	0.74	0.76	0.71	0.64	/
CroSloEngual BERT	0.66	0.69	0.63	0.66	/

this sentence, as this is the instruction added to each example in the prompt and is consequently the most common sentence in the prompt.

We evaluate our models using different values k of few-shot examples. We test values $k \in \{0, 3, 5, 10\}$. We use SARI score[5] as an evaluation metric. SARI score is commonly used to evaluate text simplification systems. It compares the system's output to both the input and reference output. It computes the F_1-score for added and preserved tokens and precision for deleted words. It is computed using the following equation:

$$\text{SARI} = \frac{F_{1,add} + F_{1,keep} + P_{del}}{3}, \tag{9}$$

where $F_{1,add}$ and $F_{1,keep}$ represent the 4-gram F_1 score for add/keep operations and P_{del} denotes the 4-gram precision score for delete operations. The goal is to have as high F_1 and precision scores as possible, meaning that higher SARI score is better.

The results are shown in Table 7. All models perform similarly (no significant differences between their SARI scores). Using a larger number of few-shot examples seems to improve the performance of the majority of the models

[5] https://huggingface.co/spaces/evaluate-metric/sari

(the exception here is LoRA-tuned OPT_GaMS INZ, which works best in the 0-shot scenario). Surprisingly, our models perform similarly or better than GPT-3.5-Turbo. Our best-performing model (WECHSEL GaMS in the 10-shot scenario) also outperforms the best-performing SloT5 model that was trained on this task. However, the differences in the SARI scores are not significant. We believe that the performance of our models could improve drastically with instruction-tuning, as the models would better understand the task instruction.

Table 7: SARI scores with 95 % confidence intervals on SENTA task. Confidence intervals were computed using quantile bootstrap method. Value of k in columns denotes the number of shown examples in few-shot prompts. The results for GPT and T5 models are copied from Žagar et al. (2024)

Model	$k = 0$	$k = 3$	$k = 5$	$k = 10$
OPT_GaMS	39.38 [38.63, 40.16]	38.51 [37.60, 39.46]	39.49 [38.67, 40.40]	39.67 [38.80, 40.60]
GaMS	39.58 [38.76, 40.47]	38.92 [37.96, 39.86]	37.98 [37.10, 38.90]	39.18 [38.37, 40.06]
WECHSEL GaMS	39.34 [38.59, 40.15]	39.53 [38.55, 40.43]	**39.87 [39.01, 40.77]**	**41.62 [40.82, 42.30]**
FOCUS GaMS	38.50 [37.77, 39.37]	**40.16 [39.27, 41.07]**	39.67 [38.81, 40.57]	41.16 [40.41, 41.89]
WECHSEL CSE	39.02 [38.28, 39.83]	39.42 [38.49, 40.35]	39.22 [38.37, 40.05]	40.54 [39.79, 41.26]
WECHSEL CSE Tied	38.77 [37.96, 39.60]	38.67 [37.79, 39.61]	39.29 [38.41, 40.20]	40.91 [40.13, 41.71]
FOCUS CSE Tied	38.93 [38.15, 39.77]	38.95 [38.02, 39.92]	39.38 [38.54, 40.25]	40.98 [40.15, 41.79]
FOCUS GaMS Tied	38.80 [37.99, 39.67]	40.05 [39.19, 40.97]	39.74 [38.86, 40.64]	41.50 [40.79, 42.20]
OPT_GaMS Quality Data	38.76 [37.97, 39.58]	37.62 [36.72, 38.47]	38.48 [37.64, 39.44]	39.02 [38.14, 39.91]
OPT_GaMS INZ	**40.29 [39.49, 41.10]**	37.88 [36.99, 38.88]	38.58 [37.72, 39.54]	38.90 [38.00, 39.85]
WECHSEL Multi-Epoch	38.80 [37.96, 39.62]	40.06 [39.23, 40.96]	39.80 [38.97, 40.62]	40.99 [40.23, 41.67]
GPT-3.5-Turbo	38.76			
SloT5-small	39.79			
mT5-small	39.09			
SloT5-large	41.01			

We can conclude that our 1 B Slovene models are not suitable for in-context learning of classification tasks but work well in generative tasks. Their performance on classification tasks with fine-tuning remains part of the future work.

6 CONCLUSION

In this work, we presented the new 1 B Slovene generative model GaMS[6], which is based on the English OPT model. The model is the first fully open-source generative language model for Slovene. Based on the analysis of different vocabulary sizes, we created a new tokenizer that was trained on Slovene, En-

[6] https://huggingface.co/cjvt/OPT_GaMS-1B

glish, and Croatian texts. We tested different embedding initialization methods and showed that they reduce both training and validation loss for next token prediction compared to random initialization.

The main challenge that we face in this work is a robust evaluation of the models. Direct comparison of training/validation losses for models using different vocabularies is not sensible, as the distributions of tokens (on which the loss is computed) are different. The comparison of models on classification benchmarking tasks is inconclusive, as the models do not really understand the tasks due to their size and lack of instruction tuning. We showed that our models perform better on generative tasks like sentence simplification but we need more tasks to get reliable conclusions on models performance.

In the future work, we will develop an instruction-following dataset and instruction-tune our models. This might improve the models performance on classification tasks, as the models will understand the evaluation tasks. For classification tasks, fine-tuning of models is also sensible. Additionally, we plan to train and release a larger model, where the differences between embedding initialization methods should be more significant.

ACKNOWLEDGMENTS

The work was supported by the Slovenian Research and Innovation Agency (ARIS) research project PoVeJMo (Adaptive Natural Language Processing with the Help of Large Language Models), core research programme P6-0411, and projects J7-3159, and L2-50070.

REFERENCES

AI-Sweden (Ed.). (2024). *GPT-SW3.* Retrieved May 28, 2024, from https://huggingface.co/AI-Sweden-Models/gpt-sw3-40b

AI@Meta. (2024). *Llama 3 Model Card.* https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbering, M., Leveling, J., ... Flores-Herr, N. (2023). Tokenizer Choice For LLM Training: Negligible or Crucial? *CoRR*. https://doi.org/10.48550/arXiv.2310.08754

Artetxe, M., Labaka, G., & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2289–2294). https://aclanthology.org/D16-1250

Bañón, M., Chichirau, M., Esplà-Gomis, M., Forcada, M. L., Galiano-Jiménez, A., García-Romero, C., ... Zaragoza-Bernabeu, J. (2023). *Slovene web corpus MaCoCu-sl 2.0.* http://hdl.handle.net/11356/1795 (Slovenian language resource repository CLARIN.SI)

Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., ... Wolf, T. (2023). *Open LLM Leaderboard.* https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard. Hugging Face.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.". https://books.google.si/books?id=KGIbfiiP1i4C

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://aclanthology.org/Q17-1010

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Cui, Y., Yang, Z., & Yao, X. (2023). *Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca.* arXiv. https://arxiv.org/abs/2304.08177

Dobler, K., & de Melo, G. (2023). FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 13440–13454). https://aclanthology.org/2023.emnlp-main.829

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., ... Robnik-Šikonja, M. (2019). *Morphological lexicon Sloleks 2.0.* http://hdl.handle.net/11356/1230 (Slovenian language resource repository CLARIN.SI)

Dragar, F. (2022). *SloBench: Slovenian Natural Language Processing Benchmark.* https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=eng&id=136261

Erjavec, T. (2023). *Corpus of combined slovenian corpora metaFida 1.0.* http://hdl.handle.net/11356/1775 (Slovenian language resource repository CLARIN.SI)

Hamborg, F., Meuschke, N., Breitinger, C., & Gipp, B. (2017). news-please: A generic news crawler and extractor. In *Proceedings of the 15th international symposium of information science* (pp. 218–223). https://api.semanticscholar.org/CorpusID:5830937

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021).

Measuring Massive Multitask Language Understanding. In *International conference on learning representations.* https://openreview.net/forum?id=d7KBjmI3GmQ

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... Sifre, L. (2022). An empirical analysis of compute-optimal large language model training. In *Advances in neural information processing systems* (Vol. 35, pp. 30016–30030). https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In *International conference on learning representations.* https://openreview.net/forum?id=nZeVKeeFYf9

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., ... Sayed, W. E. (2023). *Mistral 7b.* arXiv. https://arxiv.org/abs/2310.06825

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... Sayed, W. E. (2024). *Mixtral of Experts.* arXiv. https://arxiv.org/abs/2401.04088

Klemen, M., Žagar, A., Čibej, J., & Robnik-Šikonja, M. (2022). *Slovene natural language inference dataset SI-NLI.* http://hdl.handle.net/11356/1707 (Slovenian language resource repository CLARIN.SI)

Kosem, I., Čibej, J., Dobrovoljc, K., Erjavec, T., Ljubešić, N., Ponikvar, P., ... Krek, S. (2023). *Monitor corpus of Slovene Trendi 2023-09.* http://hdl.handle.net/11356/1879 (Slovenian language resource repository CLARIN.SI)

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 66–71). https://aclanthology.org/D18-2012

Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., & Lee, Y. T. (2023). *Textbooks are all you need ii: **phi-1.5** technical report.* arXiv. https://arxiv.org/abs/2309.05463

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., ... Li, X. (2022). Few-shot Learning with Multilingual Generative Language Models. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 9019–9052). https://aclanthology.org/2022.emnlp-main.616

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 923–929). European Language Resources Association (ELRA). https://aclanthology.org/L16-1147

Ljubešić, N., Suchomel, V., Rupnik, P., Kuzman, T., & van Noord, R. (2024). Language models on a diet: Cost-efficient development of encoders for closely-related languages via additional pretraining. In *Proceedings of the 3rd annual meeting of the special interest group on under-resourced languages @ lrec-coling 2024* (pp.

189–203). https://aclanthology.org/2024.sigul-1.23

Luo, Y., Kong, Q., Xu, N., Cao, J., Hao, B., Qu, B., ... Zeng, D. (2023). *YAYI 2: Multilingual Open-Source Large Language Models.* arXiv. https://arxiv.org/abs/2312.14862

Martins, A., & Astudillo, R. (2016). From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. In *Proceedings of the 33rd international conference on machine learning* (pp. 1614–1623). https://proceedings.mlr.press/v48/martins16.html

Minixhofer, B., Paischer, F., & Rekabsaz, N. (2022). WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 3992–4006). https://aclanthology.org/2022.naacl-main.293

Muennighoff, N., Rush, A. M., Barak, B., Scao, T. L., Tazi, N., Piktus, A., ... Raffel, C. (2023). Scaling Data-Constrained Language Models. In *Thirty-seventh conference on neural information processing systems* (pp. 50358–50376). https://proceedings.neurips.cc/paper_files/paper/2023/file/9d89448b63ce1e2e8dc7af72c984c196-Paper-Conference.pdf

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., ... Zoph, B. (2024). *GPT-4 Technical Report.* arXiv. https://arxiv.org/abs/2303.08774

OpenLMLab (Ed.). (2023). *Open-Chinese-LLaMA.* Retrieved May, 28, 2024, from https://huggingface.co/openlmlab/open-chinese-llama-7b-patch

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Alobeidli, H., Cappelli, A., ... Launay, J. (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Only. In *Advances in neural information processing systems* (Vol. 36, pp. 79155–79172). https://proceedings.neurips.cc/paper_files/paper/2023/file/fa3ed726cc5073b9c31e3e49a807789c-Paper-Datasets_and_Benchmarks.pdf

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora*. Doctoral theses, dissertations, Masaryk university, Faculty of informatics, Brno, Czech Republic. https://theses.cz/id/nqo9nn/

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*(1). https://dl.acm.org/doi/abs/10.5555/3455716.3455856

Santos, R., Silva, J. R., Gomes, L., Rodrigues, J., & Branco, A. (2024). Advancing Generative AI for Portuguese with Open Decoder Gervásio PT*. In *Proceedings of the 3rd annual meeting of the special interest group on under-resourced languages @ lrec-coling 2024* (pp. 16–26). https://aclanthology.org/2024.sigul-1.3

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem.

Psychometrika. https://doi.org/10.1007/BF02289451

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1715–1725). https://aclanthology.org/P16-1162

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A., Abid, A., Fisch, A., ... Wu, Z. (2023). Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 1–95. https://openreview.net/pdf?id=uyTL5Bvosj

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models.* arXiv. https://arxiv.org/abs/2302.13971

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models.* arXiv. https://arxiv.org/abs/2307.09288

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Text, speech, and dialogue TSD 2020* (Vol. 12284). https://doi.org/10.1007/978-3-030-58323-1_11

Ulčar, M., & Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model.. https://api.semanticscholar.org/CorpusID:250066999

Ulčar, M., & Robnik-Šikonja, M. (2022). *GPT-sl-base.* Retrieved May, 28, 2024, from https://huggingface.co/cjvt/gpt-sl-base

Ulčar, M., & Robnik-Šikonja, M. (2023). Sequence-to-sequence pretraining for a less-resourced Slovenian language. *Frontiers in Artificial Intelligence*, *6*. https://www.frontiersin.org/articles/10.3389/frai.2023.932519 doi: 10.3389/frai.2023.932519

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... Bowman, S. R. (2019). SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd international conference on neural information processing systems.* Curran Associates Inc. https://dl.acm.org/doi/10.5555/3454287.3454581

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 353–355). https://aclanthology.org/W18-5446

Wei, X., Wei, H., Lin, H., Li, T., Zhang, P., Ren, X., ... Xie, J. (2023). *Polylm: An open source polyglot large language model.* arXiv. https://arxiv.org/abs/2307.06018

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave,

E. (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 4003–4012). https://aclanthology.org/2020.lrec-1.494

Wikimedia Foundation (Ed.). (2022). *Wikimedia Downloads.* Retrieved May 8, 2024, from https://huggingface.co/datasets/wikipedia

Workshop, B., :, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., ... Wolf, T. (2023). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.* arXiv. https://arxiv.org/abs/2211.05100

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., ... Liu, T.-Y. (2020). On layer normalization in the transformer architecture. In *Proceedings of the 37th international conference on machine learning.* https://dl.acm.org/doi/abs/10.5555/3524938.3525913

Žagar, A., Kavaš, M., Robnik-Šikonja, M., Erjavec, T., Fišer, D., Ljubešić, N., ... Hrovat, G. (2022). *Abstracts from the KAS corpus KAS-abs 2.0.* http://hdl.handle.net/11356/1449 (Slovenian language resource repository CLARIN.SI)

Žagar, A., Klemen, M., Robnik-Šikonja, M., & Kosem, I. (2024). SENTA: Sentence simplification system for Slovene. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 14687–14692). https://aclanthology.org/2024.lrec-main.1279

Žagar, A., Robnik-Šikonja, M., Goli, T., & Arhar Holdt, Š. (2020). *Slovene translation of SuperGLUE.* http://hdl.handle.net/11356/1380 (Slovenian language resource repository CLARIN.SI)

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... Zettlemoyer, L. (2022). *OPT: Open Pre-trained Transformer Language Models.* arXiv. https://arxiv.org/abs/2205.01068

Zhao, J., Zhang, Z., Zhang, Q., Gui, T., & Huang, X. (2024). *LLaMA Beyond English: An Empirical Study on Language Capability Transfer.* arXiv. https://arxiv.org/abs/2401.01055

GENERATIVNI MODEL Z MILIJARDO PARAMETROV ZA JEZIK Z MANJ VIRI

Veliki jezikovni modeli so osnovna infrastruktura za sodobno obdelavo naravnega jezika. Za angleščino obstajajo številni komercialni in odprtokodni modeli, na primer ChatGPT, Llama, Falcon in Mistral. Ker so ti modeli učeni večinoma na angleških besedilih, sta njihovo znanje in poznavanje jezikov ter družb z manj viri površna. Predstavljamo razvoj novega generativnega velikega jezikovnega modela za jezik z malo viri. Za slovenski model, imenovan GaMS 1B (Generativni Model za Slovenščino), z 1 milijardo parametrov smo razvili nov tokenizator, prilagojen slovenščini, hrvaščini in angleščini, ter uporabili metodi inicializacije vektorskih vložitev FOCUS in WECHSEL za prenos vložitev iz obstoječega angleškega modela OPT. Zgrajene modele smo ovrednotili na slovenski zbirki klasifikacijskih učnih množic in na generativni nalogi poenostavljanja stavkov SENTA. Pri evalvaciji smo uporabili le učenje v kontekstu z nekaj učnimi primeri ter modele, ki še niso prilagojeni za sledenje navodilom. Pri takih nastavitvah so na klasifikacijskih nalogah zgrajeni generativni modeli zaostali za obstoječimi slovenskimi modeli tipa BERT, ki so bili prilagojeni za dane naloge. Pri nalogi poenostavljanja stavkov modeli GaMS dosegajo primerljive ali boljše rezultate kot model GPT-3.5-Turbo.

Keywords: veliki jezikovni modeli, generativni modeli, prenos znanja, OPT model, GaMS model, jezikovno prilagajanje

THE GOVORI.SI SPEECH TRANSCRIPTION PLATFORM

Klara ŽNIDERŠIČ,[1] Vid KLOPČIČ,[1] Matevž PESEK,[1] Matija MAROLT[1]
[1]University of Ljubljana, Faculty of Computer and Information Science

Despite considerable progress in automatic speech recognition, there are very few open speech transcription solutions with user-friendly graphical interfaces, especially for low-resource languages. To fill this gap, we have developed and implemented *govori.si*, a speech transcription platform specifically tailored to the Slovenian language. We trained an automatic speech recognition model that is among the best current Slovenian automatic speech recognition models and used state-of-the-art methods to tackle other transcription challenges: diarization, capitalization, punctuation, custom substitution dictionaries and numerical notation. The platform is available free of charge for research and non-commercial purposes and has been well received by users, making it a valuable tool for various applications, including legislative processes, journalism, and research.

Keywords: Slovenian language, speech recognition, natural language processing

1 INTRODUCTION

The field of automatic speech recognition has seen remarkable advancements over the years and automatic speech transcription services have become indispensable tools in numerous domains, streamlining time-consuming tasks. However, despite the ubiquity of these services, there are still challenges in extending their capabilities to non-English languages. Recent studies have mainly focused on English language processing in neutral environments (Al-Fraihat et al., 2024), but language-specific special cases, particularly for languages with limited training data and complex linguistic variations, need to be additionally considered to broaden the scope of speech recognition.

In this paper, we present a state-of-the-art speech transcription platform *govori.si*, which is specifically tailored to the Slovenian language and enables seamless speech-to-text conversion with high accuracy and efficiency. The main advantages of the proposed platform lie in the full implementation of the speech transcription pipeline and the adaptability to the different needs of end users in a wide range of applications. Whether it is the transcription of lectures, speeches, interviews, meetings or other lengthy voice recordings, *govori.si* can increase productivity by speeding up the transcription process and saving valuable time and resources.

The article is structured as follows: Section 2 presents the relevant ASR tools and an overview of related interfaces for speech transcription. Section 3 describes the methods used in each phase of the transcription process, Section 4 discusses the technologies used for the development of our platform, and Section 5 describes the user interface of *govori.si*. The results of an informal user study are presented in Section 6, and the article concludes with Section 7.

2 LITERATURE REVIEW

Automatic speech recognition (ASR) is the process of transcribing speech into text. Since the diversity of speech signals poses challenges to automatic methods, researchers have pursued different strategies to overcome them. In recent years, deep learning models trained on large corpora are the predominant choice, especially the models based on transformer blocks such as OpenAI's Whisper (Radford et al., 2023) or on convolution-augmented transformer blocks such as Google's USM (Zhang et al., 2023).

Although the Slovenian language is included in many multilingual ASR models (Pratap et al., 2020; Li et al., 2022; Radford et al., 2023; Zhang et al., 2023), it may be underrepresented in the training data, which has a negative impact on the accuracy of these models. Therefore, efforts have been made in recent years to develop ASR datasets and models specifically for the Slovenian language. The ARTUR dataset (Verdonik et al., 2023) contains 884 hours of transcribed speech and was specifically designed to train ASR models for the Slovenian language. SloBENCH (Žitnik & Dragar, 2021) serves as a central evaluation platform for benchmarking the progress of Slovenian natural language processing technologies, including machine translation between Slovenian and English, named entity recognition, universal dependency parsing, and speech recognition. Several ASR models have also been developed in recent years. The ASR system of (Gril & Dobrišek, 2022) uses a hybrid acoustic modeling approach that combines hidden Markov models (HMMs) with deep neural networks (DNNs).

Two models based on convolution-augmented transformer blocks (conformers) were also introduced (Lebar Bajec et al., 2022; *True-bar 23.02 ASR model*, 2023).

Integrated solutions for transcribing and editing speech are often found in commercial software packages. Unfortunately, the high associated costs are often an obstacle for researchers or projects with limited budgets. Manual transcription, while accurate, is time-consuming and labor-intensive. Therefore, open source solutions are the only viable option for many individuals or organizations looking to streamline the analysis of spoken text, interviews, and other audio content.

Vink (Tolle et al., 2024) is an open-source automatic transcription tool that simplifies the use of OpenAI's Whisper ASR model for non-programmers engaged in qualitative research. The tool is available as a standalone application for the Windows operating system. Users reported that transcription accuracy varied widely across the 14 languages tested, with the highest error rates for the languages with the smallest training datasets. A similar result was observed when evaluating the tool aTrain (Haberl et al., 2024), which also uses the Whisper transcription model and is implemented as an offline application for Windows. As these tools are not dependent on an internet connection and do not require the data to be uploaded to online servers, they meet the criteria of data protection, ethical guidelines, and legal compliance. While both Vink and aTrain offer a graphical user interface, they lack additional tools for text editing. Their functionality is limited to basic functions such as copying to the clipboard or exporting transcripts in various formats used for synchronizing audio and transcripts. In addition, the parameter settings for transcription are severely limited in both applications.

Another open source transcription tool, SpeechToText (Negrão & Domingues, 2021), was originally developed as a software module for the forensic software Autopsy, with the primary goal of integrating speech content into the standard workflow of digital forensic investigations. Powered by Mozilla's DeepSpeech for speech transcription, SpeechToText was tested with English recordings of non-native speakers obtained from various Android applications that can generate audio files with speech. These tests aimed to replicate scenarios that occur during forensic investigations. However, SpeechToText can also be used independently and is available under an open source license. Compared to Vink and aTrain, its advantage lies in the graphical user interface, as it also includes a transcription viewer and a keyword browser. The disadvantage, however, is the limited language support of DeepSpeech, which is no longer being actively developed.

An application closely related to *govori.si* was developed in Estonia by Olev and Alumäe (2022). The Estonian speech recognition and transcription editing system uses the Wav2Vec2 model, and its evaluation shows that even smaller languages can benefit from such a pre-trained model. It is provided as a publicly available service and offers a graphical user interface for transcription editing, interactive listening to recordings, and speaker annotation features.

3 THE GOVORI.SI SPEECH PROCESSING PIPELINE

The *govori.si* platform processes the input speech in several phases, starting with diarization, automatic speech recognition, capitalization, punctuation, replacement of regular expressions and ending with the processing of numbers. We briefly present the phases in the following subsections.

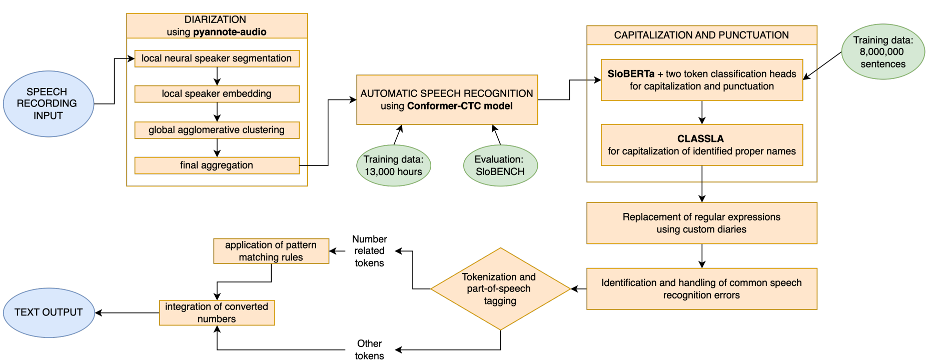


Figure 1: Graphic block scheme

3.1 Diarization

For speech diarization, we use pyannote-audio (Plaquet & Bredin, 2023; Bredin, 2023)[1], an open-source Python toolkit known for its good performance in

[1]https://github.com/pyannote/pyannote-audio

speaker diarization. The diarization model consists of a speaker segmentation model applied to short sliding windows, neural speaker embedding of each (local) speaker, and (global) agglomerative clustering. We chose pyannote because it is seamlessly implemented in the Python language and has a modular architecture that facilitates the integration of new modules and provides the flexibility to customise the parameters of existing components to our specific needs.

3.2 Speech recognition

We have trained a new speech recognition model for the Slovenian language. The model is based on convolution-augmented transformer (conformer) blocks, which combine convolutional neural networks and transformer models to utilise the best of both approaches. While transformer models are well suited to capture content-based global interactions, CNNs effectively utilise local features (Gulati et al., 2020). The Conformer-CTC model is specifically designed for ASR tasks and uses connectionist temporal classification (CTC) loss and decoding, making it a non-autoregressive model. The NeMo framework from Nvidia was used to train and evaluate the model.

We trained the model with a dataset containing 13,000 hours of speech and the corresponding transcriptions. The dataset includes existing public collections: Gos (Zwitter Vitez et al., 2021), Gos VideoLectures (VideoLectures.NET, 2019), CommonVoice, SiTEDx (Žgank et al., 2016), Sofes (Dobrišek et al., 2017) and ARTUR (Verdonik et al., 2023). We augmented the datasets with publicly available online sources, including dialectal speech from *narecja.si*, radio and television broadcasts, and parliamentary debates, as well as a smaller proportion of datasets from other Slavic languages (Croatian, Serbian, Czech, and Russian).

We evaluated the model (labelled CON-ASR-1.1) as part of the SloBENCH evaluation. The SloBENCH evaluation data for ASR consists of 15 recordings with a total duration of almost 3.5 hours, including public and private speech from southwestern and northeastern Slovenia, represented by male and female speakers. The character error rate of our model was 0.019 and the word error rate was 0.050, which is comparable to the best evaluated model (True-bar

23.02) and significantly better than the OpenAI Whisper multilingual model. For more details on the model's performance see CJVT SloBench Leaderboard[2].

3.3 Capitalization and punctuation

To capitalize and punctuate the output of the speech recognition model, we use a two-step process. In the first step, we trained two token-level classifiers on top of the monolingual Slovenian SloBERTa model (Ulčar & Robnik-Šikonja, 2020). SloBERTa is a pre-trained BERT-like model trained on corpora with over 3 billion tokens and a subword vocabulary of 32,000 tokens. We extended its embeddings with two different token classification heads—one for punctuation and one for capitalization—and jointly trained them on a set of 8,000,000 sentences from corpus Gigafida (Krek et al., 2020).

In the second step, each sentence is processed with the CLASSLA pipeline. CLASSLA (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023)[3] is the official fork of Stanza (the official Python NLP library of the Stanford NLP Group) for processing Slovenian, Croatian, Serbian, Macedonian, and Bulgarian. Although CLASSLA supports extensive linguistic processing (including tokenization, sentence splitting, lemmatization, part-of-speech tagging, dependency parsing, and named-entity recognition) for both standard and non-standard Slovenian, our main focus was on using CLASSLA to capitalize identified proper names. For this purpose, we used the morphosyntactic XPOS tags, which are more detailed and specific than standard POS tags and contain information that determines whether a word should be capitalized, e.g. whether the word is a noun and a proper name.

This two-step approach ensures that the final text maintains consistent and accurate punctuation and capitalization, which improves readability and the overall quality of the text.

3.4 Regular Expressions Replacement

Search and replace with regular expressions enables automatic correction of ASR output using custom dictionaries, allowing for customization of text output

[2] https://slobench.cjvt.si/leaderboard/view/10
[3] https://github.com/clarinsi/classla?tab=readme-ov-file

and increasing the flexibility and effectiveness of the platform. For example, colloquial terms can be replaced with their formal counterparts (e.g. *pršu* can be replaced with *prišel* etc.) In addition, at this stage we integrate the identification and handling of common speech recognition errors.

3.5 Numeric notation

The final step in the recognition pipeline is the conversion of text representations of numbers into their corresponding digit forms. While there are some online tools for conversion in widely used languages such as English[4], there is a lack of comprehensive solutions for less common languages such as Slovenian. In our solution, we have implemented a rule-based approach for converting Slovenian text numbers into their digit counterparts. The conversion tool uses a set of regular expressions and linguistic patterns to recognise and convert various number representations, including cardinals, ordinals, fractions, and decimals. It also takes into account common abbreviations, units of measurement, and contextual clues specific to the Slovenian language. First, the input text is subjected to tokenization and part-of-speech tagging to identify relevant number-related tokens. A series of pattern matching rules is then applied to extract and convert the numbers in the text. These rules cover a wide range of number formats and take into account the grammatical subtleties of the Slovenian language. Some examples are:

- Special handling of ”pol” (half) to correctly convert phrases like ”pol milijona” (half a million).
- Recognition of unconventional formats, such as ”dvajset dvajset” (2020) or ”devetnajststo dvajset” (1920).
- Correct conversion of decimal separators and fractions like ”dve celi pet” (2,5) or ”tri četrtine” (3/4).
- Contextual disambiguation that distinguishes between the use of ”sto” (hundred) as an independent number and as part of a larger number such as ”dvesto” (two hundred) or ”petsto” (five hundred).
- Conversion of number ranges, such as ”od pet do deset” (from five to ten).

[4] https://pypi.org/project/word2number/
https://www.atatus.com/tools/word-to-number
https://codebeautify.org/word-to-number-converter

- Handling abbreviations like ”mio” for million and ”mrd” for billion.

4 IMPLEMENTATION

The *govori.si* platform is implemented as a web application. The technology stack consists of several components:

1. Backend: We use the Django web framework together with Django Channels and Redis for handling real-time web functionality, and PostgreSQL for database management.
2. Frontend: Our frontend is developed with React and MobX libraries, providing a dynamic and responsive user interface.
3. Pipeline worker: Our worker module is developed with Python and includes the PyTorch machine learning library and models from Section 3 to support the functionalities of our application.
4. Containerization: We also use Docker for application containerization and Nginx for web server functionalities, resulting in its seamless deployment and efficient web traffic management.

To ensure the accuracy and coverage of the solution, a comprehensive set of test cases was developed. These test cases include a variety of real-world examples taken from various Slovenian texts. The solution has been extensively validated against these test cases, proving its robustness and reliability.

5 USER INTERFACE

The user interface of *govori.si* allows the user to upload and transcribe speech recordings, customising various parameters to their preferences, as shown in Figure 2. The user can specify punctuation and capitalization preferences, and toggle the options for transcribing numbers, segmenting speakers, and using dictionaries. The solution also offers three preset settings: automatic transcription, dictation with punctuation, and raw transcription.

After uploading, the recordings are processed according to the selected settings and the user can edit the transcribed text manually, as shown in Figure 3. The

Figure 2: User interface for uploading recordings and selecting parameters.

top frame provides an overview of the audio recording and allows the user to navigate through the recording and select the playback speed. The transcribed text is displayed segmented by speaker, with each speaker identified by a colored marker that can be renamed if required. The user has the option to replace all instances of a particular word and adjust the font size. Once editing of an individual speaker turn is complete, the user can lock it by clicking the check mark icon on the side. After editing, the text can be exported in various formats, including as an interview in .docx format, as plain text in .txt format, or as a subtitle-compatible .vtt file.

Interaction with the tool is simplified by a series of keyboard shortcuts for text editing and navigation through the audio recording. In addition to individual projects, users can also collaborate on joint projects.

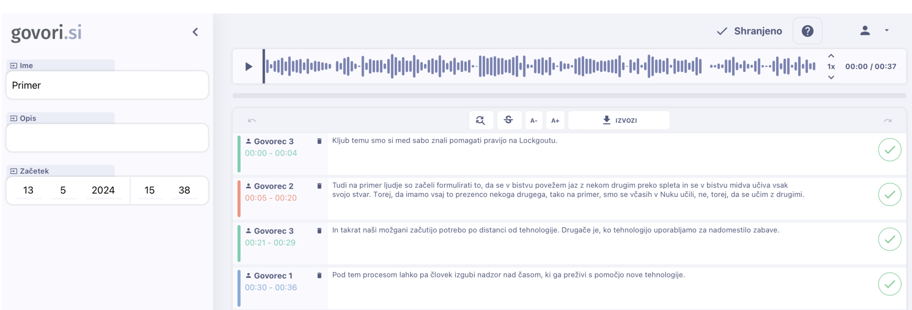


Figure 3: User interface for transcription editing.

6 EVALUATION

Our goal in developing the *govori.si* platform was to provide an easy-to-use speech recognition tool for a variety of applications, including transcription of field recordings, interviews, manual dictation in professional environments, subtitle generation, etc. We have made the tool available for research purposes and it is currently used by over 60 users from different backgrounds who have transcribed over 100 hours of recordings in the past year. It has been systematically used as part of the History of Journalism course at the Faculty of Social Sciences, where students used it to transcribe their interviews with media audiences in socialist Yugoslavia. We gathered informal feedback from 25 participants.

More than half of them emphasized the time-saving benefits of the transcription tool. Some also pointed out the ability to translate informal language into formal language, although the handling of certain dialects still needs improvement. In addition, users appreciated the availability of keyboard shortcuts and the helpful documentation, as well as the very effective processing of background noise. The export function was also praised for its user-friendliness. Overall, the tool was rated as faster, simpler, and more efficient than manual transcription methods.

However, challenges were also identified, such as difficulties with speaker segmentation, which were mentioned by more than three quarters of users. Users reported cases where the tool recognized three or more speakers when only two were involved in the conversation, as well as incorrectly tagged speakers. Some respondents also reported problems with excessive capitalization and punctuation. There were also some problems with copying and pasting text.

Suggestions for improvement included adding features such as text underlining and bolding in the editing interface, providing information on the duration of pauses in speech, and improving search functions within audio recordings. One suggested solution to the problem of speaker segmentation was the ability to specify the number of speakers in advance.

7 CONCLUSION

In this paper, we present an overview of our solution for Slovenian speech-to-text transcription. We have used state-of-the-art methods to overcome various transcription challenges. We trained our own deep speech recognition model, which is among the best current Slovenian ASR models, and integrated diarization, capitalization, punctuation, custom substitution dictionaries, and numerical notation parsing into a common platform that is available for free for non-commercial use. However, access is granted via registration credentials issued by the authors upon request. Users rated the platform very positively, although its accuracy can of course always be improved. We consider it a useful tool for a variety of use cases, including legislative processes, journalism, and research.

In the future, we aim to further improve the performance and usability of the tool. This includes improving the speaker segmentation model, integrating large language models for text summarization and correction, and continuously developing the user-friendly interface for different use cases. With all these improvements, we hope to make the *govori.si* platform a useful tool for Slovenian language processing and also contribute to advances in transcription accuracy, efficiency, and usability.

ACKNOWLEDGMENTS

This study was supported by the Slovenian Research Agency (research core funding No. P6-0436, Digital humanities: resources, tools and methods).

REFERENCES

Al-Fraihat, D., Sharrab, Y., Alzyoud, F., Qahmash, A., Tarawneh, M., & Maaita, A. (2024). Speech Recognition Utilizing Deep Learning: A Systematic Review of the Latest Developments. *Human-centric Computing and Information Sciences*, *14*, 41–76. https://doi.org/10.22967/HCIS.2024.14.015

Bredin, H. (2023). pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Interspeech 2023* (pp. 1983–1987). https://doi.org/10.21437/Interspeech.2023-105

Dobrišek, S., Žganec Gros, J., Žibert, J., Mihelič, F., & Pavešić, N. (2017). *Speech Database of Spoken Flight Information Enquiries SOFES 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1125

Gril, L., & Dobrišek, S. (2022). Akustično modeliranje z različnimi osnovnimi enotami za avtomatsko razpoznavanje slovenskega govora. In *Proceedings of Conference on Language Technologies and Digital Humanities, Ljubljana, 2022.*

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., … Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020* (pp. 5036–5040). https://doi.org/10.21437/Interspeech.2020-3015

Haberl, A., Fleiß, J., Kowald, D., & Thalmann, S. (2024). Take the aTrain. Introducing an interface for the Accessible Transcription of Interviews. *Journal of Behavioral and Experimental Finance*, *41*. https://doi.org/10.1016/j.jbef.2024.100891

Krek, S., Holdt, Š. A., Erjavec, T., Čibej, J., Repar, A., Gantar, P., … Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard slovene. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 3340–3345).

Lebar Bajec, I., Bajec, M., Bajec, Ž., & Rizvič, M. (2022). *Slovene Conformer CTC BPE E2E Automated Speech Recognition model RSDO-DS2-ASR-E2E 2.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1737

Li, B., Pang, R., Zhang, Y., Sainath, T. N., Strohman, T., Haghani, P., … Prasad, M. (2022). Massively Multilingual ASR: A Lifelong Learning Solution. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6397–6401). https://doi.org/10.1109/ICASSP43922.2022.9746594

Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 29–34). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3704

Negrão, M., & Domingues, P. (2021). SpeechToText: An open-source software for automatic detection and transcription of voice recordings in digital forensics. *Forensic Science International: Digital Investigation*, *38*. https://doi.org/10.1016/j.fsidi.2021.301223

Olev, A., & Alumäe, T. (2022). Estonian Speech Recognition and Transcription Editing Service. *Baltic Journal of Modern Computing*, *10*(3). https://doi.org/10.22364/bjmc.2022.10.3.14

Plaquet, A., & Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. In *Interspeech 2023* (pp. 3222–3226). https://doi.org/10.21437/Interspeech.2023-205

Pratap, V., Sriram, A., Tomasello, P., Hannun, A., Liptchinsky, V., Synnaeve, G., & Collobert,

R. (2020). Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters. In *Interspeech 2020* (pp. 4751–4755). https://doi.org/10.21437/Interspeech.2020-2831

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 28492–28518). https://proceedings.mlr.press/v202/radford23a.html

Terčon, L., & Ljubešić, N. (2023). *CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.* https://arxiv.org/abs/2308.04255

Tolle, H., Castro, M. d. M., Wachinger, J., Putri, A. Z., Kempf, D., Denkinger, C. M., & McMahon, S. A. (2024). From voice to ink (Vink): development and assessment of an automated, free-of-charge transcription tool. *BMC Research Notes*, *17*(1), 95. https://doi.org/10.1186/s13104-024-06749-0

True-bar 23.02 asr model. (2023). https://vitasis.si/products/truebar (Accessed: 2024-05-10)

Ulčar, M., & Robnik-Šikonja, M. (2020). *Slovenian RoBERTa contextual embeddings model: SloBERTa 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1387

Verdonik, D., Bizjak, A., Sepesy Maučec, M., Gril, L., Dobrišek, S., Križaj, J., ... Dretnik, N. (2023). *ASR database ARTUR 1.0 (transcriptions).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1772

VideoLectures.NET. (2019). *Spoken corpus Gos VideoLectures 4.0 (audio).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1222

Žgank, A., Maučec, M. S., & Verdonik, D. (2016). The SI TEDx-UM speech database: a new Slovenian Spoken Language Resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4670–4673). https://aclanthology.org/L16-1740

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., ... Wu, Y. (2023). Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. *ArXiv*. https://arxiv.org/abs/2303.01037

Žitnik, S., & Dragar, F. (2021). *SloBENCH evaluation framework.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1469

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., & Erjavec, T. (2021). *Spoken corpus Gos 1.1.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1438

PLATFORMA ZA TRANSKRIPCIJO GOVORA GOVORI.SI

Kljub velikemu napredku pri razvoju samodejne razpoznave govora je odprtokodnih integriranih rešitev z uporabniku prijaznimi grafičnimi vmesniki še vedno zelo malo. To vrzel naslavljamo z razvojem slovenskemu jeziku prilagojene platforme *govori.si* za transkripcijo govora. Razvili smo nov model za samodejno razpoznavo slovenskega govora, ki se trenutno uvršča med najboljše tovrstne modele in uporabili druge sodobne pristope k reševanju izzivov pri transkripciji. Platforma *govori.si* združuje prepoznavo govora s segmentacijo govorcev, samodejnim določanjem velikih začetnic in ločil, uporabniško definiranimi slovarji in logiko za zapisovanje števil. Za raziskovalne in nekomercialne namene je platforma prosto dostopna. Med uporabniki je bila pozitivno sprejeta in postaja dobrodošlo orodje za uporabo v zakonodajnih postopkih, novinarstvu in raziskavah.

Keywords: slovenski jezik, samodejna razpoznava govora, obdelava naravnega jezika

GENDER IDEOLOGY: A CORPUS-LINGUISTIC LOOK AT EMERGENT 'ANTI-GENDER' VOCABULARY IN SLOVENIA, CROATIA AND SERBIA

Ksenija BOGETIĆ,[1,2] Vojko GORJANC,[1,3] Jure SKUBIC,[1,3] Alenka KAVČIČ[4]

[1]Institute of Contemporary History
[2]Research Centre of the Slovenian Academy of Sciences and Arts
[3]Faculty of Arts, University of Ljubljana
[4]Faculty of Computer and Information Science, University of Ljubljana

1 INTRODUCTION

The concept of *gender ideology* has been observed as a rising and semiotically shifting linguistic innovation in public discourses across the globe (Borba, 2022; Kuhar & Zobec, 2018; Baran, 2022). Essentially a co-optation of feminist and scholarly terminology by the populist right (Graff, 2020) used to denote a threat to traditional gender/social relations, the phrase has come across issues in theoretical interpretation, leaving linguists to often resort to describing it as an 'empty signifier' with fluid meaning (Gal, forthcoming). Empirical analyses of the phrase's usage in different language contexts are yet to provide a deeper picture of the term's social meanings.

The phrase itself offers an important point of departure for understanding the discourse underpinning the recent rise of the so-called 'anti-gender movements' (Kuhar & Patternotte, 2017), which define themselves as transnational, anti-feminist and anti-LGBTQ social mobilisations against *gender ideology* (Corredor, 2019). Beyond a merely populist term of the far-right, the idea of *gender ideology* as a threat or conspiracy has permeated important political discussions on matters such as reproductive rights and violence over women, providing legitimization for anti-feminist positionings and trivialization of gender-related inequalities. Understanding its use in more 'official' contexts of political debate – such as those of national parliaments – especially as the concept is in many parts of the world yet emerging, is thus an important prerequisite to engaging with anti-gender motions in meaningful ways.

Corpus data can in this regard provide important insights into processes of semiotic resignification by pointing to specific patterns of usage and association. Multilingual corpora as in the case of the ParlaMint project (Erjavec et al., 2023) are particularly useful for cross-linguistic and cross-national analyses. In particular, such resources may thus help us further understand the circulation of global (anti-gender or other) discourses across geopolitical space (Gal, 2021), and across social settings, among others by allowing comparisons of more general usage and adoption in parliamentary or political discourse in particular.

As a step in this direction, in this talk we will present the first preliminary results of a wider study on the concept of *gender ideology*, looking at three former-Yugoslav states: Serbia, Croatia and Slovenia.

2 BACKGROUND AND SOCIAL CONTEXT

The rise of the so-called anti-gender mobilizations (Kuhar & Paternotte, 2017), has brought forth a range of disturbing events in Europe and beyond, ranging from new motions to limit abortion rights, through bannings of gender studies from curricula, to new waves of violence against LGBTQ+ persons and organizations. Most simply, the notion of anti-genderism is used to refer to transnational social mobilisations against feminism and gender equality, built around the notion of *gender ideology* as a new threat to tradition, nation and the family (Corredor, 2019; Bogetić 2022). The targets of anti-genderism can differ in scope, and are molded to local national concerns, but overall span concepts of LGBTQ+ rights, reproductive rights, sex and gender education, gender studies, and democracy in a broadest sense (Kuhar & Paternotte, 2017).

Gender ideology is a comparatively recent coinage rising to prominence in the 21st century (Kuhar & Paternotte, 2017). Spreading across languages of central and eastern Europe like in many others, the phrase itself, based in a linguistic, semiotic transformation of both gender and ideology is now used to refer to an imported (Western) threat to social relations (Graff, 2020). It has been noted in similar form in different countries and languages across this space, as e.g. *gendernaia ideologiia* in Russia, *ideologija džendar* in Bulgaria – but also sometimes with differing realizations, as with *teorija spola* and *ideologija spola* as both existing in use in Slovenia (Popič & Gorjanc, 2023) – and as such has become a central point in local populist rhetorics, symbolizing perceptions of social decline stemming from 'liberal' or 'leftist' ideas. Its basis lies in the conviction that society in Western Europe faces imminent downfall because fundamental values like the family, homeland, and nation are no longer being upheld (Verseck, 2018). This in turn implies a need for attention and protection by mobilizing national and/or religious resources.

However, as Paternotte and Kuhar (2017) warned, the automatic entanglement of anti-gender campaigns to the far-right populism might miss on many important regional and national specificities, and the same goes for understanding the language of anti-genderism. Though the term '*gender ideology'* seems to feature in many Slavic-language states in the region, alternative phrases appear in parallel use, and also carry somewhat different connotations, and may also occur in different forms in different public contexts (such as media texts and parliamentary debates). Generally, the area of eastern and central Europe has often been described as a hub of anti-genderism, seen especially with recent developments in Poland and Hungary (Graff, 2020; Peto 2020); other post-socialist areas such as the former Yugoslav space, where gender is long argued to hold a key symbolic role in political life (Bonfiglioli et al. 2015), are less explored.

Also, as some scholars in the post-Yugoslav area have noted the emergence of anti-gender discourses and vocabulary in debates in national parliaments (cf. Kuhar, 2020), parliamentary discourse becomes an important and under-represented discourse setting in which to observe the adoption of the anti-genderist shibboleth of *gender ideology*. This is where recent corpus developments in multilingual, cross-country parliament datasets provide an invaluable point of analysis of language of the parliaments, and discourses of politics and gender. Tracing emergent concepts also allows us to observe some potentials, limitations and challenges in bringing corpus-based analyses together with critical analyses of discourse.

3 AIMS

In this study we aim to investigate the use of the concept of *gender ideology* in Serbian, Croatian and Slovene. Indirectly, we also aim to use the analysis to obtain insights into aspects of anti-genderist rhetoric in the three states of interest. In our preliminary analysis, we find it productive to compare the uses of the term in a web corpus of more general scope and in a corpus of parliamentary discourse, as well as to look across three national contexts once belonging to the same state, now taking some differing political and EU-related positionings. Specifically, our research questions include:

RQ1: How may the frequencies of occurrence compare across the countries and corpus/discourse types?

RQ2: What sets of social/political concepts (sexuality, religion, political affiliation, etc.) may be observed among the top collocates of *'gender ideolog*y' in general and parliamentary discourse?

The multiple levels of comparison would of course warrant further and carefully controlled corpus analyses. At this point, we focus on the broadest picture of frequencies and collocation, while finding that a combination of corpus-based analysis and analysis of concordances from a discursive angle is telling of social meanings (see Methodology).

Zooming in on parliamentary discourse as our particular point of interest, we explore two further questions:

RQ 3: As an essentially anti-feminist term, can differences be observed in the distribution of the use of *gender ideology* between male and female MPs?

RQ 4: What may sentiment analysis of the statements on *gender ideology* tell us about the phrase's use, in terms of negative/positive/neutral attitudes?

4 METHODOLOGY AND CORPORA

In our analysis, we draw on two main corpus resources. The first involves CLASSLA-web, as massive web corpora for Slovenian (CLASSLA-web.sl), Croatian (CLASSLA-web.hr) and Serbian (CLASSLA-web.sr), each comprising around 2 billion words, and including texts published up to 2021 (CLASSLA-web.sl) and 2022 (CLASSLA-web.hr and CLASSLA-web.sr). Collected by crawling primarily the national top-level internet domains (see https://www.clarin.si/info/k-centre/classla-web-bigger-and-better-web-corpora-for-croatian-serbian-and-slovenian-on-clarin-si-concordancers/), CLASSLA-web is aimed to encompass all the texts written on the web, and thus allows large-scale insights into contemporary language use. Second, we use ParlaMint 4.0 (Erjavec et al., 2023), a multilingual set of comparable corpora containing parliamentary debates from 29 parliaments and covering at least the period from 2015 to 2022. For this analysis, we use ParlaMint's latest versions of parliamentary corpora of Serbian, Croatian and Slovene (ParlaMint-SR 4.0, ParlaMint-HR 4.0, ParlaMint-SI 4.0); given the somewhat differing corpus timespans (e.g. going back to 1997 in SR, and late 2003 in HR), we select the 2004-2022 timespan for our analysis.

Using the noSketchEngine concordancer, we perform a search of the node 'language ideology' in the three language corpora (i.e. rodna ideologija Sr / rodna ideologija Hr / ideologija spola / teorija spola Slo), namely CLASSLA-web-sr, CLASSLA-web-hr, CLASSLA-web-slo, and ParlaMint-SR 4.0, ParlaMint-HR 4.0, and ParlaMint-SI 4.0. Collocation analysis (McEnery and Hardie 2011), is performed with the 5L-5R node span (cf. Baker 2010 on discourse-oriented analyses), and in SketchEngine calculated with the logDice measure. This allows comparisons across the corpus/discourse type (general/Web and Parliaments), and across the three states. Qualitative analysis of the concordance liness for the top-ranking collocates provides more insights into these uses, in line with our perspective on combining the corpus-based approach with contextually situated (critical) discourse analysis, while the perspectives can be additionally enriched with sentiment analysis.

5 FINDINGS OVERVIEW

Looking at the CLASSLA-web corpora, the phrase *gender ideology* is found to have different levels of prominence in the three national/language contexts examined. In Croatia, *rodna ideologija* can be seen to represent an established term, with several thousand occurrences across different genres (2.2 hits per mill. tokens). By contrast, in the Serbian corpus, *rodna ideologija* is far less infrequent, mostly belonging to scholarly discourse; it appears to be a concept only emerging in the past several years, in relation to educational and gender-sensitive language debates in particular (0.1 hits per mill. tokens). In Slovenia, our analysis confirms that two variants exist in use – '*teorija spola*' and '*ideologija spola*' (cf. Popič & Gorjanc, 2020) – with the first strongly preferred (0.6 hits per mill. tokens, over a thousand occurrences), and the latter more sporadic (less than 0.1 hits per mill. tokens). The social meanings and associations of the phrase become clearer in collocation analysis.

While notable similarities can be observed across the collocate lists for the three sub-corpora, in line with the more global anti-genderist discourses, we can also observe interesting differences when looking at the top collocates. For Croatia, for example, these include a more procedural focus on the Instanbul Convention and its ratification, and varied sexuality related terms (*LGBT*, *LGTB*, *homoseksualnost*); in Slovenia, the top collocates refer to activism, with a more specific political focus on 'leftism' and 'cultural Marxism', in line with some recent observations on the role of memory politics in anti-gender discourses in the country (Kuhar & Shevtsova, forthcoming). For Serbia, we note a mix of academic terms, and those suggesting colonization and outside threat to nation and tradition, whose meanings become clearer in discourse-based concordance analysis.

Turning to the ParlaMint corpus, we find that the phrase has entered parliamentary usage, however at very different levels in the differing countries (again by far most prominent in ParlaMint-HR, and almost entirely absent in ParlaMint-SR). Collocation analysis points to some differing foci, which also show differences from the CLASSLA-web general discourse; still, we note some limitations in tracing and comparing the meanings of emergent terms such as the one in question. For Slovenian, thus, the majority of the collocates are grammatical/function words, of little use for interpreting social meanings. Further, using the affordances of SketchEngine and ParlaMint metadata, our investigation into the gendered aspects of the use of our phrase of interest shows a general preference for the term to be used by male MPs over the female MPs. Sentiment analysis, finally, gives information on the attitudes surrounding the usage of the term, which are unsurprisingly negative; still, evaluating the results we also note some issues in automatic sentiment assignment which may be especially problematic in corpus analyses of emergent terms, coloured by significant levels of irony and sarcasm.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support from the Slovenian Research and Innovation Agency, research core funding P6-0436: Digital Humanities: Resources, Tools and Methods and the EU Commission Horizon 2020 WIDERA Research Project CRISMET, Grant agreement ID: 101038047.

REFERENCES

Baran, D. (2022). 'Rainbow plague' or 'rainbow allies'? TĘCZA 'rainbow' as a floating signifier in the contestation of Poland's national identity. Gender & Language, 16(3).

Bogetić, K. (2022). Politics of resignification: central and eastern European perspectives on language and gender. *Gender & Language*, 16(3).

Bonfiglioli, C., Kahlina, K., & Zaharijević, A. (2015). Transformations of gender, sexuality and citizenship in South East Europe. In Women's Studies International Forum, 49, 43-47. Pergamon.

Borba, R. (2022). Enregistering 'gender ideology': the emergence and circulation of a transnational anti-gender language. Journal of Language and Sexuality 11(1), 57–79.

Corredor, E. (2019). Unpacking 'gender ideology' and the global right's antigender countermovement. Signs: Journal of Women in Culture and Society 44(3), 613–638.

Erjavec, T.; et al., (2023), Multilingual comparable corpora of parliamentary debates ParlaMint 4.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1859.

Gal, S. (2021). Gender and the discursive authority of far right politics. Gender and Language 15(1), 96–103.

Ljubešić, N.; Rupnik, P. and Kuzman, T. (2024). Slovenian web corpus CLASSLA-web.sl 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1882.

Ljubešić, N.; Rupnik, P. and Kuzman, T. (2024). Croatian web corpus CLASSLA-web.hr 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1929.

Ljubešić, N.; Rupnik, P. and Kuzman, T. (2024). Serbian web corpus CLASSLA-web.sr 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1931.

Kuhar, R., & Paternotte, D. (eds.) (2017). Anti-gender Campaigns in Europe: Mobilizing against Equality. Lanham: Rowman & Littlefield.

Kuhar, R., & Zobec, A. (2017). The anti-gender movement in Europe and the educational process in public schools. CEPS journal, 7(2), 29-46.

Popič, D., & Gorjanc, V. (2023). Corpus-Linguistic Analysis of Speech Communities on Anti-Gender Discourse in Slovene. Gender a výzkum/Gender and Research, 23(2), 140-166.

EXPANDING THE EUROPEAN PARLIAMENT TRANSLATION AND INTERPRETING CORPUS: A MODULAR PIPELINE FOR THE CONSTRUCTION OF COMPLEX CORPORA

Alice FEDOTOVA,[1] Adriano FERRARESI,[1] Maja MILIČEVIĆ PETROVIĆ,[1] Alberto BARRÓN-CEDEÑO[1]
[1]University of Bologna

1 INTRODUCTION

The present paper introduces an expanded version of the European Parliament Translation and Interpreting Corpus (EPTIC), a multimodal parallel corpus comprising speeches delivered at the European Parliament along with their official interpretations and translations (see Bernardini et al., 2016; Bernardini et al., 2018). Constructing multimodal and parallel corpora for translation and interpreting studies (TIS) has been acknowledged as a "formidable task" (Bernardini et al., 2018), which – if automated, as we propose – involves a number of subtasks such as automatic speech recognition (ASR), multilingual sentence alignment, and forced alignment, each of which poses its own challenges. Yet tackling these subtasks also offers a unique way to evaluate state-of-the-art natural language processing (NLP) tools against a unique, multilingual benchmark. In this paper we discuss the development of a modular pipeline adaptable for each of these subtasks and address the broader implications of this work for the field of corpus construction.

While multilingual sentence alignment is particularly relevant for translation and interpreting corpora, transcriptions of spontaneous or planned speech aligned with recordings are also essential for linguistic research more broadly, in particular for spoken corpora used in phonology, conversational analysis, dialectology and so forth (Lemmenmeier-Batinić, 2023). However, the adoption of NLP tools for corpus construction has often lagged behind due to technological hesitancy and tool requirements, i.e. lack of re-use of tools developed for specific corpus construction needs, which impeded interoperability.

To address these challenges, this paper aims to raise awareness of the software available for the purposes of corpus construction, create reusable resources, and facilitate the adoption of NLP tools for researchers interested in speech transcription, sentence alignment, and more generally multimodal corpora. We highlight the potential benefits of automatic alignment and transcription for different types of corpora, elaborating on the increasing interest in tools such as OpenAI's Whisper (Radford, 2022) and their suitability for linguistic research. We find that satisfactory results can be achieved with ASR, although challenges remain, especially with regards to the verbatimness of the transcription, where by verbatimness we mean the level of detail where all words are transcribed, along with disfluencies and some extra-linguistic information (Wollin-Giering, 2023). Sentence alignment can be facilitated through state-of-the-art embedding-based tools, whereas forced alignment can be considered a largely solved problem. This makes the construction of EPTIC more streamlined and requiring less human intervention, with wider implications for multilingual corpus construction in the field of TIS and beyond.

2 EPTIC 1.0 AND TOOLS FOR ITS EXPANSION

Within EPTIC, the corpus construction process revolves around individual speech events, where edited "verbatim" reports published by the European Parliament and transcriptions of the speeches are accompanied by transcriptions of interpretations and official translations. Figure 1 uses parallel boxes to represent, both vertically and horizontally, different facets of the same events (source or target, written or spoken). Empty space in the English subcorpus represents the potential for more languages to be added. Corpora containing translations in both directions (e.g., from English to French and from French to English) are referred to as bidirectional, while those with translations in only one direction are referred to as unidirectional.

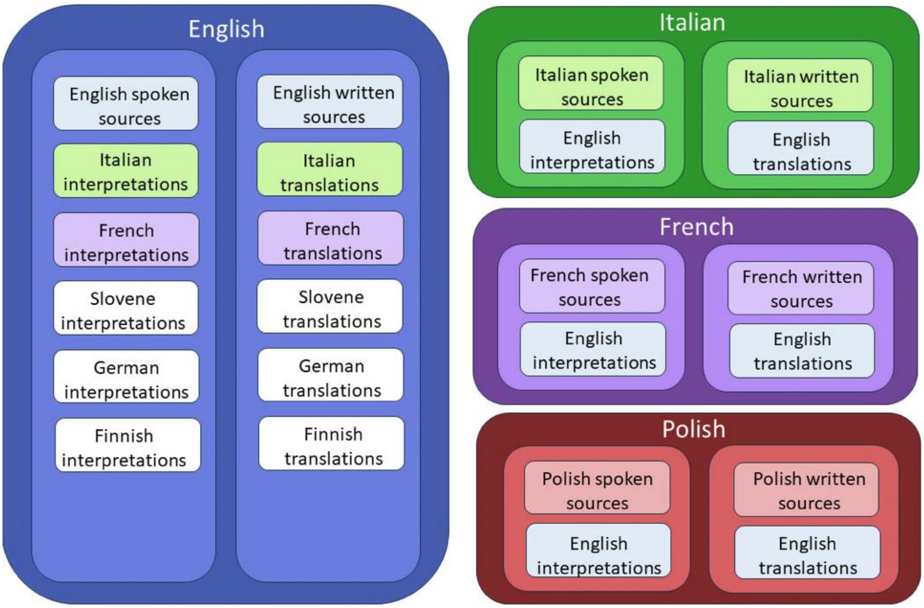


Figure 1: The structure of EPTIC.

The languages included and the size of the previously published EPTIC corpus, before the expansion discussed in the present work, are shown in Table 1.

Our approach to corpus expansion began with a review of previous guidelines for developing the EPTIC corpus (Bernardini et al., 2018; Kajzer-Wietrzny and Ferraresi, 2020). The first step in the construction of the previous version involved obtaining transcripts, verbatim reports, translations, and interpretations of European Parliament speeches from the official website.[1] Transcripts of the original speeches and interpretations were manually adapted following editing conventions to annotate features of orality such as disfluencies and timestamped using Aegisub. Then, the texts were automatically segmented into sentences and aligned across languages using the Intertext Editor alignment tool.

The creation of the new workflow started with the previous procedure as a basis. It was first subdivided into separate tasks, the main ones being

[1] https://www.europarl.europa.eu/plenary/en/debates-video.html

Table 1: Token counts, by language, of the previously published version of EPTIC.

	Sources		**Targets**	
	Spoken	*Written*	*Interpretations*	*Translations*
English	24,136	22,782	53,615	58,561
French	27,713	26,674	23,185	25,855
Italian	20,016	19,591	20,352	23,234
Polish	11,011	10,616	*TBA*	*TBA*

automatic speech recognition, multilingual sentence alignment, and forced alignment. Software selection was based on criteria such as ease of use and installation, compatibility with the Python programming language, linguistic coverage, and compatibility with Sketch Engine, an established corpus query tool for teaching and research (Rychlý, 2007; Kilgarriff, 2014). Python version 3.11.5 was used along with the Poetry[2] package manager for portability. The resulting workflow is planned for release by the end of 2024 as an open-source GitHub repository and a Python package installable via pip.[3] The following paragraphs discuss the tasks and the considerations made in order to design a new pipeline for EPTIC.

Automatic Speech Recognition has seen recent advancements, with the introduction of Whisper (Radford, 2022) and Wav2Vec 2.0 (Baevski, 2020). However, achieving a reasonable level of transcription quality is complex and context-dependent, as it can be interpreted and evaluated differently depending on the specific domain, task, and application (Kuhn et al., 2024). For EPTIC, we require an ASR system to produce a verbatim transcription where all words are transcribed, along with disfluencies and some extra-linguistic information. "Verbatimness" is, however, also a broad concept (Wollin-Giering, 2023), given the variety of transcription conventions existing in the field of linguistics, and Whisper has been observed to produce transcripts "often almost comparable to the final read through of a manual (verbatim to gisted) transcript" (Wollin-Giering, 2023). We further explore this

[2] https://python-poetry.org/
[3] https://pypi.org/project/pip/

claim by testing a variant of Whisper, WhisperX,[4] on our data. Given its exceptional performance in long-form transcription (Bain et al., 2023), we hypothesize that WhisperX could be especially beneficial when dealing with parliamentary speeches. In this regard, EPTIC serves as a challenging and unique benchmark due to its multilingual nature, inclusion of non-native speech, and interpreted text, which has been found to be particularly difficult to transcribe (Wang and Wang, 2024).

Sentence Alignment involves identifying and aligning parallel sentences, both monolingually and multilingually. For this task, we used Bertalign (Liu & Zhu, 2022), a tool for aligning parallel corpora based on sentence embeddings. Unlike predecessors like Hunalign[5] that rely on lexical translation probabilities, Bertalign employs sentence embeddings to identify parallel sentences, providing a more robust approach for handling semantic similarities across languages.[6] Some changes were necessary as the default settings were not appropriate out-of-the-box. For instance, we were required to change the value of the variable is_split to False, as the corpus was already sentence-split in a previous step. The tool produces alignments in the format of a list of tuples, and it has been extended with an additional Python script to convert its output into the Sketch Engine alignment format based on corpus-internal indexing.

3 EPTIC 2.0: PIPELINE AND CORPUS PROPERTIES

The Python pipeline, aimed at facilitating the expansion of the EPTIC corpus, has been structured in a modular fashion. This process begins with the extraction of text and video data, either manually or through the use of ad-hoc scripts, depending on the amount of data that the researcher intends to add.[7]

---

[4] https://github.com/m-bain/whisperX

[5] https://github.com/danielvarga/hunalign

[6] The settings for Bertalign are not documented, but information about the code is available as part of comments in its aligner.py script, which is provided in a GitHub repository.

[7] Download of the video through the European Parliament’s interface can be hindered by the lengthy process, requiring a personal e-mail address which is then used to obtain a URL for the purpose.

Transcription is then performed using WhisperX, which concurrently provides timestamp information. To remove mistranscriptions and to ensure adherence to the transcription guidelines, the transcripts undergo manual review to incorporate disfluencies and rectify potential mistranscriptions. Ongoing research is being conducted to determine whether official transcripts can be leveraged to enhance the quality of the transcriptions in terms of word recognition, and whether it is possible to include disfluency markers such as hesitations, false starts, and repetitions.

Once all required texts have been transcribed, they undergo sentence splitting and sentence alignment using Bertalign. Subsequently, relevant metadata, encompassing session topics, are automatically retrieved from the European Parliament website. The only metadata item requiring manual input is the speech type, which can be defined as impromptu, read out, or mixed if both delivery types are present. After exporting the alignments in the Intertext format[8] and performing Part-of-Speech tagging with Sketch Engine, the texts and metadata are converted to the .vert format, rendering them ready for indexing in Sketch Engine (Rychlý, 2007; Kilgarriff, 2014). The following part-of-speech taggers were used to annotate the texts within Sketch Engine: MULTEXT-East Slovenian (version 4)[9] for Slovene, Polish NKJP[10] for Polish, FreeLing[11] for Italian and French, TreeTagger[12] for Finnish and English, and German RFTagger[13] for German.

We now highlight the substantial expansion of the EPTIC corpus across different languages and all subcorpus types, as compared to the original corpus presented in Table 1. Table 2 shows the size, in tokens, of the updated bidirectional English, French, Italian subcorpora, with the English subcorpus exhibiting the largest increases. The Polish unidirectional subcorpus has not yet received an expansion, though text-to-video alignments were added as an

[8] https://wanthalf.saga.cz/intertext
[9] http://nl.ijs.si/ME/V6/msd/html/msd-sl.html
[10] https://nkjp.pl/poliqarp/help/ense2.html
[11] https://freeling-user-manual.readthedocs.io/en/latest/tagsets/
[12] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[13] https://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/

Table 2: Token counts, by language, of the expanded version of EPTIC.

	Sources		**Targets**	
	Spoken	*Written*	*Interpretations*	*Translations*
English	43,138	41,047	55,109	58,651
French	35,648	34,063	31,935	35,566
Italian	21,208	20,646	27,329	31,816
Polish	9,458	9,193	*TBA*	*TBA*
Slovene	*TBA*	*TBA*	19,717	22,476
German	*TBA*	*TBA*	18,258	19,822
Finnish	*TBA*	*TBA*	11,624	12,045

additional feature. Furthermore, new additions to EPTIC include the unidirectional Slovenian, German, and Finnish subcorpora, with the Slovenian interpretations (19,717 words) and translations (22,476 words) representing the largest target subcorpus to be added as part of this update.

WhisperX has demonstrated robust performance across diverse speech types, languages, and accents. Consistent with prior research findings (Wollin-Giering, 2023), it does not provide verbatim transcriptions, though the word error rate (WER) metric alone may be insufficient to fully ascertain the extent of this limitation.[14] We are currently investigating additional evaluation metrics and conducting qualitative analyses to comprehensively assess WhisperX's transcription capabilities. The results in Table 3 indicate that good performance can be achieved for certain languages, such as French, which exhibits a low WER of 0.118, while performance degrades for Slovenian. Furthermore, we evaluated WhisperX on a subset of English speech data to examine whether factors such as speaker nativeness or interpreted speech influence WER. Our findings indicate a WER of 0.104 for native English speakers, 0.110 for non-native speakers, and a notably higher WER of 0.222 for interpreted speech. This is consistent with Wang and Wang (2024), where the higher WER in interpreted speech is attributed to the increased presence of disfluencies such as filled and unfilled pauses and the challenges posed by

[14] WER evaluates the accuracy of transcribed text compared to the ground truth, with lower WER indicating better performance.

mispronunciations, which are more prevalent in interpreted speech, leading to greater difficulties in accurate ASR transcription (see Table 4).

Table 3: Performance of WhisperX by language, expressed in WER.

	English WER	**Italian WER**	**French WER**	**Slovenian WER**
Whisper small	0.212	0.219	0.162	0.463
Whisper medium	*0.196*	*0.173*	0.213	0.327
Whisper large-v2	*0.194*	*0.152*	0.118	0.262

The task of aligning transcriptions with their corresponding audio files across multiple languages and subcorpora within EPTIC has formerly been carried out manually, a laborious and error-prone process, especially at a large scale. Leveraging Bertalign appears to be effective in tackling this complex challenge. This allowed for the introduction of automatic alignment across all subcorpora within EPTIC – a significant improvement over the previous version of the corpus, where text-to-text alignments were often absent due to the manual efforts required.

4 CONCLUSIONS AND FUTURE WORK

In conclusion, the development of a state-of-the-art, NLP-based modular pipeline has resulted in a significant expansion of the European Parliament Translation and Interpreting Corpus. The resulting corpus features a significant increase in size and language coverage compared to the previous version of EPTIC (Bernardini et al., 2016; Bernardini et al., 2018), making it a more comprehensive and valuable resource for research in TIS, as well as related fields such as linguistics and NLP. The development of this pipeline has demonstrated the potential for automating various aspects of corpus construction by including ASR, multilingual sentence alignment, and forced alignment. By integrating tools like WhisperX and Bertalign, the process of transcribing, aligning, and timestamping the audio-video data has been streamlined, reducing the time and effort required for manual intervention.

Table 4: Disfluency types and transcription errors made by Whisper large-v2.

	Human transcription	**Whisper large-v2**
Contraction	**I'm** encouraged that the interim leadership has promised substantial reforms because embarking on such a path will greatly strengthen Tunisia's relationship with the European Union.	**I am** encouraged that the interim leadership promised substantial reforms because embarking on such a path will greatly strengthen Tunisia's relationship with the European Union.
Truncation	E.g., polygamy was banned, veils were banned, foreign direct **in-** ehm **investment** was encouraged, tourism was encouraged.	E.g. polygamy was banned, veils were banned, foreign direct **investment** was encouraged, tourism was encouraged.
Discourse marker	Presidente, la sommossa in Tunisia è senz'altro un riflesso della frustrazione della gente di fronte alla situazione politica, poi c'è anche la corruzione la la con- e la conduzione della famiglia regnante **diciamo**.	La sommossa in Tunisia è senz'altro un riflesso della frustrazione della gente di fronte alla situazione politica, poi c'è anche la corruzione e la conduzione della famiglia regnante.
Filled pause	**Ehm** importanti sono le riforme… solo questo potrà rinforzare le relazione con la Tunisia.	Importanti sono le riforme solo questo potrà rinforzare le relazioni con la Tunisia.
Empty pause	Il Parlamento europeo deve condannare queste azioni che **…** rivelano il volto opprimente e aggressivo della Turchia a Cipro.	Il Parlamento europeo deve condannare queste azioni che rivelano il volto frimente e aggressivo della Turchia a Cipro.

Limitations remain, such as the relatively small overall corpus size and the challenges in evaluating punctuation prediction by WhisperX, i.e. how closely it aligns with the sentence boundaries that a human would conceive of as natural and appropriate for a given spoken content. Additionally, ongoing work includes conducting a more thorough evaluation of the tools' performance using metrics including the F1 score in the case of sentence alignment, as well as experimenting with methods aimed at improving ASR performance, for instance by fine-tuning Whisper on our transcriptions. The inclusion of automatic alignment between all subcorpora in EPTIC, facilitated by Bertalign's robust performance, represents a significant advancement, enabling analyses and comparisons that were previously impractical or impossible due to the substantial manual effort required.

Looking ahead, future work could involve adding numerical features to represent prosodic aspects of speech, which could enable more research avenues leveraging the spoken EPTIC data. These numerical features, such as mean pitch, pitch range, intensity contours, duration of speech segments, voice quality measures like jitter and shimmer, and spectral characteristics, could be obtained using acoustic analysis tools like Praat (Boersma & Weenink, 2024). Additionally, the release of the Python-based software will provide a valuable resource for other researchers and corpus builders working with the Sketch Engine platform or similar tools. By lowering the barrier to entry and increasing the efficiency of multimodal corpus development, this work paves the way for more comprehensive and representative language resources, ultimately driving progress in various domains of linguistic and computational research.

ACKNOWLEDGMENTS

The work of Alice Fedotova is supported by the NextGeneration EU programme, ALMArie CURIE 2021 - Linea SUpER, Ref. CUPJ45F21001470005.

REFERENCES

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460. Retrieved May 19, 2024, from https://10.48550/arXiv.2006.11477

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. *arXiv preprint.* Retrieved May 20, 2024, from https://arxiv.org/pdf/2303.00747

Bernardini, S., Ferraresi, A., Russo, M., Collard, C., & Defrancq, B. (2018). Building interpreting and intermodal corpora: A how-to for a formidable task. In M. Russo, C. Bendazzoli & B. Defrancq (Eds) *Making Way in Corpus-based Interpreting Studies* (pp. 21-42). Springer, Singapore. https://doi.org/10.1007/978-981-10-6199-8_2

Bernardini, S., Ferraresi, A., & Miličević, M. (2016). From EPIC to EPTIC—Exploring simplification in interpreting and translation from an intermodal perspective. *Target*, *28*(1), 61-86. https://10.1075/target.28.1.03ber

Boersma, P. & Weenink, D. (2024). Praat: Doing Phonetics by Computer [Computer program]. Version 6.4.12. Retrieved May 14, 2024, from http://www.praat.org/

Della Corte, G. (2020). *Text and Speech Alignment Methods for Speech Translation Corpora Creation: Augmenting English LibriVox Recordings with Italian Textual Translations* [Master's thesis]. Retrieved May 14, 2024, from http://www.diva-portal.org/smash/get/diva2:1440026/FULLTEXT01.pdf

Jones, C., Li, W., Almeida, A., & German, A. (2019). Evaluating cross-linguistic forced alignment of conversational data in North Australian Kriol, an under-resourced language. *Language Documentation and Conservation*, *13*, 281-299. Retrieved May 15, 2024, from http://hdl.handle.net/10125/24869

Kajzer-Wietrzny, M. & Ferraresi, A. (2020). *Guidelines for EPTIC collaborators.* Unpublished manuscript.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2024). Measuring the accuracy of Automatic Speech Recognition solutions. *ACM Transactions on Accessible Computing, 16*(4), 1-23. https://doi.org/10.1145/3636513

Lei, L. & Zhu M. (2022). Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities, 38*(2), 621-634. https://doi.org/10.1093/llc/fqac089

Lemmenmeier, D. (2023). *Spoken Language Corpora: Approaches for Facilitating Linguistic Research* [Doctoral dissertation]. Retrieved May 13, 2024, from https://www.zora.uzh.ch/id/eprint/235310/

Pettarin, A. (2018). Forced-alignment-tools [Computer program]. Version 1.0.9. Retrieved May 16, 2024, from https://github.com/pettarin/forced-alignment-tools

Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C. & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning* (pp. 28492-28518). Retrieved May 13, 2024, from https://proceedings.mlr.press/v202/radford23a.html

Rychlý, P. (2007). Manatee/Bonito–A modular corpus manager. *RASLAN*, 65-70. Retrieved May 14, 2024, from https://www.sketchengine.eu/wp-content/uploads/Manatee-Bonito_2007.pdf

Wang, X., & Wang, B. (2024). Exploring automatic methods for the construction of multimodal interpreting corpora. How to transcribe linguistic information and identify paralinguistic properties? *Across Languages and Cultures.* Retrieved 14 May, 2024, from https://scholarspace.manoa.hawaii.edu/items/406b1b2a-8aab-4f57-a44f-d9c736d5e655

Wollin-Giering, S., Hoffmann, M., Höfting, J., & Ventzke, C. (2023). Automatic transcription of qualitative interviews. *Sociology of Science Discussion Papers*. https://10.13140/RG.2.2.14480.38404

Wu, H., Yun, J., Li, X., Huang, H., & Liu, C. (2023). Using a forced aligner for prosody research. *Humanities and Social Sciences Communications, 10*(1), 1-13. https://doi.org/10.1057/s41599-023-01931-4

SLOVENSKI TERMINOLOŠKI PORTAL

Mateja JEMEC TOMAZIN
Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti, Inštitut za slovenski jezik Frana Ramovša

1 UVOD

V okviru projekta Razvoj slovenščine v digitalnem okolju (v nadaljevanju RSDO) je bil razvit tudi Slovenski terminološki portal, dostopen na URL https://terminoloski.slovenscina.eu/, ki združuje več funkcionalnosti in je namenjen tako tistim uporabnikom, ki želijo sami urejati terminologijo, kot tudi tistim, ki bi radi na enem mestu iskali in preverjali terminološke odločitve. Uporabnikom sta na voljo slovenski in angleški vmesnik, omogočena je tudi povezava z drugimi terminološkimi portali. Pri oblikovanju so bile upoštevane prilagoditve za osebe, ki imajo težave z vidom. Na portalu deluje luščilnik terminoloških kandidatov iz povezanega korpusa znanstvenih besedil Open Science Slovenia http://hdl.handle.net/11356/1774 (v nadaljevanju korpusa OSS), integriranega tudi v konkordančnika noSketch in KonText ter uporabnikovih lastnih besedil. Osnovni iskalnik poleg zadetkov v slovarjih ponuja tudi zadetke med terminološkimi odgovori. Celotna koda je za samostojno namestitev in prilagoditev dostopna na portalu GitHub https://github.com/clarinsi/rsdo_term_portal.

2 TERMINOLOŠKI PORTAL IN RSDO

Projekt RSDO je natančneje predstavljen v monografiji Arhar Holdt in Krek, ur., (2023). Glavni cilj projekta je bil strateško načrtovanje razvoja in tudi izdelava jezikovnih virov, od korpusov, govornih tehnologij do obogatitve leksikalne baze in izboljšanja prevajalnikov tako v slovenščino kot v druge jezike, čeprav je bila v ospredju angleščina. Na področju terminologije je bilo treba ugotoviti, kateri so primarni uporabniki terminologije in kakšne potrebe imajo. Strokovna področja za nove pojme uveljavijo nove termine, vendar pa terminološki viri ne nastajajo dovolj hitro, da bi se ta poimenovanja lahko hitro razširila zunaj kroga specializiranih uporabnikov, včasih celo zunaj posameznih raziskovalnih skupin. Specializirani uporabniki se strinjajo (prim. Fajfar in Žagar Karer[2], 2015), da potrebujejo zanesljivo in verodostojno terminologijo, torej terminologijo, ki ima med strokovnjaki določeno stopnjo soglasja, s katero se lahko uspešno sporazumevajo med seboj in ki je objavljena na primerno dostopnem odprtem mestu. Obstoječi spletni slovarski portali, ki delujejo v Sloveniji, imajo različne prednosti in nekatere pomanjkljivosti. Največ različnih terminoloških in splošnojezikovnih virov je dostopnih na portalu Termania (https://www.termania.net/), ki ponuja tudi preprost urejevalnik slovarjev zainteresiranim uporabnikom, vendar prav množica virov lahko oteži iskanje in prepoznavanje najustreznejšega poimenovanja. Terminološki viri so dostopni tudi na Terminologišču (https://isjfr.zrc-sazu.si/sl/terminologisce/slovarji), kjer so objavljeni znanstveni terminološki slovarji, nastali v okviru Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, poleg tega pa na Terminologišču že od 2013 deluje Terminološka svetovalnica.

Pred začetkom dela smo preverili portale s terminološkimi viri v drugih državah (podrobno v Jemec Tomazin in Romih, 2023), od katerih smo natančneje ovrednotili rešitve, ki jih prinašajo Hrvatski terminološki portal (http://nazivlje.hr/), Eurotermbank (https://www.eurotermbank.com/) in katalonski Cercaterm (https://www.termcat.cat/en/cercaterm).

V projektu RSDO smo zato želi oblikovati novo spletno mesto, ki bo namenjeno izključno terminologiji in bo omogočilo združevanje terminoloških virov z različnih spletnih mest, ponujalo terminološko svetovanje, hkrati pa omogočilo preprosto izdelavo novih terminoloških katerikoli skupini ali posamezniku, zato so bila dodana tudi navodila za samostojno izdelavo virov ali se obrnejo na urednike portala. Pri razvoju terminološkega portala je sodelovalo 33 raziskovalcev s petih raziskovalnih ustanov, in sicer Oddelka za terminologijo Inštituta za slovenski jezik ZRC SAZU, podjetja Amebis, Instituta Jožef Stefan, Filozofske fakultete UL ter Fakultete za elektrotehniko, računalništvo in informatiko UM.

2.1 Funkcionalnosti terminološkega portala

Na osnovi analize drugih spletišč s terminološkimi vsebinami (prim. Jemec Tomazin in Romih, 2023) smo pri zasnovi terminološkega portala želeli: a) združiti čim več specializiranih virov na enem mestu ter pri tem oblikovati enotno shemo terminoloških virov z mednarodno primerljivimi metapodatki; b) vključiti čim bolj standardizirane in odprtodostopne terminološke vire; c) ponuditi odprtodostopno specializirano ter hkrati preprosto orodje za izdelavo novih terminoloških virov; č) ponuditi možnost izvoza in izmenjave odprtodostopnih podatkov vsem registriranim uporabnikom terminološkega portala; d) omogočiti programsko povezljivost podatkov na novem terminološkem portalu za potrebe zunanjih spletnih storitev, npr. pomnilnike prevodov; e) uporabnikom ponuditi tudi terminološko svetovanje v primerih, ko v obstoječih virih ne bi našli odgovora na svoje terminološko vprašanje; f) z registracijo uporabnikov omogočiti preglednejše urejanje vsebin.

2.1.1 ISKALNIK

Osnovna funkcija terminološkega portala je iskanje podatkov po terminoloških virih. Pri razvoju iskalnika je bilo glavno vodilo enostavnost in hitrost iskanja ter pregleden prikaz podatkov o najdenih zadetkih. Za iskanje ni potrebna registracija ali prijava uporabnika na portalu.

Osnovno iskanje je privzeta oblika iskanja. Iskalnik ima klasično iskalno polje, v katero se vpiše niz znakov. Iskalnik išče ločeno po slovenskih terminih, terminih v drugih jezikih in preostali vsebini slovarskega sestavka. Poleg iskanja ene besede je v terminologiji pogosto iskanje večbesednih terminov, pri čemer iskalnik išče vse slovarske sestavke, v katerih se pojavljajo vse iskane besede, tudi če je med njimi še več drugih besed. Zapis večbesednih terminov med narekovaji omogoča točno določene zadetke. Omogočena je tudi uporaba nadomestnih znakov – vprašaja (?) in zvezdice (*). Znaki, kot so klicaj (!) (izločanje niza), ključnik (#) (od–do) in drugi zaradi redke uporabe na ostalih slovenskih slovarskih portalih (npr. na Franu, Termanii) niso omogočeni.

Napredno iskanje omogoča še filtriranje zadetkov po jezikih, področjih, slovarjih in tudi terminoloških virih, ki so na portalu dostopni prek povezanih portalov. Zadetki z drugih slovarskih portalov so vidni v testnem okolju, Slovenski terminološki portal pa trenutno še ne prikazuje podatkov z drugih slovarskih portalov, kar je ena od nalog za prihodnji razvoj. V času oddaje prispevka (maj 2024) je na portalu objavljenih 7 slovarjev s skupno 2839 slovarskimi sestavki, v izdelavi pa je še 36 slovarjev, ki se jih uredniki še niso odločili objaviti.

2.1.2 LUŠČILNIK

Eno od pomembnih izhodišč pri zasnovi novega terminološkega portala je bila izdelava pomoči za samostojno oblikovanje novega terminološkega vira, kamor sodi tudi enotavna izdelava manjših specializiranih korpusov in luščenje terminoloških kandidatov. Pri tem ima uporabnik možnost, da po določenih kriterijih izbere manjšo skupino besedil iz obstoječega označenega korpusa OSS, ki je nastal kot eden od rezultatov projekta RSDO.

Pred luščenjem terminoloških kandidatov lahko uporabnik pripravi svoj nabor strokovnih besedil za specializirani korpus. Podprti formati besedil so .txt, .docx in .pdf. Še pred začetkom luščenja je uporabnik opozorjen, če katerega od besedil program ne prepozna in ga zato izloči.

Druga možnost je, da uporabnik izbere predmetno področje v korpusu OSS, v katerem so združena besedila, objavljena na Nacionalnem portalu odprte znanosti do konca leta 2022, ter zažene luščenje. O koncu luščenja je uporabnik obveščen na elektronski naslov, ki ga je navedel ob registraciji.

Modul za luščenje terminoloških kandidatov je kot programska koda in samostojno orodje dostopen tudi na repozitoriju GitHub (https://github.com/clarinsi/rsdo_luscilnik). Rezultat luščenja sta seznam terminoloških kandidatov v kanonični obliki in uporabniški korpus, ki je samo uporabniku, ki je luščenje zagnal, na voljo v konkordančniku, razvitem na osnovi konkordančnika, ki ga uporablja korpus Gigafida 2.0 (več o zasnovi konkordančnika v Krek idr. 2019), v katerem uporabnik lahko preverja pojavljanje terminoloških kandidatov in primere rabe v posameznem besedilu. Vsak uporabnik ima lahko shranjenih do pet rezultatov luščenja in prav toliko uporabniških korpusov.

Seznam kanoničnih oblik terminoloških kandidatov omogoča strokovnjaku, ki ni jezikoslovno usposobljen, da prepozna termin, saj je na seznam uvrščen najpogosteje v imenovalniku ednine, ki je najpogostejša iztočnična oblika v terminološkem viru. Kanonizator omogoča dobro prepoznavo slovenske terminologije in besednih vrst, saj uporablja oblikoskladenjske oznake korpusa OSS za iskanje termina v imenovalniku, pri uporabniških korpusih, ki morajo biti pred začetkom luščenja še oblikoskladenjsko označeni, pa ustrezne zunanje jezikovne vire (npr. Sloleks oz. sezname terminov, ki so že vključeni v repozitorij Clarin.si).

S statističnimi podatki (npr. TF, TF-IDF) lahko uporabnik filtrira in ureja terminološke kandidate. Hibridni sistem luščenja je zasnovan na podlagi prepoznavanja izbranih besednovrstnih vzorcev in statističnega razvrščanja kandidatov (Hong Hanh idr. 2022). Preverjeno so bile metode, ki so upoštevale klasične statistične mere stabilnosti besednih zvez (npr. MI, PMI), mere za primerjavo specializiranega in referenčnega korpusa (npr. LUIZ-CF), statistične mere specifičnosti za izbrano strokovno področje glede na širšo zbirko besedil (npr. TF-IDF) in druge klasične mere za luščenje terminoloških kandidatov (C/NC value). Za primerjavo z referenčnim korpusom luščilnik interno uporablja korpus Gigafida 2.0, ki je označen z istimi orodji oz. s frekvenčnim seznamom, izdelanim na podlagi tega korpusa (frekvenčni seznam n-gramov lem). Za primerjavo s širšim korpusom besedil so uporabljena besedila Nacionalnega portala odprte znanosti.

Preverjene so bile tudi naprednejše metode z uporabo kontekstualnih besednih vložitev, kjer je bilo preizkušeno nenadzorovano učenje s primerjavo kontekstualnih vektorjev splošnega in domenskega korpusa ter nadzorovano učenje s prepoznavanjem terminov v učnih korpusih. Orodje za luščenje, ki deluje na portalu, združuje obe metodi (več Hong Hanh idr. 2022).

V učnem korpusu besedil RSDO5 1.1 (Jemec Tomazin idr., 2021) smo razpoznali tudi stavke, ki lahko uporabniku pomagajo poiskati čim več informacij za tvorbo definicije. Metoda temelji na razpoznavanju dobrih primerov glede na informacijo o dolžini stavka, informacijo o mestu termina v stavku in skladenjskih informacijah, značilnih za strokovna in znanstvena besedila (več Pollak 2014 in Hong Hanh idr. 2023).

Rezultate luščenja uporabnik lahko izvozi lokalno na svoj računalnik v formatu *.json*, lahko pa jih uvozi v urejevalnik in nadaljuje z urejanjem novega terminološkega vira na terminološkem portalu.

2.1.3 UREJEVALNIK

Uporabnikom je na voljo orodje, ki omogoča prilagajanje elementov slovarskega sestavka v terminološkem viru od najbolj preprostega z iztočnico v slovenskem jeziku in definicijo ali tujim terminom do kompleksnejšega z večjim številom elementov. Nujna sta najmanj dva elementa, termin v slovenskem jeziku je obvezna sestavina.

Urejevalnik omogoča uvoz podatkov iz običajnih datotečnih formatov (.xml, .csv, .tsv, .tbx), prav tako pa lahko terminološki vir v katerem koli trenutku izvozi lokalno v prej naštetih formatih. Pomembna funkcionalnost urejevalnika je večuporabniško delo pri enem terminološkem viru, pri čemer urednik vira lahko drugim registriranim uporabnikom omogoči različne pravice, npr. spreminjanje vseh podatkov, strokovni pregled. Urednik terminološkega vira mora zato poznati uporabniška imena drugih uporabnikov, v nadgradnji pa smo predvideli dodajanje spustnega seznama z obstoječimi uporabniki.

Struktura slovarskega sestavka predvideva naslednje elemente: *termin* v slovenskem jeziku, *področno oznako*, ki termin uvrsti na neko ožje področje, *pojasnilo*, ki sporoča okoliščine, *definicijo*, ki opiše glavne značilnosti pojma, morebitni *sinonim*, ki označuje isti pojem, vendar z drugim poimenovanjem, *povezani termin*, ki omogoča razvijanje ontoloških povezav v terminološkem viru, polje *drugo* je vsebinsko nedefinirano in omogoča vključitev podatkov, ki si jih uporabnik želi, npr. zgledov rabe, vire, sledi pa del, v katerem uporabnik določi, v katerem *tujem jeziku* bodo navedeni elementi tujejezičnega dela slovarskega sestavka, pri čemer je možno dodati *tuji termin*, *definicijo* in morebitni *sinonim*. Sledijo še polja za *slike*, *zvok* in *video*, vendar omogočajo vnos povezav na druga spletna mesta, saj strežniške zmogljivosti ne omogočajo hranjenja izvirnih datotek na portalu. Elemente slovarskega sestavka v terminološkem viru je treba določiti pred začetkom urejanja, je pa možno v pri urejanju lastnosti slovarja elemente slovarskega sestavka dodati ali odstraniti tudi po začetku urejanja vsebine.

Pri strukturi velja opozoriti na dve posebnosti, dodana je kategorija *povezani termin*, ki označuje sorodni, vendar ne isti pojem, saj je za sinonime predvidena samostojna kategorija. Uporabnik lahko izbere tudi vrsto povezave med termini, in sicer *sorodni*, *ožji*, *širši*, pri čemer mora izbirati med termini, ki jih je že vključil v svoj terminološki vir. Čeprav so te informacije zelo dragocene, so vendar redko sistemsko izpeljane po celotnem viru. Druga posebnost je element *Drugo*, ki nima določene vsebine, če pa je uporabnik uvozil rezultate luščenja terminoloških kandidatov, pa so v tem polju izpisani tudi primeri rabe, ki vključujejo termin in predstavljajo pomoč za oblikovanje definicije. Primeri rabe v terminoloških slovarjih niso tipična informacija, ki bi jo potrebovali strokovnjaki, hkrati pa se vsak avtor terminološkega vira na portalu lahko odloči, da jo želi vključiti.

2.1.4 SVETOVANJE

V programski kodi je razvit tudi samostojni modul za svetovanje. Ima obliko foruma, ki omogoča, da administrator portala dodeli določenemu registriranemu uporabniku pravico urednika svetovalnica, ta pa med drugimi registriranimi uporabniki izbere svetovalce ter jim dodeljuje nova terminološka vprašanja. Določeni svetovalec pripravi mnenje in ga pošlje v pregled uredniku, ki ga po potrebi dodeli v dopolnitev še drugim svetovalcem, lahko pa končno verzijo terminološkega odgovora tudi potrdi, zavrne ali pripravo odgovora dodeli drugemu svetovalcu. Modul je na voljo za morebitne samostojne namestitve terminološkega portala.

Terminološko svetovanje na Slovenskem terminološkem portalu pa ima integrirano povezavo s Terminološko svetovalnico Inštituta za slovenski jezik (prim. Fajfar, Žagar Karer 2023). Prijavljeni uporabniki portala lahko zastavijo novo vprašanje na posebnem obrazcu, ki ima že predizpolnjena polja z imenom in priimkom ter e-naslovom, uporabnik pa dopolni druge podatke, zlasti je zaželeno, da čim bolj natančno opiše svoj problem. Oddano vprašanje na portalu prispe na e-naslov Oddelka za terminologijo, priprava odgovora pa poteka na povsem enak način, kot če je oddano v Terminološki svetovalnici na Terminologišču (https://isjfr.zrc-sazu.si/sl/terminologisce/svetovanje). Svetovanje je namenjeno pomoči pri terminoloških dilemah na področjih, kjer še ni izdelanih terminoloških virov ali pa uporabnik odgovora ne more najti v obstoječih.

Analiza delovanja kaže, da poslanih vprašanj ni veliko, saj so bila doslej poslana samo testna vprašanja ob predstavitvah portala, uporabniki, ki potrebujejo pomoč, pa se obrnejo neposredno na Terminološko svetovalnico

ISJFR, kar je razvidno iz naslova prejetega sporočila (*Novo vprašanje za Slovenski terminološki portal*).

2.2 Upravljanje portala

Terminološki portal je administriran. Administrator portala ima pravice, da dodeljuje pooblastila izbranim uporabnikom, ki določi za skrbnike slovarjev ter skrbnike svetovalnice. Edino administrator ima možnost, da portal poveže z drugim portalom in tako se med zadetki iskanja lahko prikazujejo tudi tisti, ki so objavljeni drugje. Prav tako lahko ureja slovarje, dopolnjuje njihove lastnosti, npr. kolikšno je minimalno število slovarskih sestavkov, da je vir lahko objavljen (prednastavljene vrednosti so 1, 20, 50, 100), ima možnost vklopa ali izklopa, da vir, ki ga uredniki želijo objaviti, potrebuje potrditev skrbnika slovarjev in dovoljuje objavo na zunanjem delu portala. Te nastavitve veljajo za vse terminološke vire na portalu.

Uporabnik z nekoliko manj pravicami je skrbnik slovarjev, ki lahko dovoljuje objavo podatkov na zunanjem delu portala ter dovoljuje uporabnikom urejanje enega od začetih terminoloških virov, kar je sicer pravica glavnega urednika posameznega terminološkega vira. Z uredniki vira lahko komunicira preko t. i. notranjih komentarjev ob posameznem viru. Ker je portal zasnovan tako, da bo omogočal preglednost tudi pri večjem številu virov, imata skrbnik slovarjev in administrator portala možnost potrjevati nova podpodročja, ki so jih predlagali uporabniki.

Po registraciji lahko uporabniki začnejo uporabljati vse funkcionalnosti Slovenskega terminološkega portala. Stik z uredniki potrebujejo samo v primeru, da želijo svoj vir objaviti na zunanjem, javnem delu portala.

3 SKLEP

Slovenski terminološki portal ponuja možnost za samostojno urejanje terminoloških virov. Ponuja možnost takojšnje objave tudi krajših slovarjev in v juniju 2024 je bil objavljen Terminološki slovar odprte znanosti. Prednost portala je tudi lažje sodelovanje skupin strokovnjakov, ki ne delujejo v isti ustanovi.

Slovarji, ki še niso objavljeni, kažejo na razmeroma veliko potrebo po orodju za urejanje terminologije, bi pa morali več truda vložiti v diseminacijo.

Vprašanje, ki smo si ga zastavili snovalci, je tudi, ali dovoliti popravljanje posameznih slovarskih sestavkov vsem registriranim uporabnikom, ali to pravico pustimo samo urednikom terminološkega vira. Odločili smo se, da bo omogočeno komentiranje vsakega slovarskega sestavka, pri čemer imajo uredniki posameznega vira možnost, da komentarje tudi izklopijo, nismo pa dovolili popravljanja besedila vsem uporabnikom.

S komentarji imajo uredniki možnost razmisleka in tudi odgovora na komentar, hkrati pa lahko presodijo, ali bodo popravek upoštevali ali ne. V urejevalniku imajo uredniki možnost, da vidijo zadnjih 10 shranjenih sprememb slovarskega sestavka, drugim uporabnikom pa se v iskalniku prikazuje zadnja objavljena različica. Komentarji pa po drugi strani omogočajo neposredno komunikacijo med uredniki terminološkega vira in (prijavljenim) uporabnikom, kar doslej na tak način v slovenski terminologiji še ni bilo mogoče.

ZAHVALA

Prispevek je nastal v okviru projekta Razvoj slovenščine v digitalnem okolju (C3340-20-278001), ki ga je financiralo Ministrstvo za kulturo RS in v okviru raziskovalnega programa Slovenski jezik v sinhronem in diahronem razvoju (P6-0038), ki ga financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS). Zahvaljujem se vsem sodelavcem v projektni skupini, zlasti Miru Romihu, dr. Senji Pollak in dr. Milanu Ojsteršku za sodelovanje pri projektu.

LITERATURA

Arhar Holdt, Š. in Krek, S. (Ur.). (2023). *Razvoj slovenščine v digitalnem okolju.* Založba Univerze v Ljubljani. https://doi.org/10.4312/9789612972561.

Fajfar, T. in Žagar Karer, M. (2015). Strokovnjaki in prepoznavanje terminov v strokovnih besedilih. *Jezikoslovni zapiski*, 21(1), 7–21. https://doi.org/10.3986/JZ.21.1.6843.

Fajfar, T. in Žagar Karer, M. (2015). Strokovnjaki in drugi uporabniki terminologije kot naslovniki terminoloških slovarjev. *Jezikoslovni zapiski*, 21(1), 23–35. https://doi.org/10.3986/JZ.21.1.6861.

Fajfar, T. in Žagar Karer, M. (2023). Ad hoc terminologija: analiza uporabe terminoloških načel v Terminološki svetovalnici na Terminologišču. *Slavistična* *revija*, 71(4), 369–384. https://doi.org/10.57589/srl.v71i4.4146.

Hong Hanh, T., Martinc, M., Repar, A., Doucet, A., Pollak, S. (2022). A Transformer-based sequence-labeling approach to the Slovenian cross-domain automatic term extraction. V D. Fišer, T. Erjavec (Ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (196–204). https://nl.ijs.si/jtdh22/pdf/JTDH2022_Tran-et-al_A-Transformer-based-Sequence-labeling-Approach-to-the-Slovenian-Cross-domain-Automatic-Term-Extraction.pdf. Pridobljeno 31. maja 2024.

Hong Hanh, T., Podpečan, v., Jemec Tomazin, M., Pollak, S. (2023). Definition Extraction for Slovene: Patterns, Transformer Classifiers and ChatGPT. V Medveď, M., Měchura, M., Tiberius, C., Kosem, I., Kallas, J., Jakubíček, M., Krek, S. (Ur.) (2023), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. Brno, 27–29 June 2023. Brno: Lexical Computing CZ s.r.o. https://elex.link/elex2023/wp-content/uploads/19.pdf. Pridobljeno 31. maja 2024.

Jemec Tomazin, M. in Romih, M. (2023). Slovenski terminološki portal – nova priložnost za urejanje slovenske terminologije. V Š. Arhar Holdt, S. Krek (Ur.), *Razvoj slovenščine v digitalnem okolju* (str. 211–247). Založba Univerze v Ljubljani. https://doi.org/10.4312/9789612972561.

Jemec Tomazin, M., Trojar, M., Žagar Karer, M., Atelšek S., Fajfar, T. in Erjavec, T. (2021). *Corpus of term-annotated texts RSDO5 1.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1470.

Krek, Simon, Arhar Holdt, Špela, Erjavec, Tomaž, Čibej, Jaka, Repar, Andraž, Gantar, Polona, Ljubešić, Nikola, Kosem, Iztok in Dobrovoljc, Kaja (2019). *Specifikacije izdelave korpusa Gigafida 2.0, v1*. Ljubljana: Center za jezikovne vire in tehnologije, Univerza v Ljubljani. https://www.cjvt.si/gigafida/wp-content/uploads/sites/10/2019/06/Gigafida2.0_specifikacije.pdf. Pridobljeno 20. avgusta 2024.

Pollak, S. (2014). Luščenje definicijskih kandidatov iz specializiranih korpusov. Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave, 2(1), 1–40. https://doi.org/10.4312/slo2.0.2014.1.1-40.

Žagar Karer, M. in Fajfar, T. (2015). Strokovnjaki in prepoznavanje terminov v strokovnih besedilih. *Jezikoslovni zapiski*, 21(1), 7–21. https://ojs.zrc-sazu.si/jz/article/view/6843/6433.

Slovenski terminološki portal. Pridobljeno 31. maja 2024, https://terminoloski.slovenscina.eu/

Žagar, K., Ferme, M., Ojsteršek, M., Jemec Tomazin, M., Erjavec, T. (2023). *Corpus of scientific texts from the Open Science Slovenia portal OSS 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1774.

UTILIZING FORCED ALIGNMENT FOR PHONETIC ANALYSIS OF SLOVENE SPEECH

Janez KRIŽAJ,[1] Jerneja ŽGANEC GROS,[2] Simon DOBRIŠEK[1]
[1] Laboratory for machine intelligence, Faculty of electrical engineering, University of Ljubljana, Tržaška cesta 25, 1000 Ljubljana, Slovenia
[2] Alpineon R&D, d. o. o., Ulica Iga Grudna 15, 1000 Ljubljana

1 INTRODUCTION

Forced alignment is a process in speech technology where given transcriptions are automatically aligned with the corresponding audio. This technique finds extensive application in various speech processing tasks, including automatic speech recognition (ASR), speech synthesis, subtitle generation, audio anonymization and phonetic research where it is used to study phenomena like speech reduction (Adda-Decker & Lamel, 2018), historical sound changes in languages (Labov et al., 2013), diagnosing speech sound disorders (Y. Li et al., 2023), and speech disfluencies (Kouzelis et al., 2023).

Forced alignment techniques play a pivotal role in the field of phonetic analysis by providing precise and efficient tools for aligning phonetic transcriptions with audio recordings. This capability is essential for researchers who seek to analyze and understand the intricate details of speech production and variation across different languages and dialects. The Montreal Forced Aligner (MFA) is a widely used tool in this domain, known for its robustness and accuracy in aligning phonetic segments with corresponding audio (Wu et al., 2023), (Chodroff et al., 2024).

We have chosen the MFA as the primary method for forced alignment due to its comprehensive feature set and proven track record in various linguistic studies. Our contributions include the forced alignment of the Slovene speech corpus GOS2.1 (Verdonik, Zwitter Vitez, et al., 2023) at the word, syllable, and phone levels, and its application to basic acoustic measurements. We also compare its performance with the more recent NeMo Forced Aligner (NFA) (Rastorgueva et al., 2023) to evaluate its comparability with state-of-the-art methods. The source code containing alignment procedures and acoustic measurements is freely available at https://github.com/jan3zk/forced_alignment.

The paper is structured as follows: we begin with a review of related work, then describe the forced alignment methodology and our experiments, and conclude with potential avenues for future research.

2 RELATED WORK

The application of forced alignment to phonetic research has been increasingly popular, as demonstrated by studies like (Yuan et al., 2023), which used forced alignment for phonetic segmentation and investigating speech variation. Contribution by (Young & McGarrah, 2023), focuses on the study of forced alignment on a lesser-known language with a limited amount of training data. The work of (Adda-Decker & Lamel, 2018) explores discovering speech reductions across different speaking styles and languages, underlining the versatility of forced alignment in diverse linguistic contexts. Moreover, (Kouzelis et al., 2023) emphasized the alignment of disfluent speech, eliminating the need for verbatim transcription. Additionally, (Huang et al., 2024), (J. Li et al., 2022) and (Sun, 2023) present examples of recent deep learning-based approaches, furthering the innovation in the field of forced alignment.

On the software front, several forced alignment solutions exist. PraatAlign (Lubbers & Torreira, 2013-2018), is effective but not scalable for larger datasets. Deep learning approaches such as NVIDIA NeMo (Rastorgueva et al., 2023) and WhisperX (Bain et al., 2023) offer accurate and efficient word-level alignment but lack alignments at the phone-level. For our research, we utilized the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) for its proficiency in providing phone-level alignments and its robust performance in our experimental setup.

3 FORCED ALIGNMENT FOR PHONETIC RESEARCH

3.1 Montreal Forced Aligner

In the MFA alignment process, the pronunciation dictionary is first used to convert the transcription into phones. Next, the acoustic model determines which parts of the audio recording correspond to the specific phones from the dictionary. The pronunciation dictionary provides the phones to search for, while the acoustic model assesses how well the audio patterns match these phones.

The pronunciation dictionary contains a list of words along with their phonetic transcriptions, indicating which phones compose each word. This transcription is crucial, as it informs the acoustic model of the sounds to search for in the audio signal. The acoustic model is trained to recognize specific phones from audio features such as Mel-Frequency Cepstral Coefficients (MFCCs). It models how individual phones sound in different contexts, such as monophones and triphones, as well as with varying speech styles or accents. The acoustic model is based on Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs), which determine how likely a given set of audio features corresponds to a specific phone or sequence of phones.

In finding the most likely path through these models, which allows aligning the audio signal with the text transcription, the MFA employs the Viterbi algorithm, optimized for efficient search of the most probable sequences of phones that align with the observed audio features.

The MFA process outputs a detailed information about the timing of each phone and word in the input speech audio, as depicted in the flowchart (Figure 1).

3.2 Additional Tiers for Phonetic Research

The output of MFA initially provides only two tiers in the output files, containing alignments at the word and phone levels. To enhance the utility of these files for linguistic and phonetic research, we have introduced additional tiers into the annotation files, as shown in Figure 2. These tiers are designed to facilitate more detailed analyses of Slovene speech allowing researchers to delve into various aspects of speech, including prosody, discourse, and conversational dynamics, which are pivotal for comprehensive linguistic research.

Several tiers are derived from data in the transcript files, which are provided as corpus metadata, and combined with time intervals obtained through forced alignment. These tiers offer information regarding speaker IDs, word IDs, speaker changes, and conversational text transcriptions.

Other tiers result from further processing of the corresponding audio file and/or data within the existing tiers. These tiers include information on discourse

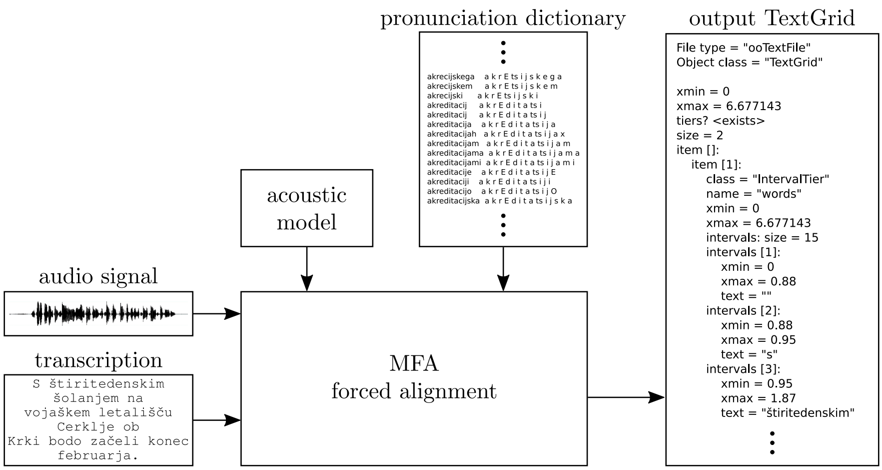


Figure 1: Schematic representation of the forced alignment process. On the left are the input files: a speech audio and its transcription. Given the acoustic model and the pronunciation dictionary, the MFA produces an output TextGrid file, which contains time intervals at both the word and phone levels.

markers, syllable intervals, pitch resets, intensity resets, speech rate reductions, and pauses, enriching the scope of phonetic analysis.

3.3 Acoustic Measurements

Utilizing the synchronized audio and transcript files, our study involves the computation of a variety of acoustic metrics essential for subsequent speech analysis. For each phone, we calculate various acoustic features including the duration of each phone, average pitch values, pitch trend, formant frequencies, intensity levels, voice onset time, and the center of gravity.

4 EXPERIMENTS

4.1 Databases

The Gos 2.1 corpus (Verdonik, Zwitter Vitez, et al., 2023) is a comprehensive reference speech corpus for the Slovenian language, comprising about 300 hours of speech (2.4 million words and 127 thousand utterances from 1,500 texts), enhanced with word-level temporal information where possible. It amalgamates

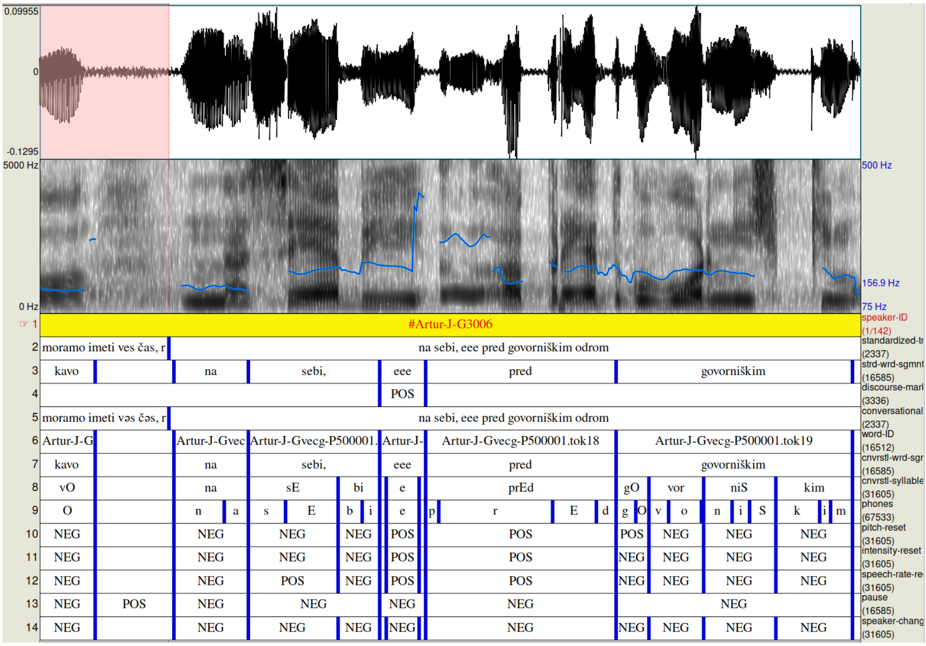


Figure 2: Praat software displaying additional TextGrid tiers for an in-depth analysis of Slovene speech.

data from three sources: the Gos 1.1 corpus (112 hours, 1 million words), the Gos VideoLectures 4.2 (22 hours, 179,000 words), and selections from the ARTUR 1.0 ASR database (185 hours, 1.2 million words), including varied content like media recordings, dialogues, and transcriptions from the Slovene National Assembly. Transcriptions are provided in two forms: pronunciation-based and standardised orthographic transcriptions. This edition features improvements such as unified casing and normalization, re-introduction of punctuation, additional temporal data, and uniform encoding across subcorpora. Corpus transcriptions are presented in TEI (XML) format and include part-of-speech tagging and lemmatisation, automated by CLASSLA (Ljubešić & Dobrovoljc, 2019).

4.2 Experimental Setup

We adopted the training approach outlined by (McAuliffe et al., 2017) for MFA alignment, training the acoustic model on the Artur-B subset (Verdonik, Bizjak, et al., 2023). This subset includes 485 hours of read speech and is distinct from the GOS2.1 dataset, which is used for subsequent alignment evaluations.

4.3 Alignment Accuracy Evaluation

In the absence of ground truth time intervals, we conducted a comparative analysis of the MFA with two other forced alignment methods: Nemo Forced Aligner (NFA) (Rastorgueva et al., 2023) and temporal intervals obtained by Kaldi recipe which are part of GOS2.1 database metadata. The NFA uses a pretrained Conformer Connectionist Temporal Classification (CTC) ASR model, trained on the Artur corpus as described by (Lebar Bajec et al., 2022), while the Kaldi approach utilizes the Artur-P subset for training (Verdonik et al., 2024).

For our evaluation, each method was alternately used as the reference (pseudo-ground truth) to assess the relative performance of the others. The models were assessed by calculating the mean and median absolute errors between the predicted and reference temporal boundaries at the word level. Additionally, we computed the proportion of boundaries falling within specific tolerance thresholds (10 ms, 50 ms, 100 ms). These metrics are restricted to the word level due to the lack of phone-level alignments in the Kaldi and NFA methods.

Each of the assessed methods served as the reference (pseudo-ground truth) in turn allowing us to evaluate the relative performance of the other methods. The evaluation of the models is conducted by assessing the mean and median of absolute errors between the predicted temporal boundaries and reference values at the word-level temporal data as well as share of boundaries that fall within specific tolerance thresholds (10 ms, 50 ms, 100 ms). These evaluation metrics are calculated at the word level due to the absence of phone level annotations in case of Kaldi and NFA methods.

The results, summarized in Table 1, reveal significant differences in alignment accuracy across different subcorpora. Not surprisingly, we observed higher error rates in the Artur-N and GosVL corpora, which feature more complex audio elements like spontaneous conversations, overlapping speech, and speakers wearing masks. When compared against the Kaldi reference, the MFA predicts over 90% of time boundaries within a 100 ms margin across all examined subcorpora. While the MFA method generally demonstrated higher accuracy compared

Table 1: Performance assessment of the evaluated methods at the word level across different subsets of the GOS2.1 database.

Corpus	Method	Mean abs. err. [ms]	Median abs. err. [ms]	Share within 10 ms [%]	Share within 50 ms [%]	Share within 100 ms [%]
Artur-P	MFA vs Kaldi	28	11	47	90	96
	NFA vs Kaldi	71	49	11	51	86
	MFA vs NFA	77	50	10	50	84
Artur-J	MFA vs Kaldi	48	14	40	84	92
	NFA vs Kaldi	86	50	11	50	83
	MFA vs NFA	86	52	10	48	52
Artur-N	MFA vs Kaldi	208	14	41	83	90
	NFA vs Kaldi	450	50	11	50	80
	MFA vs NFA	457	52	10	48	79
GosVL	MFA vs Kaldi	212	12	44	84	90
	NFA vs Kaldi	184	91	3	22	56
	MFA vs NFA	190	91	4	22	57

to the NFA when evaluated against the Kaldi references, this may be attributed to the close methodological similarities between the MFA and Kaldi approaches. Upon manual inspection, we observed that NFA has favourable performance in several challenging scenarios, particularly in cases of overlapping indistinct speech and poor transcriptions.

Considering the effective performance of the Montreal Forced Aligner (MFA) and the requirement for significant adaptations in the NeMo Forced Aligner (NFA) tokens to align with phones from our pronunciation dictionary for phone and syllable level alignment, we opted to employ MFA for our subsequent phonetic studies.

5 CONCLUSION

This study demonstrates the effective application of the Montreal Forced Aligner for phonetic segmentation and acoustic analysis of Slovene speech. Our alignment of the GOS2.1 corpus provides phone level alignments, and together with additional tiers and acoustic measurements contributes to Slovene phonetics by providing a faster alternative to manual alignments.

Future research will focus on automatic detection of speech disfluencies and examining phonetic variations in relation to corpus metadata, different speech styles, and different dialects. This effort aims to deepen our understanding of speech dynamics and phonetic characteristics in Slovene language.

ACKNOWLEDGMENTS

This research was made possible through the support of the MEZZANINE project (teMeljnE raZiskave Za rAzvoj govorNih vIrov in tehNologij za slovEnščino – Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language), project ID: J7-4642, financed by the Slovenian Research Agency.

REFERENCES

Adda-Decker, M., & Lamel, L. (2018). Discovering speech reductions across speaking styles and languages. In F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, & M. Zellers (Eds.), *Rethinking reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation* (pp. 101–128). Berlin, Boston: De Gruyter Mouton. doi: doi:10.1515/9783110524178-004

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Chodroff, E., Ahn, E., & Dolatian, H. (2024). Comparing language-specific and cross-language acoustic models for low-resource phonetic forced alignment. *Language Documentation & Conservation*.

Huang, R., Zhang, X., Ni, Z., Sun, L., Hira, M., Hwang, J., ... Khudanpur, S. (2024). Less peaky and more accurate ctc forced alignment by label priors. In *Icassp 2024 - 2024 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 11831-11835). doi: 10.1109/ICASSP48485.2024.10446111

Kouzelis, T., Paraskevopoulos, G., Katsamanis, A., & Katsouros, V. (2023, 08). Weakly-supervised forced alignment of disfluent speech using phoneme-level modeling. In (p. 1563-1567). doi: 10.21437/Interspeech.2023-1887

Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, *89*(1), 30–65. Retrieved 2024-04-08, from http://www.jstor.org/stable/23357721

Lebar Bajec, I., Bajec, M., Bajec, Ž., & Rizvič, M. (2022). *Slovene conformer CTC BPE E2E automated speech recognition model RSDO-DS2-ASR-E2E 2.0.* http://hdl.handle

.net/11356/1737 (Slovenian language resource repository CLARIN.SI)

Li, J., Meng, Y., Wu, Z., Meng, H., Tian, Q., Wang, Y., & Wang, Y. (2022). Neufa: Neural network based end-to-end forced alignment with bidirectional attention mechanism. In *Icassp 2022 - 2022 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 8007-8011). doi: 10.1109/ICASSP43922.2022.9747085

Li, Y., Wohlan, B. J., Pham, D.-S., Chan, K. Y., Ward, R., Hennessey, N., & Tan, T. (2023). Improving text-independent forced alignment to support speech-language pathologists with phonetic transcription. *Sensors*, *23*(24). https://www.mdpi.com/1424-8220/23/24/9650 doi: 10.3390/s23249650

Ljubešić, N., & Dobrovoljc, K. (2019, August). What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th workshop on balto-slavic natural language processing* (pp. 29–34). Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-3704 doi: 10.18653/v1/W19-3704

Lubbers, M., & Torreira, F. (2013-2018). *Praatalign: an interactive praat plug-in for performing phonetic forced alignment.* https://github.com/dopefishh/praatalign. (Version 2.0)

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. interspeech 2017* (pp. 498–502). doi: 10.21437/Interspeech.2017-1386

Rastorgueva, E., Lavrukhin, V., & Ginsburg, B. (2023). NeMo Forced Aligner and its application to word alignment for subtitle generation. In *Proc. interspeech 2023* (pp. 5257–5258).

Sun, L. (2023). Unsupervised forced alignment on syllable and phoneme with universal phonetics transcriptions. In *2023 ieee international conference on signal processing, communications and computing (icspcc)* (p. 1-5). doi: 10.1109/ICSPCC59353.2023.10400287

Verdonik, D., Bizjak, A., Žgank, A., Bernjak, M., Antloga, Š., Majhenič, S., ... Bordon, D. (2023). *ASR database ARTUR 1.0 (audio).* http://hdl.handle.net/11356/1776 (Slovenian language resource repository CLARIN.SI)

Verdonik, D., Dobrovoljc, K., Erjavec, T., & Ljubesic, N. (2024). Gos 2: A new reference corpus of spoken slovenian. In *International conference on language resources and evaluation.* https://api.semanticscholar.org/CorpusID:269804654

Verdonik, D., Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., ... Rupnik, P. (2023). *Spoken corpus gos 2.1 (transcriptions).* http://hdl.handle.net/11356/1863 (Slovenian language resource repository CLARIN.SI)

Wu, H., Yun, J., Li, X., Huang, H., & Liu, C. (2023, 07). Using a forced aligner for prosody research. *Humanities and Social Sciences Communications*, *10*. doi: 10.1057/s41599-023-01931-4

Young, N. J., & McGarrah, M. (2023). Forced alignment for nordic languages: Rapidly constructing a high-quality prototype. *Nordic Journal of Linguistics*, *46*(1), 105–131. doi: 10.1017/S033258652100024X

Yuan, J., Lai, W., Cieri, C., & Liberman, M. (2023). Using forced alignment for phonetics research. In C.-R. Huang, S.-K. Hsieh, & P. Jin (Eds.), *Chinese language resources: Data collection, linguistic analysis, annotation and language processing* (pp. 289–301). Cham: Springer International Publishing. doi: 10.1007/978-3-031-38913-9_17

DEGREES OF BELONGING TO EUROPE IN PARLIAMENTARY DISCOURSE: A COMPARATIVE CORPUS-ASSISTED STUDY

Anna KRYVENKO
Institute of Contemporary History, Slovenia
National Institute for Strategic Studies, Ukraine

1 INTRODUCTION

The question of what it means to belong to Europe reemerges in public discourse in times of crises. Answers vary significantly depending on the following key factors: what is implied by Europe, how belonging is understood, what the speakers' vantage point (national or international, political, ideological, etc.) and communicative aims are. The word *Europe* is highly polysemous and the complexity of the concept of EUROPE has been evolving over time (Heinemann et al., 2022, pp. 7-11). In contemporary public discourse produced by EU political actors, the ambiguity of *Europe* has been aggravated by: 1) a deliberate convergence between Europe as a broadly understood geographical, historical or cultural space and the EU as a political organization with a set of institutions (Krzyzanowski, 2010, pp. 91-94); 2) conceptualizations of EUROPE as either a single imaginary space with rather fuzzy boundaries or as a sum of individual countries located on the continent irrespective of their membership in the EU; 3) multiple regional divisions habitually loaded with complex connotations, which may include the east-west divide, with eastern Europe marked more commonly than western Europe, central Europe, or continental Europe (e.g., see Williams et al., 2012, pp. 68-80).

The discursive construction of belonging to any of these 'Europes' inevitably involves considerations of inclusion and exclusion, which are traditionally expressed as the binary distinction between Us and Them and employed into identity building (Wodak et al., 2009, p. 35). However, recently it has been pointed out that this well-established dichotomy stemming from social identity theory fails to capture fuzzy areas of belonging to Europe located between Us and Them (Le, 2021, p. 206), which embrace such countries as the UK and Ukraine. Importantly, considerations of belonging to Europe and the extent of this belonging are dynamic and subject to change. They are influenced by the attitudes of individual or collective actors, which may evolve over time. These attitudes are affected by shifts in the geopolitical landscape and can be challenged by various crises.

National parliamentary settings are particularly well suited for exploring the recurrent discursive construction of belonging to Europe over time due to the representative nature of parliaments, their ideological diversity, rootedness in national contexts and openness to the public. Political discourse scholars have been increasingly interested in corpora of parliamentary proceedings as a data source (see Fišer and Lenardič, 2018, for an overview). However, until now there has been a dearth of cross-linguistic and cross-cultural research drawing on parliamentary corpora (Truan & Romary, 2021). Moreover, the Ukrainian parliamentary data have been put in the spotlight of the international research community only recently due to the release of the first full-text corpus of parliamentary proceedings from the Verkhovna Rada of Ukraine (Kryvenko & Kopp, 2023), along with the comparable corpus of proceedings in the UK Parliament as well as multiple other parliaments under the ParlaMint project (Erjavec et al., 2024). The availability of comparable and interoperable linguistic corpora is a prerequisite for addressing the problem of elaborating qualitative and “quantitative factors that can be used to measure and demonstrate different levels of belonging” (Williams et al., 2012, p. 56) to EUROPE from cross-linguistic and cross-cultural perspectives.

The major research questions are:

1) Have references to Europe changed over time in Ukrainian and British parliamentary discourse amidst recent crises and geopolitical shifts (Russia’s war against Ukraine, Covid-19 pandemic, Brexit, and the granting of candidate status for EU membership to Ukraine)?

2) How do different political actors in Ukrainian and British parliamentary settings discursively construct their countries’ relatedness to Europe?

2 DATA AND METHODS

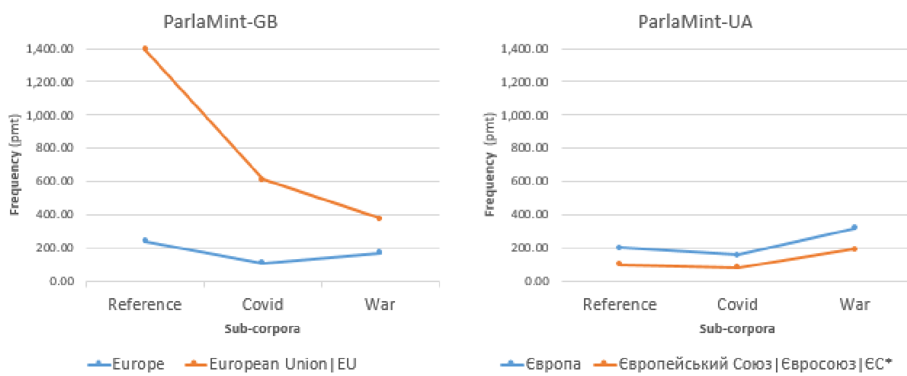
Methodologically, the research reported here combines corpus linguistics, (critical) discourse studies within the framework of MD CADS (Partington et al., 2013; Marchi, 2018; Räikkönen, 2023), and insights from cognitive linguistics including spatial semantics in terms of image schemas (Croft and Cruse, 2004) as well as conceptual metaphor theory in its application to an analysis of the political debate on Europe and European identity (Schäffner, 1996; Musolff, 2001; Zhabotynska, 2018). This study utilizes distant and close readings of concordance lines or larger fragments of the transcripts, when necessary. The NoSketch Engine concordancer was used to interrogate ParlaMint-GB and ParlaMint-UA v. 4.0.

To address the research questions posed above the following steps were taken. First, the relative frequencies of the proper names *Europe* and *Європа* ‘Europe’ in English and Ukrainian were compared to the relative frequencies of the proper names *the European Union, the EU, Європейський Союз* ‘European Union’, *ЄС* ‘EU’, and the clipped compound *Євросоюз* ‘(lit.) *Eurounion’ in the built-in sub-corpora splitting the data into the reference period (until 30 January 2020), the period marking the formal declaration of the public health emergency for Covid-19 as well as the withdrawal of the UK from the EU (from 31 January 2020), and the period after the beginning of Russia’s full-scale war of aggression against Ukraine (from 24 February 2022 onwards) in ParlaMint-GB and ParlaMint-UA. Second, to retrieve contexts potentially relevant for further analysis, the corpora were queried for concordance lines containing the co-occurrence of the first-person plural personal and possessive pronouns (the lemmas *we* and *our* in English and the lemmas *ми* ‘we’ and *наш* ‘our’ in Ukrainian) or the country names *(Great) Britain, the UK, the United Kingdom* and *Україна* 'Ukraine' with the proper names *Europe* and *Європа* ‘Europe’ in a span of nine words to the left and right. The retrieved results were filtered to remove named entities containing the word *Europe* or *Європа* ‘Europe’ respectively (e.g. *Council of Europe, Horizon Europe*) and saved as separate sub-corpora preserving the temporal distinctions described above. Third, the metadata including the speakers’ gender and party affiliation were compared across the sub-corpora and the corpora. Fourth, the filtered results were manually analyzed with respect to the discursively constructed spatial relation of the nations under study to Europe according to the following model: position in Europe, destination to Europe, destination (away) from Europe, and position out of Europe. This working model loosely stems from (Quirk et al. 1985, p. 674).

3 RESULTS AND BRIEF DISCUSSION

Figure 1 illustrates three notable findings: a sharp decline in the aggregated references to the European Union in ParlaMint-GB between 2015 and 2023, a consistent preference for *Європа* ‘Europe’ over the aggregated proper names of the European Union in ParlaMint-UA throughout the observed period, and an increase in references to *Europe* during Russia’s large-scale aggression compared to the Covid-19 pandemic in both corpora.

Figure 1: Comparison between relative frequencies (pmt) of the proper names for Europe and the European Union in English and Ukrainian across the sub-corpora of ParlaMint-GB and ParlaMint-UA.[1]



As seen in Figure 2, the proportion between relative frequencies of the country names *(Great) Britain | the UK | the United Kingdom* on the one hand and *we | our* on the other hand stayed relatively unchanged across the sub-corpora of ParlaMint-GB, while the proportion between relative frequencies of *Україна* 'Ukraine' and *ми* ‘we’ | *наш* ‘our’ slightly shifted in the Covid sub-corpus of

[1] Instances of the homonymous abbreviation “ЄС”, which stands for “Європейська солідарність” ‘European Solidarity’ – the name of a parliamentary party and its corresponding faction during the 9th term of the Rada – were excluded from the data.

ParlaMint-UA. Also, although personal and possessive pronouns co-occur with *Europe* and *Європа* 'Europe' much more readily in both corpora, the country names are less common in this context in ParlaMint-GB.

Figure 2: Comparison between the proportions of relative frequencies (pmt) in the sub-corpora of personal and possessive pronouns in aggregate vs. the proper names for the UK and Ukraine.

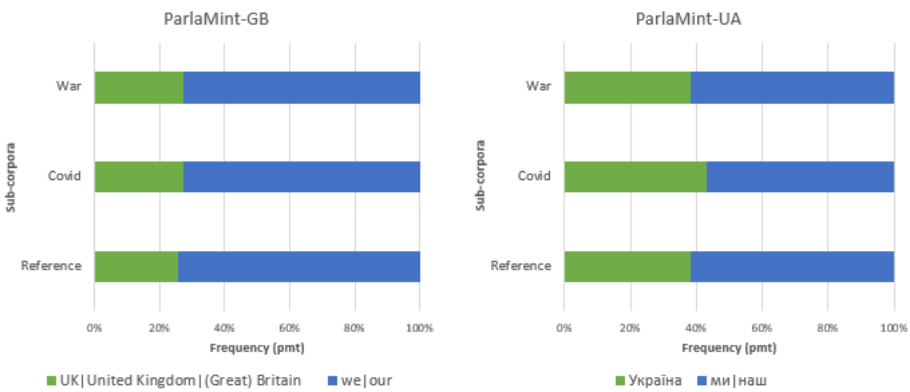
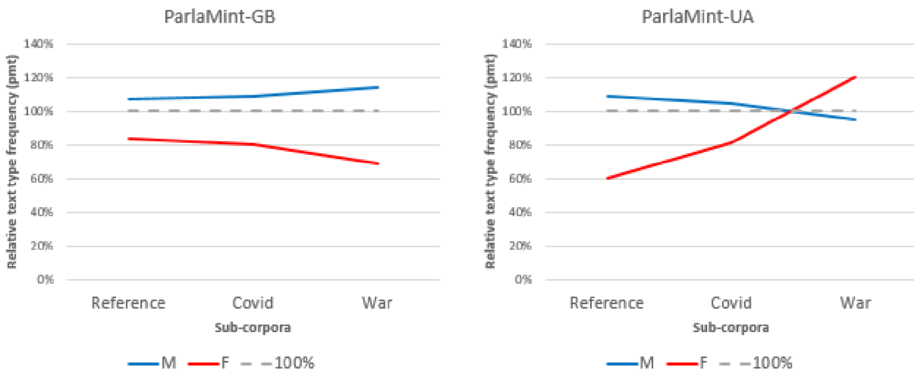


Figure 3 compares the relative text type frequency of the pronoun / country name co-occurrences with *Europe* and *Європа* 'Europe' in the sub-corpora by gender. 100% means that the result in the text type is as frequent as in the whole corpus and deviations from this value point to a higher or lower frequency in the text type compared to the entire corpus. A higher frequency of females talking about Europe in the analyzed contexts in the Covid sub-corpus of ParlaMint-UA can be explained by an increase in the number of female MPs in the Rada from 9.6% during the 7th term to 20.9% during the 9th term, which started in August 2019. However, neither Ukraine nor the UK held general elections after the beginning of Russia's large-scale invasion, so the gender balance among the MPs has not changed since then. Further analysis revealed that about one-third of the female speakers who mentioned Ukraine in relation to Europe in the war sub-corpus were foreign guest politicians who spoke in the Rada, *inter alia*, about Ukraine's place in Europe and steps towards its further European integration. The availability of these data provide an opportunity to take a closer look at parliamentary practices of positioning Self and positioning the Other, known as reflexive and interactive positioning (Ilie, 2010). However, this research avenue exceeds the scope of this contribution.

Figure 3: Relative text type frequency of the pronoun / country name co-occurrences with *Europe* and *Європа* 'Europe' by gender.



The application of the working model based on the semantics of spatial prepositions, which are frequently used with the word *Europe* in various languages (e.g., see Williams et al., 2012, pp.61-62 on 'prepositional' Europe) and motivated primarily by the perceptual schemas of containment (CONTAINER) and scale (PATH) (cf. Langacker's (2008, p. 33) conceptual archetypes), enabled the identification of all four envisioned types in both corpora. Commonly for ParlaMint-UA, Europe was conceived as an enclosed space with Ukraine / us in it or on our way to it (cf. goal-oriented collocates of the NP European integration in the Ukrainian parliamentary discourse in Kryvenko, 2018). In ParlaMint-GB, the UK / we were localized in Europe more frequently than out of it, although Europe as a negative or positive destination was used relatively rarely after Brexit. Also, the distinction between the image schema of containment (CONTAINER) and the image schema of unity/multiplicity (PART-WHOLE) was made, where applicable, to differentiate between Europe conceived as a single enclosed space and Europe conceived as a sum of countries or regions, as in (1) and (2).

(1) <u>We</u> work among the longest hours **in Europe**, and <u>we</u> very often retire later than people **in other European countries**.

(2) Мир завжди пануватиме **в** <u>нашому</u> **регіоні**, **в** <u>нашій</u> **Європі** ...

‘Peace will always prevail **in** our **region**, **in** our **Europe** …’

Further close reading of the concordance lines suggested that the collective actors in the British parliament commonly conceive of the UK as belonging or not belonging to Europe, while the collective actors in the Ukrainian parliament commonly construct Ukraine as already belonging to Europe or becoming Europe. The positioning of Ukraine as not belonging to Europe is marginal in the data; however, not being European enough due to particular policies or practices is an established pattern of criticism among the political opponents. In the British parliament, attitudes to Europe are not clearly divided along political lines between the two major parties, the Conservatives and Labour (Räikkönen, 2023, p. 153); however, the Europhile sentiment of the Scottish National Party is evident. The recurrently constructed UK identities in relation to Europe include a leader, an outlier, a partner, a neighbour, a friend, or a competitor (cf. the UK’s roles in the EU suggested by Riihimäki, 2019, p. 418). Expressions of Ukraine’s identities related to Europe are often motivated by metaphor: a shield, a breadbasket, a traveller, a student, or a family member (cf. Yavorska & Bohomolov, 2010, pp. 51-80).

ACKNOWLEDGMENTS

The study reported in this talk is funded by the Slovenian Research and Innovation Agency project N6-0288: The Changing Discursive Semantics of EU Representations: Identity, Populism, Propaganda.

SELECTED REFERENCES

Croft, W., & Cruse, D. A. (2004). *Cognitive linguistics*. Cambridge University Press.

Erjavec, T. et al. (2024). ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*, (preprint version). https://doi.org/10.21203/rs.3.rs-4176128/v1.

Erjavec, T. et al. (2023). Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0. Slovenian language resource repository CLARIN.SI. ISSN 2820-4042. http://hdl.handle.net/11356/1860.

Fišer, D., & Lenardič, J. (2018). CLARIN Corpora for Parliamentary Discourse Research. In D. Fišer, M. Eskevich, & F. de Jong (Eds), *Proceedings of the LREC2018*

Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora (pp. 2–7). European Language Resources Association. http://lrecconf.org/workshops/lrec2018/W2/summaries/14_W2.html.

Heinemann, S., Helfrich, U., & Visser, J. (2022). Europe under Construction: Introduction and Overview. In S. Heinemann, U. Helfrich, & J. Visser (Eds), *On the discursive deconstruction and reconstruction of Europe* (pp. 7–26). J.B. Metzler. https://doi.org/10.1007/978-3-662-64893-3_1

Illie C. (2010). Identity co-construction in parliamentary discourse practices. In: C. Illie (Ed*), European parliaments under scrutiny: Discourse strategies and interaction practices* (pp.57–79). John Benjamins.

Kryvenko, A. (2018). Constructing a Narrative of European Integration in the Verkhovna Rada of Ukraine: A Corpus-Based Discourse Analysis. *Cognition, Communication, Discourse*, 17, 56–74. https://doi.org/10.26565/2218-2926-2018-17

Kryvenko, A., & Kopp, M. (2023). Workflow and metadata challenges in the ParlaMint project: Insights from building the ParlaMint-UA corpus. In K. Lindén, J. Niemi, & T. Kontino (Eds), *CLARIN Annual Conference Proceedings* (pp. 67–70). CLARIN ERIC, 2023. https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf

Krzyżanowski, M. (2010). *The Discursive construction of European identities: A Multi-level approach to discourse and identity in the transforming European Union*. Peter Lang.

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195331967.001.0001

Le, É. (2021). *Degrees of European belonging: The fuzzy areas between Us and Them*. John Benjamins.

Marchi, A. (2018). Dividing Up the Data: Epistemological, Methodological and Practical Impact of Diachronic Segmentation. In C. Taylor, & A. Marci (Eds), *Corpus approaches to discourse: A critical review* (pp. 174–196). Oxon/New York: Routledge.

Musolff, A. (2001). The Metaphorisation of European politics: Movement on the road to Europe. In A. Musolff, C. Good, R. Wittlinger, & P. Points (Eds), *Attitudes towards Europe: Language in the unification process* (pp. 179–200). Ashgate.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse:*

Theory and practice in corpus-assisted discourse studies (CADS). John Benjamins.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive grammar of the English language*. Longman.

Räikkönen, J. (2023). Leaving the EU out of the ingroup: A diachronic analysis of the use of we and us in British parliamentary debates (1973–2015). In M. Korhonen, H. Kotze, & J. Tyrkkö (Eds), *Exploring Language and Society with Big Data: Parliamentary discourse across time and space* (pp. 142–165). John Benjamins. https://doi.org/10.1075/scl.111.05rai

Riihimäki, J. (2019). At the heart and in the margins: Discursive construction of British national identity in relation to the EU in British parliamentary debates from 1973 to 2015. *Discourse & Society*, 30(4), 412–431. https://doi.org/10.1177/0957926519837394

Schäffner, C. (1996). Building a European house? Or at two speeds into a dead end? Metaphors in the debate on the united Europe. In A. Musolff, C. Schäffner, & M. Towson (Eds.), *Conceiving of Europe: Diversity in unity* (pp. 31–59). Dartmouth.

Truan, N., & Romary, L. (2021). Building, encoding, and annotating a corpus of parliamentary debates in TEI XML: A cross-linguistic account. *Journal of the Text Encoding Initiative 14*. https://doi.org/10.4000/jtei.4164

Williams, G., Piazza, R., & Giuliani, D. (2012). Nations and supernations: A tale of three Europes. In P. Bayley, & G. Williams (Eds), *European identity. What the media say* (pp. 55–83). Oxford University Press.

Wodak, R., de Cillia, R., Reisigl, M., & Liebhart, K. (2009). *The discursive construction of national identity* (2nd ed.). Edinburgh University Press.

Yavorska, H., Bohomolov, O. (2010). *Nepevnyi obyekt bazhannya: Yevropa v ukrayins'komu politychnomu dyskursi [An Uncertain object of desire: Europe in Ukrainian political discourse]*. Dmytro Burago Publ. and A. Krymskyi Institute of Oriental Studies of the National Academy of Sciences of Ukraine (in Ukrainian).

Zhabotynska, S. (2018). Images of Ukraine-EU relations in conceptual metaphors of Ukrainian mass media. In *Cognition, Communication, Discourse*, 17, 118–140. https://doi.org/10.26565/2218-2926-2018-17

NA POTI K SKLADENJSKIM ANALIZAM ŠOLSKEGA PISANJA: SKLADENJSKI VZORCI V KORPUSU ŠOLAR 3.0

Tina MUNDA,[1] Špela ARHAR HOLDT[1, 2]
[1]Filozofska fakulteta, Univerza v Ljubljani
[2]Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

1 UVOD

Razvojni korpus Šolar vsebuje besedila, ki so jih učenke in učenci slovenskih osnovnih in srednjih šol samostojno napisali pri pouku, s tem pa empirične podatke o značilnostih šolske pisne produkcije. Korpus omogoča pripravo učinkovitejših, k uporabniku usmerjenih jezikovnih učnih gradiv in pristopov ter razvoj jezikovnotehnoloških orodij, ki podpirajo pisanje in opismenjevanje. Korpus, ki je po obsegu in sestavi primerljiv jezikovnim virom za druge evropske jezike (npr. Abel in sod., 2014; Barbagli in sod., 2016; Glaznieks in sod., 2022; Granger, 1998; Pala in sod., 2003), se razvija že dobro desetletje (Rozman in sod., 2012; Kosem in sod., 2016; Arhar Holdt in Kosem, 2024) in je trenutno dostopen v različici 3.0 (Arhar Holdt in sod., 2022). V okviru projekta *Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti* (ARIS, J7-3159)[1] razvijamo metodologijo pridobivanja korpusnih podatkov za različne raziskovalne in razvojne namene, pri čemer nas zanimajo tudi podatki za skladenjske analize šolskega pisanja.

Če je še pred kratkim za slovenščino vladalo pomanjkanje skladenjsko označenih podatkov, pa tudi prosto dostopnih orodij za njihovo analizo in vizualizacijo (Ledinek, 2018), se v zadnjih letih stanje bistveno izboljšuje. Skladenjsko označevanje, ki je del označevalnika CLASSLA-Stanza 2.1 (Terčon in Ljubešić, 2023; Arhar Holdt in sod., 2023), dosega 90-odstotno natančnost skladenjskega označevanja po sistemu Universal Dependencies in skoraj 94-odstotno po sistemu JOS-SYN. (Obliko)skladenjske oznake so na voljo v različnih besedilnih korpusih za slovenščino in so bile uporabljene za napredno pridobivanje podatkov, npr. na ravni kolokacij (Krek in sod., 2022), in za jezikoslovne analize, npr. analize skladenjskih značilnosti govorjenega jezika (Dobrovoljc, v objavi), računalniško

[1]https://www.cjvt.si/prop/

posredovane slovenščine (Arhar Holdt, 2018) in udeleženskih vlog (Gantar, 2023). Na voljo so tudi zmogljiva orodja za luščenje oblikoskladenjskih in skladenjskih podatkov, kot so denimo Q-CAT (Brank, 2023), LIST (Krsnik in sod., 2019), Drevesnik (Štravs in Dobrovoljc, 2024) in novonastali STARK (Krsnik in sod., 2023), ki ga uporabljamo in opišemo tudi v tem prispevku.

V nadaljevanju predstavljamo frekvenčne sezname skladenjskih struktur iz korpusa Šolar 3.0 in orišemo postopek njihovega luščenja, ki predstavlja enega prvih poskusov pridobivanja skladenjskih podatkov z orodjem STARK. Cilj je bil pridobiti besednozvezne skladenjske vzorce, kjer je jedro vsaka izmed besednih vrst izbranega označevalnega sistema, in medstavčne vzorce: priredja, soredja in podredja, skupaj s pripadajočimi vezniki (kar ponuja možnost podrobnejšega razvrščanja gradiva in na veznikih temelječe raziskave).

2 PRIPRAVA PODATKOV

2.1 O orodju STARK

STARK[2] je prosto dostopno orodje, zgrajeno na osnovi Pythona in upravljano prek ukazne vrstice, ki iz vhodne datoteke – skladenjsko razčlenjenega korpusa (drevesnice) – proizvede izhodno datoteko – frekvenčni seznam skladenjskih struktur (dreves) (Dobrovoljc in sod., 2023). Vsebino izhodne datoteke narekujejo prilagodljivi luščilni kriteriji, ki omogočajo tako splošna kot podrobno usmerjena iskanja. Določiti je možno dolžino skladenjskega vzorca, npr. pri nastavitvi *size = 2-5* orodje vrne vse možne skladenjske strukture z 2–5 besedami (vozlišči). Iskanje lahko poljubno nadgrajujemo: določiti je možno lastnosti jedra strukture, in sicer je to lahko lema in/ali oblikoskladenjska oznaka (podprti so operatorji '|' (ali), '&' (in), '!' (ne)); prikaz vozlišč v izluščenih strukturah, ki so lahko oblike, leme, oblikoskladenjske oznake, morfološke lastnosti ali skladenjske relacije – odvisno od namena (če za prikaz določimo obliko, bo drevo v izhodu zapisano npr. *Poiščeš >dve rešitev*, če pa določimo oznako MULTEXT-East, pa *Ggdsde >dve Sozet*); iskati je možno drevesa, pri čemer dovolimo le nekatere skladenjske povezave, lahko pa poljubne relacije v izluščenih vzorcih

[2] Več o delovanju orodja in nastavitvah luščenja je na voljo na https://github.com/clarinsi/STARK in v (Dobrovoljc in sod., 2023).

tudi prepovemo. Možno je tudi iskanje s poizvedbo (ang. *query*), ki temelji na poizvedovalnem jeziku dep_search, ki ga uporablja tudi Drevesnik.[3]

STARK je bil sprva zasnovan za skladenjsko razčlenjevanje korpusov, označenih z jezikovno univerzalno shemo Universal Dependencies[4] (v nadaljevanju: UD), v različici 3.0 pa je omogočeno luščenje tudi prek drugih shem skladenjskega označevanja. Če je prejšnja verzija orodja podpirala luščenje podatkov po sistemu z natanko enim korenskim elementom (ang. *root*) v povedi (kar je inherentna značilnost skladenjske sheme UD), nova različica podpira luščenje prek shem z več koreni v povedi. Med te sodi tudi slovenščini prilagojen sistem JOS-SYN[5] (Erjavec in sod., 2010), s katerim je mdr. označen korpus Šolar.

2.2 Besednozvezne skladenjske strukture

Besednozvezne skladenjske strukture iz korpusa Šolar 3.0 smo luščili prek oblikoskladenjskih oznak (v nadaljevanju: MSD-oznake) MULTEXT-East (Erjavec, 2012). Vozlišča so v pridobljenih frekvenčnih seznamih ponazorjena z MSD-oznakami, skladenjske relacije pa so po sistemu JOS-SYN (Erjavec in sod., 2010), ki je komplementarna dopolnitev omenjenega sistema oblikoskladenjskega označevanja. Uporabili smo STARK 3.0, saj ta v nasprotju s prejšnjima različicama omogoča luščenje prek označevalne sheme z več kot enim korenskim elementom (gl. razdelek 2.1).

Pred luščenjem smo korpus predprocesirali. Orodje namreč oblikoskladenjske oznake, ki naj jih uporabi za luščenje, pričakuje v stolpcu, kjer so v izvorni datoteki Šolarja UD-oznake. Ker želimo luščiti po sistemu MULTEXT-East, smo vrednosti v tem stolpcu prepisali z oznakami MULTEXT-East iz sosednjega stolpca, vendar so bile tja shranjene le prve črke oznake, ki kažejo na besedno vrsto. Tako smo se pri določanju jedra besedne zveze v nastavitvah luščenja izognili zvrščanju vseh obstoječih MSD-oznak za vsako besedno vrsto (samostalnik jim ima 104, npr. Somei, Somer itd.). Stolpec s celotnimi MSD-oznakami je ostal nespreme-

[3] Več o sestavljanju iskalnih nizov je na voljo na: https://orodja.cjvt.si/drevesnik/help/en/ in https://github.com/clarinsi/STARK/blob/master/settings.md.
[4] https://universaldependencies.org/; več o označevalnem sistemu je na voljo na: https://wiki.cjvt.si/books/07-universal-dependencies.
[5] Več o označevalnem sistemu je na voljo na: https://wiki.cjvt.si/books/06-odvisnostna-skladnja-jos-syn.

njen, kar pomeni, da bodo v izvozu kot vozlišča skladenjske strukture prikazane celotne MSD-oznake.

V Tabeli 1 je prikazan primer besednozvezne skladenjske strukture v izhodni datoteki. V nadaljevanju predstavljamo tipe podatkov, ki jih najdemo v izvozu.

- *Tree* prikazuje eno izmed izluščenih skladenjskih relacij (odvisnostnih dreves), ki ustreza iskalnemu nizu. Predstavljena je v obliki `vozlišče A <relacija vozlišče B`,[6] kar nam pove: (a) vozlišče A je podrejeno vozlišču B (puščica nakazuje smer odvisnosti) in (b) med vozliščema A in B je skladenjsko razmerje [relacija].

- *Node A-xpos* in *Node B-xpos* izpišeta vsako vozlišče izluščene relacije posebej.

- *Absolute frequency* je absolutna pogostost izluščene relacije v korpusu.

- *Relative frequency* je relativna pogostost izluščene relacije v korpusu.

- *Order* prikazuje vrstni red vozlišč v izluščeni strukturi.

- *Number of nodes* prikazuje število vseh vozlišč v izluščeni strukturi.

- *Head node* izpiše vozlišče, ki predstavlja koren v izluščeni strukturi.

- *Example* izpiše naključno korpusno poved, v kateri je bila prepoznana skladenjska struktura iz prvega stolpca; v povedi so vozlišča strukture tudi označena, tako da se vidi, katere pojavnice zavzemajo katero mesto v skladenjski strukturi.

- *MI*, *MI3*, *Dice*, *logDice, t-score* in *simple-LL* so mere povezanosti, ki nam povejo moč statistične povezanosti med vozlišči drevesa znotraj korpusa, ki ga analiziramo. Te mere so tipično v rabi za merjenje jakosti kolokacij, lahko pa podajo uporabne informacije tudi za analizo skladenjskih struktur.

[6] Vozlišča so lahko prikazana kot besedna oblika, lema ali oblikoskladenjska oznaka.

Tabela 1: Ponazoritev najpogostejšega skladenjskega vzorca s samostalniškim jedrom v korpusu Šolar 3.0, skupaj s podatki, ki jih v izhodni datoteki vrne orodje. Tipi podatkov, ki so v izvozu v stolpcih, so za jasnejšo ponazoritev tukaj predstavljeni v vrsticah, v stolpcu pa so vrednosti tipa podatka za eno strukturo. Najpogostejši skladenjski vzorec s samostalniškim jedrom je kombinacija predloga v mestniku (Dm) in občnoimenskega samostalnika ž. sp. edn. v mestniku (Sozem).

Tree	Dm <dol Sozem
Node A-xpos	Dm
Node B-xpos	Sozem
Absolute frequency	6486
Relative frequency	3400.2
Order	AB
Number of nodes	2
Head node	Sozem
Example	Samo A[v] B[smrti] je videla rešitev.
MI	3.90
MI3	29.22
Dice	0.19
logDice	11.59
t-score	75.14
simple-LL	3121.30

Končni rezultat je 11 izvozov z besednozveznimi skladenjskimi vzorci iz korpusa Šolar, po en izvoz za vsako od besednih vrst po sistemu MULTEXT-East: samostalnik, glagol, pridevnik, prislov, zaimek, števnik, predlog, veznik, členek, medmet, okrajšava (kategoriji neuvrščeno in ločilo nista vrnili rezultatov). Vsakemu izvozu pripada še izhodna datoteka s prepoznanimi drevesi v vseh povedih korpusa.

2.3 Medstavčne skladenjske strukture

Medstavčne strukture smo luščili z verzijo orodja 2.0 na osnovi skladenjskih oznak UD, saj so oznake sistema JOS-SYN, ki so na voljo v korpusu Šolar 3.0, zasnovane za souporabo z oblikoskladenjskimi oznakami MULTEXT-East, STARK pa ne podpira možnosti, da bi jih bilo mogoče učinkovito uporabiti pri izdelavi luščilnih parametrov. Korpus smo pred luščenjem stavčnih struktur zato morali strojno označiti še s skladenjskimi relacijami po sistemu UD, s CLASSLO-Stanza 2.1 (Ljubešić in Dobrovoljc, 2019; Terčon in Ljubešić, 2023).

Za namene bodočih jezikoslovnih analiz na ravni medstavčne skladnje smo v korpusu identificirali razmerja, ki kažejo na sestavo povedi: priredje, soredje in podredje. Pri podredjih smo posebej luščili osebkov odvisnik, predmetni odvisnik in prislovne odvisnike. Slednji po sistemu skladnje UD združuje vse tipe prislovnih odvisnikov, zato v iskalni pogoj kot vozlišča vključujemo tudi veznike, s pomočjo katerih je možno natančneje urejati in analizirati izvožene podatke.

Definiranje poizvedbe za luščenje osebkovega (`csubj`), predmetnega (`ccomp`) in prislovnih odvisnikov (`advcl`) skupaj z vezniki je bilo relativno preprosto, saj v osnovi povezujejo samo povedke in ni treba dodatno definirati izvora in cilja povezave. Upoštevati pa je treba tudi večbesedne veznike. Pri luščenju smo tako uporabili tri poizvedbe, kot kaže spodnji primer za prislovne odvisnike:

a. `_ >advcl (_ >mark _)` – vrne prislovni odvisnik z veznikom (cilj relacije `mark`), čemur ustreza skladenjski vzorec v Tabeli 2.

b. `_ >advcl (_ >mark (_ >fixed _))` – dodatna relacija `fixed` povezuje dele večbesednih enot, kamor po UD spadajo tudi večbesedni vezniki; kot vozlišči sta določeni dve vezniški besedi (npr. *Jazonu je ostala zvesta in mu A[pomagala] do prestola, B[tako] C[da] je D[ubila] kralja.*).

c. `_ >advcl ((_ >fixed _ >fixed _) <mark _)` – dve relaciji `fixed` ustrezata trobesednim veznikom (npr. *Antigona je A[pokopala], B[kljub] C[temu] D[da] je E[vedela] kakšne so posledice sodbe.*).

Ker se naštete poizvedbe delno prekrivajo, se lahko iste povedi pojavijo v več izvozih. Poizvedba a. namreč določa kot vozlišče veznik (cilj povezave `mark`), ne določa pa naslednjega vozlišča (ki je lahko tudi druga beseda dvo- ali trobesednega veznika). Tako se lahko v rezultatu a. pojavi tudi stavek z dvobesednim veznikom, pri čemer prva beseda ima vozlišče, druga pa ne. Primer v Tabeli 2 je rezultat poizvedbe a., vendar isto poved najdemo tudi v izvozu b., kjer sta vozlišči obe besedi veznika. Zato smo uvedli fazo poprocesiranja, kjer smo iz predstavljenih treh izvozov izbrisali duplikate in obdržali le unikatno poved z največjim številom vezniških vozlišč.

Tabela 2: Ponazoritev skladenjskega vzorca s prislovnim odvisnikom v korpusu Šolar 3.0, skupaj s podatki, ki jih v izhodni datoteki vrne orodje. Vzorec (*Tree*), predstavljen z UD-oznakami in relacijami kot vozlišči, prinaša povezavo glagola glavnega stavka z glagolom podrednega stavka (advcl je povezava za prislovni odvisnik), slednji pa je povezan s podrednim veznikom (cilj relacije mark). Tipi podatkov, ki so v izvozu v stolpcih, so za jasnejšo ponazoritev tukaj predstavljeni v vrsticah, v stolpcu pa so vrednosti tipa podatka za eno strukturo; za razlago tipov podatkov gl. razdelek 2.2.

Tree	(ADV <mark VERB) <advcl VERB
Node A-upos	ADV
Node B-upos	VERB
Node C-upos	VERB
Absolute frequency	212
Relative frequency	111.1
Order	ABC
Number of nodes	3
Head node	VERB
Example	A[Medtem], ko sta se B[prepirala], je vonj po piščancu C[zavohal] tudi jež.
MI	-2.69
MI3	12.77
Dice	0,00
logDice	4.23
t-score	-79.36
simple-LL	1967.72

Definiranje poizvedbe za priredje (`conj`) in soredje (`parataxis`) je bilo zahtevnejše, saj ti relaciji po shemi UD označujeta tudi druge, ne nujno stavčne strukture. Tako je bilo treba za pridobitev priredij in soredij v iskalnem pogoju določiti tudi izvor in cilj povezave, ki morata biti glagolska. Pokriti je bilo treba tako stavke s polnopomenskim glagolom (upos=VERB) kot tiste s povedkovim določilom (ta vzorec smo definirali s povezavo `cop`, ki ima na izvoru relacije vezni glagol *biti*, na cilju pa besedo v vlogi povedkovega določila). Da smo pokrili vse možne kombinacije, smo uporabili štiri poizvedbe, kot kaže spodnji primer za soredja:

a. `upos=VERB >parataxis upos=VERB` za soredja, kjer je polnopomenski glagol (VERB) na izvoru in cilju povezave.

b. `upos=VERB >parataxis (_ <cop _)` za soredja, kjer je na izvoru povezave polnopomenski glagol, na cilju pa povedkovo določilo.

c. `_ >cop _ >parataxis upos=VERB` za soredja, kjer je na izvoru povezave povedkovo določilo, na cilju pa polnopomenski glagol.

d. `_ >cop _ >parataxis (_ <cop _)` za soredja, kjer je povedkovo določilo na izvoru in cilju povezave.

Pri luščenju priredij smo dodali še vozlišča za veznike, ki so spet lahko eno- ali večdelni. Za ponazoritev navajamo poizvedbo za priredje, kjer je na izvoru povezave polnopomenski glagol, na cilju povedkovo določilo, povezana pa sta z dvodelnim prirednim veznikom: `upos=VERB >conj ((_ <fixed _) <cc _ <cop _)`. Tudi pri izvozih priredij smo izvedli zgoraj opisani postopek deduplikacije. To vpliva na statistično sliko izvoženih podatkov, kar je treba upoštevati pri interpretaciji rezultatov.

Končni rezultat je 5 izvozov z medstavčnimi skladenjskimi vzorci iz korpusa Šolar. Več izvozov za istovrstno medstavčno relacijo, ki so nastali zaradi podrobne obravnave veznikov ali vključitve povedkovega določila kot jedra stavka, je bilo po fazi poprocesiranja združenih v enega. Vsakemu izvozu pripada še izhodna datoteka s prepoznanimi drevesi v vseh povedih korpusa.

2.4 Dostopnost podatkov

Izvozi skladenjskih struktur iz korpusa Šolar 3.0, ki so v obliki frekvenčnih seznamov, bodo po koncu projekta na voljo na repozitoriju CLARIN.SI. Dodali bomo tudi konfiguracijske datoteke z nastavitvami luščenja in predlog za analizo oz. vizualizacijo podatkov s programom MS Excel.

2.5 Omejitve

Uporaba predstavljene metodologije luščenja skladenjskih korpusnih podatkov predpostavlja dobro poznavanje označevalnih shem UD in JOS-SYN in osnovno uporabo ukazne vrstice za zagon programa oz. uporabo uporabniku prijaznejšega integriranega razvojnega okolja (ang. *integrated development environment;*

IDE), kot je npr. PyCharm. Nadalje je za luščenje skladenjskih dreves treba usvojiti poizvedovalni jezik dep_search, ki ga uporablja STARK. Osnovni napotki so na voljo na strani GitHub.[7] Ker poizvedovalni jezik temelji na poizvedovalnem jeziku Drevesnika, spletnega vmesnika za iskanje po slovenskih skladenjskih drevesnicah, je dobro poznati tudi tega.[8]

Pri orodju STARK je za naš scenarij rabe umanjkala možnost uporabe operatorja za negacijo ('!') pred puščico (pri sestavi poizvedbe (ang. *query*)). Na voljo je negacija relacije, ki v primeru `_ >!fixed _` prepoveduje pojav druge besede večbesedne enote, ni pa možnosti negacije katere koli podrejene relacije. Npr. z iskalnim nizom `_ >advcl (_ >mark (_ !>fixed _))` bi izločili vse večbesedne veznike oz. izluščili le enobesedne veznike. Ta možnost bi v našem primeru zmanjšala število poizvedb in s tem čas ter procesno breme, hkrati pa ne bi bilo potrebno poprocesiranje izvozov z vezniki (gl. razdelek 2.3).

Upoštevati je treba, da luščimo na osnovi strojno označenih korpusnih podatkov, ki lahko vsebujejo napake v pripisanih skladenjskih povezavah. Upoštevati je treba tudi specifike sheme UD. Primerjalni odvisniki s 'kot', npr. *Kakor vem je A[več] pogovora o problemu, B[kot] C[pa] da bi bili takoj D[nasilni]*, so označeni kot prislovni odvisniki (`advcl`), pri čemer pa je izvor povezave lastnost, na katero se primerjava nanaša, in ne jedro povedka glavnega stavka, kot velja pri drugih primerih. Za navedeno poved se tako na mestu, kjer bi pričakovali glagol, pojavi kot vozlišče 'več' (označen kot izvor povezave `advcl`).

3 NADALJNJI KORAKI

Naslednji korak je pregled skladenjskih podatkov iz korpusa Šolar 3.0 in njihova ureditev glede na stopnjo šolanja. Po predstavljeni metodologiji bo opravljeno tudi luščenje iz korpusa učbenikov (Kosem in sod., 2022), kar bo omogočilo primerjalne analize produkcijskega in pričakovanega recepcijskega dometa na različnih stopnjah šolanja. Izvoženi, urejeni in dokumentirani podatki, ki jih je v bodoče mogoče bogatiti še z drugimi informacijami, npr. merami skladenjske kompleksnosti, bodo skupnosti odprto na voljo in bodo kot taki lahko služili kot empirična podstat za različne namene, od korpusnih skladenjskih analiz, pre-

[7] https://github.com/clarinsi/STARK/blob/master/README.md, https://github.com/clarinsi/STARK/blob/master/settings.md
[8] https://orodja.cjvt.si/drevesnik/

verjanj hipotez o značilnostih šolskega pisanja do priprave didaktičnih smernic, učnih gradiv in drugih aplikativnih ciljev.

ZAHVALA

Prispevek je nastal v okviru raziskovalnega projekta *projekta Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti* (J7-3159) in raziskovalnega programa Jezikovni viri in tehnologije za slovenski jezik (P6-0411), ki ju financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS). Zahvaljujeva se dr. Kaji Dobrovoljc za pomoč pri delu s programom STARK in recenzentom za koristne predloge.

LITERATURA

Abel, A., Glaznieks, A., Nicolas, L. in Stemle, E. (2014). Koko: An l1 learner corpus for german. V *Proceedings of the 9th international conference on language resources and evaluation* (str. 2414–2421). Reykjavik, Iceland.

Arhar Holdt, Š. (2018). Korpusni pristop k skladnji računalniško posredovane slovenščine. V D. Fišer (Ur.), *Viri, orodja in metode za analizo spletne slovenščine* (Zv. 1, str. 228–252). Ljubljana: Znanstvena založba Filozofske fakultete. https://doi.org/10.4312/9789610600701

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Erjavec, T., Gantar, P., Krek, S., ... Žitnik, S. (2023). Nadgradnja učnega korpusa ssj550k v SUK 1.0. V Š. Arhar Holdt in S. Krek (Ur.), *Razvoj slovenščine v digitalnem okolju* (str. 119–156). Ljubljana: Založba Univerze. https://doi.org/10.4312/9789612972561

Arhar Holdt, Š. in Kosem, I. (2024). Šolar, the developmental corpus of slovene. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-024-09758-4

Arhar Holdt, Š., Rozman, T., Stritar Kučuk, M., Krek, S., Krapš Vodopivec, I., Stabej, M., ... Kosem, I. (2022). *Developmental corpus šolar 3.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1589

Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S. in Venturi, G. (2016). Cita: An l1 italian learners corpus to study the development of writing competence. V *Proceedings of the 10th international conference on language resources and evaluation* (str. 88–95). Portorož, Slovenia.

Brank, J. (2023). *Q-CAT corpus annotation tool 1.5.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1844

Dobrovoljc, K. (v objavi). Uporaba drevesnice sst v raziskavah govorjene slovenščine:

prednosti in omejitve. *Jezik in slovstvo*.

Dobrovoljc, K., Krsnik, L. in Šikonja, M. R. (2023). Stark: a tool for dependency tree extraction and analysis. V *Unidive 1st general meeting: universality, diversity and idiosyncrasy in language technology* (str. [1-3]). Paris-Saclay University. https://unidive.lisn.upsaclay.fr/lib/exe/fetch.php?media=meetings:2023-saclay:abstracts:62_dobrovoljc_et_al_stark_a_tool_for_dependency_tree.pdf

Erjavec, T. (2012). Multext-east: morphosyntactic resources for central and eastern european languages. *Language Resources and Evaluation*, *46*(1), 131–142. https://doi.org/10.1007/s10579-011-9174-8

Erjavec, T., Fišer, D., Krek, S. in Ledinek, N. (2010). The jos linguistically tagged corpus of slovene. V *Proceedings of the seventh conference on international language resources and evaluation (lrec'10).* Valeta, Malta. https://aclanthology.org/L10-1087/

Gantar, P. (2023). Analiza udeleženskih vlog s skladenjskega, pomenskega in leksikalnega vidik. V M. Smolej in M. Schlamberger Brezar (Ur.), *Prispevki k preučevanju slovenske skladnje* (Zv. 1, str. 77–97). Ljubljana: Založba Univerze. https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/421/764/8503

Glaznieks, A., Frey, J. C., Stopfner, M., Zanasi, L. in Nicolas, L. (2022). Leonide: A longitudinal trilingual corpus of young learners of italian, german and english. *International Journal of Learner Corpus Research*, *8*(1), 97–120.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for sla research. V S. Granger (Ur.), *Learner english on computer* (str. 3–18). Addison Wesley Longman.

Kosem, I., Pori, E., Žagar, A. in Arhar Holdt, Š. (2022). *Corpus of slovenian textbooks ccUčbeniki 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1693

Kosem, I., Rozman, T., Arhar Holdt, Š., Kocjančič, P. in Laskowski, C. A. (2016). Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov. V T. Erjavec in D. Fišer (Ur.), *Zbornik konference jezikovne tehnologije in digitalna humanistika* (str. 95–100). Ljubljana: Znanstvena založba Filozofske fakultete. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Kosem-et-al_Solar-2-0-nadgradnja-korpusa-solskih-pisnih-izdelkov.pdf

Krek, S., Gantar, P. in Kosem, I. (2022). Extraction of collocations from the gigafida 2.1 corpus of slovene. V A. Klosa (Ur.), *Euralex 2022, proceedings of the xx euralex international congress* (str. 240–252). Mannheim, Germany: IDS-Verlag. https://euralex.org/publications/extraction-of-collocations-from-the-gigafida-2-1-corpus-of-slovene/

Krsnik, L., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Ključevšek, A., Krek, S. in Robnik-

Šikonja, M. (2019). *Corpus extraction tool LIST 1.2.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1276

Krsnik, L., Dobrovoljc, K. in Robnik-Šikonja, M. (2023). *Dependency tree extraction tool STARK 2.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1899

Ledinek, N. (2018). Skladenjska analiza slovenščine in slovenski jezikoslovno označeni korpusi. *Jezik in slovstvo, 63*(2/3), 103–116. https://doi.org/10.4312/jis.63.2-3.103-116

Ljubešić, N. in Dobrovoljc, K. (2019, August). What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. V *Proceedings of the 7th workshop on balto-slavic natural language processing* (str. 29–34). Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-3704

Pala, K., Rychlý, P. in Smrž, P. (2003). Text corpus with errors. V V. Matoušek in P. Mautner (Ur.), *Text, speech and dialogue (tsd 2003) lecture notes in computer science* (Zv. 2807, str. 90–97). Springer.

Rozman, T., Stritar, M. in Kosem, I. (2012). Šolar – korpus šolskih pisnih izdelkov. V T. Rozman in sod. (Ur.), *Empirični pogled na pouk slovenskega jezika.* Ljubljana: Trojina, digitalna različica Založba Univerze v Ljubljani. https://doi.org/10.4312/9789610603511

Štravs, M. in Dobrovoljc, K. (2024). *Service for querying dependency treebanks drevesnik 1.1.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1923

Terčon, L. in Ljubešić, N. (2023). *Classla-stanza: The next step for linguistic processing of south slavic languages.* arXiv. https://doi.org/10.48550/arXiv.2308.04255

KOST 2.0: PREDSTAVITEV KORPUSA IN POTEK OZNAČEVANJA JEZIKOVNIH NAPAK

Mojca STRITAR KUČUK
Univerza v Ljubljani, Filozofska fakulteta

1 UVOD

Pisni korpus slovenščine kot tujega jezika KOST od leta 2019 nastaja na Filozofski fakulteti Univerze v Ljubljani. Prva različica korpusa, KOST 1.0, je bila objavljena spomladi 2023, njena zasnova in vsebina sta podrobno dokumentirani v Stritar Kučuk, 2024.[1] Jeseni 2023 je bil objavljen povečani in dodatno označeni KOST 2.0.[2] Ker gre za zbirko besedil neprvih govorcev slovenščine, za katera so značilne pogoste jezikovne napake, je bil eden od pomembnih vidikov snovanja in gradnje tega korpusa označevanje teh napak (Granger, 2003). V tem prispevku bosta zato predstavljena nadgrajeni KOST in jezikoslovni pogled na nekatere vidike označevanja napak v njem.

2 KOST 2.0

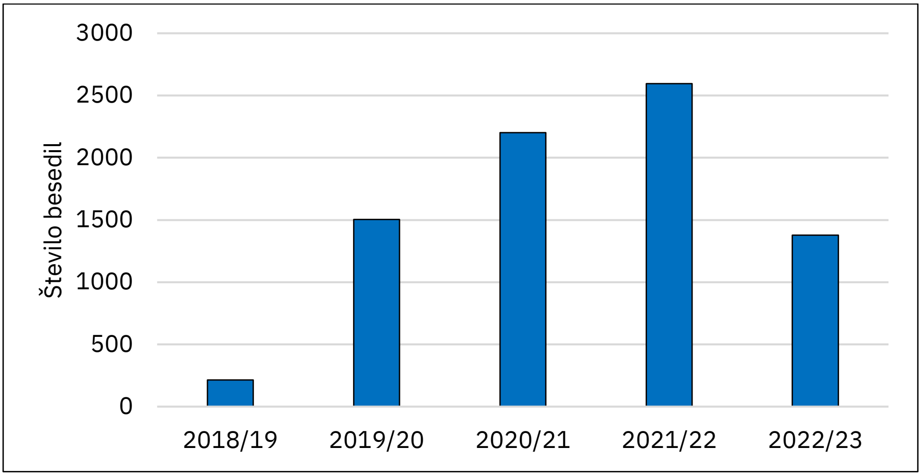
Trenutna različica korpusa, KOST 2.0, obsega 8347 besedil oz. 1.514.476 pojavnic. Zbiranje besedil se je pričelo v študijskem letu 2018/19 (Grafikon 1).[3]

[1] Korpus KOST 1.0 je bil razvit v okviru projekta Razvoj slovenščine v digitalnem okolju, ki sta ga financirala Ministrstvo za kulturo Republike Slovenije in Evropski sklad za regionalni razvoj.

[2] Nadgradnja v KOST 2.0 je potekala s sredstvi Ministrstva za kulturo Republike Slovenije, prim. https://www.cjvt.si/korpus-kost/.

[3] Številka za študijsko leto 2022/23 je nižja, saj se je zbiranje besedil zaključilo sredi leta.

Grafikon 1: Količina zbranih besedil po študijskih letih.



Ker je za vključitev pisnih besedil v korpus nujno, da njihovi tvorci podpišejo soglasje,[4] so bila vsa besedila za KOST pridobljena v različnih programih za organizirano poučevanje slovenščine kot drugega oz. tujega jezika, v katerih smo lahko dobili tudi ta soglasja: dobre tri četrtine v modulu Leto plus,[5] preostanek pa na različnih tečajih in lektoratih Centra za slovenščino kot drugi in tuji jezik v Sloveniji in po svetu.[6] Pri pridobivanju besedil je sodelovalo več kot 40 pedagoških in drugih sodelavcev teh programov.

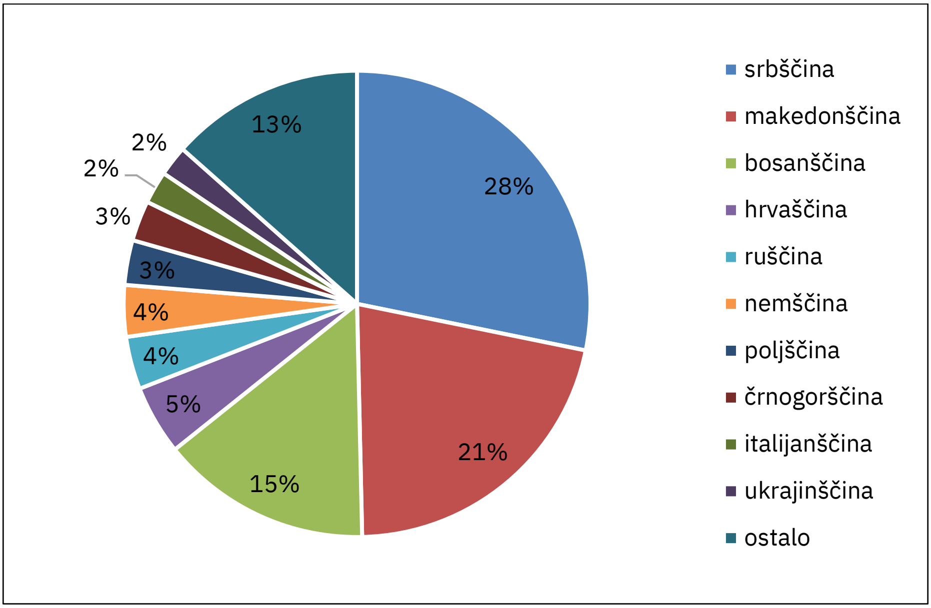
V KOST vključena besedila so napisali 1303 različni tvorci, od tega sta dve tretjini žensk in tretjina moških. Vsi so anonimizirani, njihovi osebni podatki, ki se pojavijo v besedilih, pa so zakriti s kodami, npr. »Moje ime je [XImeX] [XPriimekX], in sem študentka prvega letnika«. Tvorci govorijo več kot 30 različnih prvih jezikov, največ jih je z južnoslovanskega govornega območja (Grafikon 2).

[4] V prispevku je moški spol uporabljen kot slovnično nevtralen.

[5] Prim. https://www.uni-lj.si/studij/leto-plus/.

[6] Prim. https://centerslo.si/.

Grafikon 2: Delež tvorcev besedil glede na njihov prvi jezik.

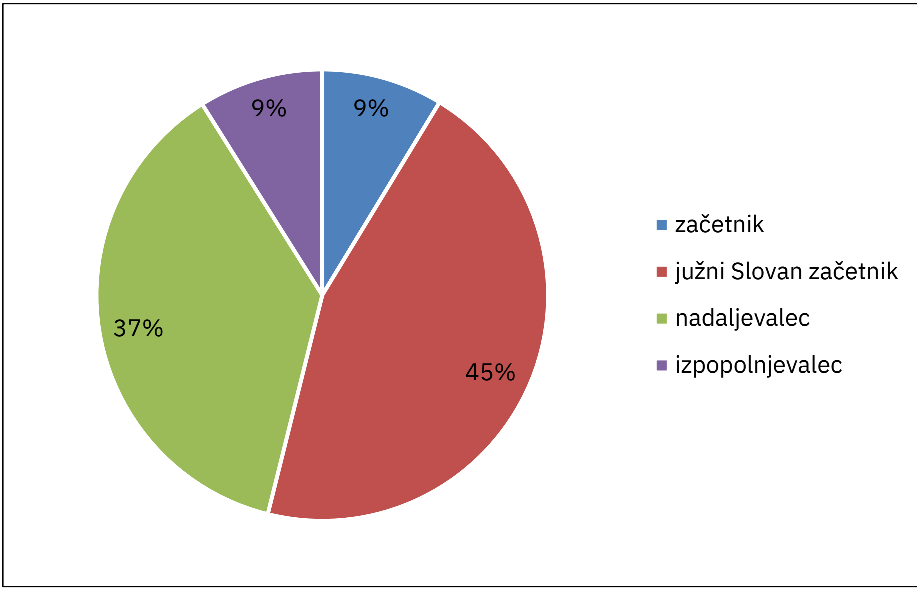


Med jeziki, navedenimi v kategoriji »ostalo« na Grafikonu 2, so španščina, nizozemščina, angleščina, slovaščina, češčina, kitajščina, slovenščina (v zamejstvu), madžarščina, japonščina, francoščina, romunščina, bolgarščina, portugalščina in grščina. Po dva ali en govorec govori albansko,[7] hebrejsko, igbojsko, indonezijsko, kirgiško, kirunsko, korejsko, švedsko in turško.

Poleg prvega jezika je za tiste, ki raziskujejo slovenščino kot neprvi jezik, pomemben podatek o tvorčevi trenutni jezikovni zmožnosti (Grafikon 3). Ta je zabeležena ob vsakem besedilu in označena s štirimi stopnjami: začetnik, nadaljevalec, izpopolnjevalec, južni Slovan začetnik (prim. Stritar Kučuk, 2024, str. 100–101). Vendar je treba upoštevati, da gre zgolj za približno oceno, namenjeno grobi orientaciji med besedili.

[7] Govorci albanščine predstavljajo eno od večjih priseljenskih skupin v Sloveniji (Razpotnik, 2024). Ker pa se ne udeležujejo organiziranih oblik jezikovnega poučevanja na Filozofski fakulteti Univerze v Ljubljani, graditelji KOST-a do besedil, ki jih morda pišejo v slovenščini, nimamo dostopa.

Grafikon 3: Deleži besedil glede na ocenjeno stopnjo jezikovne zmožnosti tvorcev.



Gradivo iz korpusa KOST je lematizirano in oblikoskladenjsko označeno, dostopno je v konkordančnikih noSketch[8] in KonText[9] ter v repozitoriju Clarin.[10] Izpostaviti pa moram novo razviti spletni konkordančnik,[11] ki omogoča uporabniku prijazen prikaz konkordanc tako v besedilih z označenimi jezikovnimi napakami kot na neoznačenih besedilih, skupaj z relevantnimi metapodatki. Nazorno so prikazane povezave med izvornim, tvorčevim besedilom in njegovo popravljeno verzijo ter oznake vrste jezikovnih napak. Poleg tega je mogoče iskanje po tipih napak (Slika 1), po okolici in po seznamih. Konkordančnik je prosto dostopen[12] in z manjšimi prilagoditvami uporaben tudi za druge korpuse, ki vsebujejo jezikovne popravke. Med drugim je uporabljen tudi za slovenski korpus šolskih besedil Šolar 3.0.[13]

[8] Prim. https://www.clarin.si/ske/#dashboard?corpname=kost20_orig.

[9] Prim. https://www.clarin.si/kontext/query?corpname=kost20_orig.

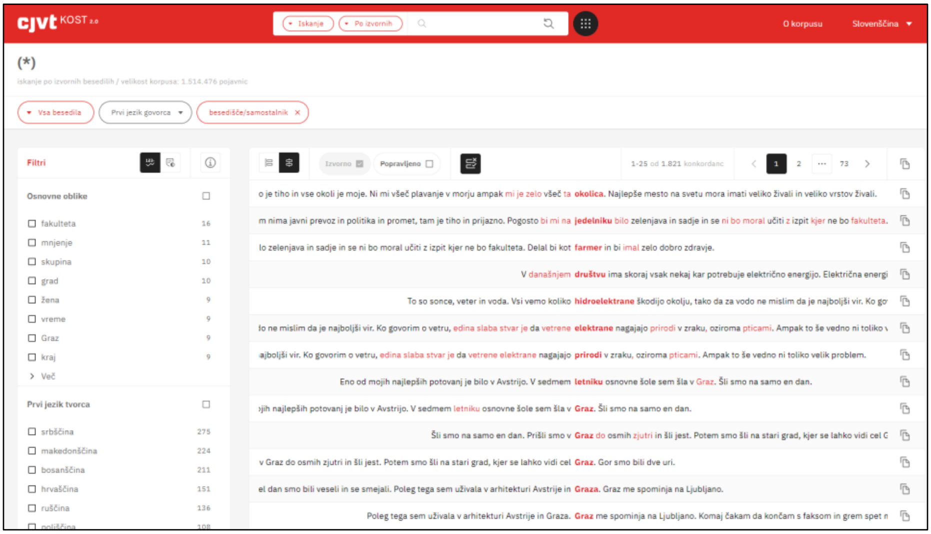
[10] Prim. http://hdl.handle.net/11356/1887.

[11] Prim. https://kost.cjvt.si/.

[12] Prim. https://github.com/TheRokUrbanc/Kost-CJVT .

[13] Prim. https://solar.cjvt.si/.

Slika 1: Prikaz iskanja po napakah besedišča pri samostalniku.



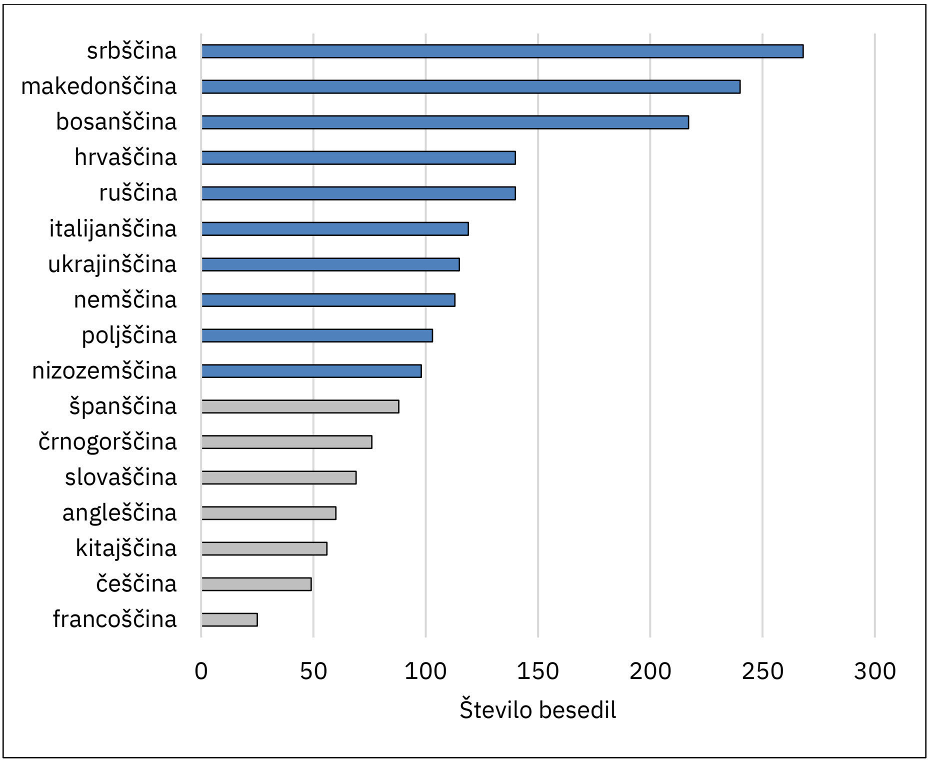
3 KRITERIJI ZA IZBIRO BESEDIL ZA OZNAČEVANJE JEZIKOVNIH NAPAK

Označevanje jezikovnih napak v besedilih KOST-a sem zasnovala in usklajevala urednica korpusa in avtorica tega prispevka. Sprva sem besedila, na katerih sem označila napake, izbirala naključno. S povečanjem količine označenih besedil in vključevanjem drugih označevalcev pa je izbor začel potekati po dveh kriterijih. Oba sta neposredno povezana z izkušnjami in potrebami učiteljev slovenščine kot drugega jezika, ki so ena od pomembnejših skupin uporabnikov KOST-a.

3.1 Prvi jezik tvorca

Prvi jezik tvorca je poleg predznanja ključni dejavnik, na podlagi katerega organizatorji jezikovnega poučevanja za slovenščino kot drugi jezik razvrščajo udeležence v skupine (Klinar in sod., 2022, str. 187), saj vpliva na didaktiko poučevanja, npr. pri odpravljanju posledic negativnega jezikovnega prenosa. Zato je to, podobno kot v številnih obstoječih korpusih usvajanja (prim. Mikelić Preradović, 2020; Rosen in sod., 2020; Volodina in sod., 2019; Gilquin, 2015, str. 17; Hammarberg, 2010), tudi najpomembnejši kriterij v korpusu KOST. Pri izbiri besedil za označevanje napak smo se omejili na jezike, za katere imamo dostop do zadostnega števila besedil, po možnosti so čim bolj raznoliki, hkrati pa so relevantni z vidika organiziranega poučevanja slovenščine.

Grafikon 4: Velikost podkorpusov z označenimi jezikovnimi napakami glede na prvi jezik tvorca.

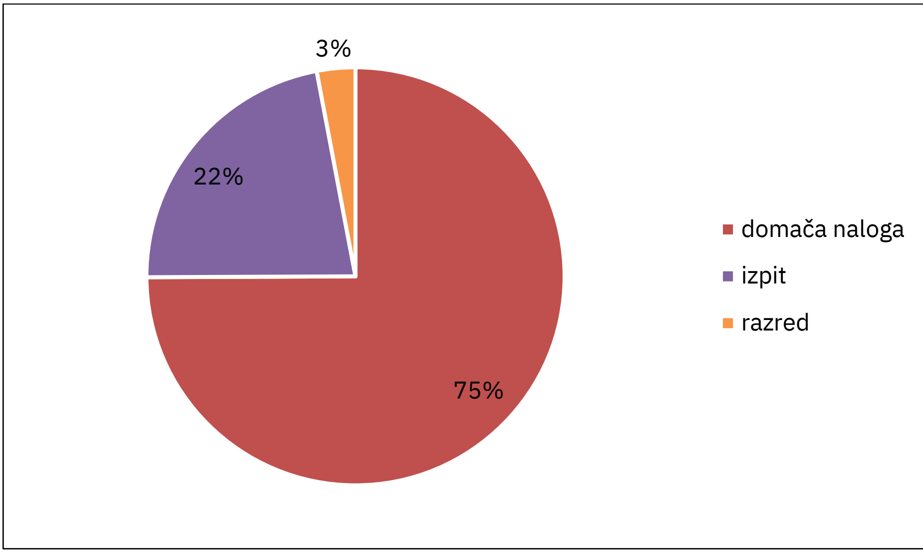


Kot je razvidno z Grafikona 4, je trenutno največ označenih besedil govorcev srbščine, makedonščine in bosanščine. Da bi bile mogoče čim bolj zanimive in zanesljive primerjave tudi za govorce jezikov iz drugih jezikovnih skupin, je bil postavljen cilj vsaj sto označenih besedil za govorce vsakega prvega jezika. Le s približno enako velikimi podkorpusi bodo namreč ti uravnoteženi in s tem primerljivi. Pri tem je uravnoteženo število besedil zgolj pragmatičen, najlažje avtomatsko dostopen podatek. Sam po sebi še ne zagotavlja primerljivih velikosti podkorpusov, saj so besedila lahko zelo različnih dolžin. Na Grafikonu 4 so s sivo barvo označeni prvi jeziki, pri katerih želimo pridobiti še dodatna besedila in pri katerih bo to v prihodnosti zaradi razmeroma stalnega dotoka besedil predvidoma izvedljivo. Če se bo občutno povečal delež besedil govorcev še katerega prvega jezika, pri katerem trenutno nimamo večje količine zbranih tekstov, pa bo seveda vključen v označeni del korpusa.

3.2 Okoliščine nastanka besedil

Drugi kriterij za izbor tekstov, ki se je pokazal kot ključen med samim procesom označevanja in ga graditelji ostalih primerljivih korpusov izrecno ne izpostavljajo (prim. Mikelić Preradović, 2020; Rosen in sod., 2020; Volodina in sod., 2019; Hammarberg, 2010), so okoliščine njihovega nastanka. Skoraj tri četrtine v KOST vključenih besedil so domače naloge (Grafikon 5), ki jih tvorci napišejo doma, brez nadzora učitelja. Večinoma so napisane v digitalni obliki in s tem enostavne za vključitev v digitalno zbirko besedil. Bistveno manj je besedil z izpitov, ki nastajajo v kontroliranih pogojih, na internih izpitih ali testih na tečajih ali lektoratih slovenščine, in besedil, napisanih v okviru različnih dejavnosti med samim jezikovnim poukom. Čeprav so takšna besedila napisana na roko in jih je za KOST treba ročno digitalizirati, jim dajemo prednost pri označevanju napak. Pri teh besedilih so namreč okoliščine, v katerih so nastala, transparentne in enake za vse. V besedilih prvega tipa posebej pri manj motiviranih tvorcih skozi leta opažamo naraščanje uporabe strojnih prevajalnikov in drugih jezikovnih orodij (Stritar Kučuk, 2024, str. 98), zato se postavlja vprašanje, koliko odražajo dejansko jezikovno zmožnost njihovih tvorcev v slovenščini (Stritar Kučuk, 2023c). Tudi zaradi tega v označeni del KOST-a ne vključujemo doma napisanih besedil, ki so popolnoma brez napak ali z zelo malo napakami, saj je pri njih velika verjetnost, da so bila napisana z (izključno) uporabo strojnih prevajalnikov.

Grafikon 5: Deleži besedil v KOST-u glede na okoliščine nastanka.



4 POSTOPEK OZNAČEVANJA JEZIKOVNIH NAPAK

Za razliko od večine korpusnih oznak je oznake jezikovnih napak v korpuse treba dodati ročno. To ne velja samo za KOST; po dostopnih podatkih še ni korpusa usvajanja za druge jezike, v katerem bi ta proces zadovoljivo potekal avtomatsko (prim. Rosen in sod., 2020). Takšno označevanje je zamudno, zato so napake običajno označene samo na delu korpusa. V KOST-u 2.0 je to ok. 1740 besedil oziroma slabih 24 % vseh besedil.

4.1 Načelo minimalnega popravka

Vsaka neustrezna pojavitev v korpusu dobi oznako glede na taksonomijo napak (Tabela 1), pripisana pa ji je ciljna hipoteza oz. popravljena/normalizirana oblika (Lüdeling, Hirschmann, 2015, str. 143). Pri tem se držimo načela minimalnega popravka (Volodina in sod., 2019, str. 81): s popravki čim manj posegamo v besedilo in popravljamo čim manj napak, pa čeprav popravljena verzija ne zveni stoodstotno tako, kot bi jo tvorili domači govorci ciljnega jezika (Granger in sod., 2022, str. 3). Popravljamo predvsem napake na ravni zapisa, besedišča in oblike besed. V skladnjo skušamo posegati čim manj, prav tako se izogibamo stilističnim popravkom.

Tabela 1: Taksonomija jezikovnih napak v korpusu KOST.

Krovna kategorija	*Kategorija napake*	*Primer iz KOST-a*[14]
Napake zapisa	Ločilo	Naslednje jutro**,** babica se ne počuti zelo dobro > Naslednje jutro se babica [...]
	Črkovanje	**uporavljam** slovnice > uporabljam slovnice
	Skupaj oz. narazen	sem **več krat** padla > sem večkrat padla
	Mala oz. velika začetnica	vidiš Trst in **Hrvaško** obalo > [...] in hrvaško obalo
	Krajšave	z jogurtom, **i. t. n.** > [...] z jogurtom, itn.
Napake besedišča	Samostalnik	Potovale smo v **Venecijo** > [...] v Benetke
	Glagol	Radijo **slušamo** > Radio poslušamo
	Zaimek	odločil sem se da nadaljujem **moj** študij > [...] nadaljujem svoj študij
	Pridevnik	**amerkanski** zdravniki > ameriški zdravniki
	Prislov	**Izprva** sem bil bolj skeptičen > Sprva [...]
	Predlog	grem **v** sprehod > grem na sprehod
	Veznik	Upam da ne bo težav naslednje leto **kdaj** pridem > [...] leto, ko pridem
	Ostalo	**Več** znam govoriti Englesko > Že znam [...]
Napake oblike	Samostalnik	sem bila v samoizolaciji 28 **dna** > [...] dni
	Glagol	sem še **živil** > sem še živel
	Zaimek	Po **mom** mnenju > Po mojem mnenju
	Pridevnik	Film mi je bil všeč ker ni bil **težki** > [...] težek
	Prislov	**najraji** igram šah > najraje [...]
	Ostalo	Čez **štirimi** leta > Čez štiri leta
Napake skladnje	Struktura	Okoli 8 **mi je prišel** fant > Okoli 8 je k meni prišel fant
	Besedni red	**Tudi rada grem** v nakupovalno središče > Rada grem tudi v nakupovalno središče
	Izpuščeni jezikovni elementi	Druge jezike učimo, ker je to zanimivo > Druge jezike **se** učimo [...]
	Odvečni jezikovni elementi	priporočam da **si** ga obiščete > priporočam da ga obiščete

[14] Vsi primeri v članku so iz KOST-a. V tej tabeli je napaka odebeljena, za znakom > pa je navedena popravljena oblika.

Kadar se zgodi, da napačni pojavitvi ne znamo pripisati popravljene, to označimo s tremi vprašaji v oglatem oklepaju: *[???]*. Takih primerov je malo, v KOST-u 2.0 le 160, npr. *To mi izboljša kakovost življenja, da lahko izgubim svoje obveznosti in ostanem v* ***čaravni*** *sreči*.

Pri označevanju napak se srečamo tudi z vprašanjem, kateri je ciljni jezik korpusa oz. norma, h kateri naj bi težili tvorci v korpus vključenih besedil. Čeprav se pri korpusu slovenščine odgovor zdi preprost, se zaplete npr. pri pogovornih izrazih. Lahko gre samo za vprašanje zapisa, npr. *tut* namesto *tudi*, *zgubiti* nam. *izgubiti*, ali dejanske leksikalne izbire, npr. *štamprl* namesto *šilce*. Ker je načeloma ciljni jezik tistih, ki se učijo slovenščino kot drugi oz. tuji jezik v eni od organiziranih oblik poučevanja, standardna oz. knjižna slovenščina, so tovrstne pojavitve v KOST-u označene kot napake (podobno je v češkem korpusu CzeSL, prim. Rosen in sod., 2020, str. 68). Pogovorno zaznamovano besedišče, npr. *faks, bus, flet*, pa dopuščamo v manj formalnih besedilnih vrstah, npr. v elektronskih pismih.

4.2 Označevalci jezikovnih napak

V uvodnih fazah sem jezikovne napake v besedilih iz KOST-a označevala sama, sčasoma pa se je ekipa označevalcev širila. V označevanje so se vključili študenti tretjega letnika slovenistike, ki so v študijskih letih 2021/22, 2022/23 in 2023/24 v okviru seminarja pri predmetu Slovenščina kot tuji jezik označili manjšo količino korpusnih besedil. V treh letih je sodelovalo 54 študentov, ki so skupaj označili 197 besedil. Čeprav so delo ocenili pozitivno – kot zanimivo, a razmeroma zahtevno – se tovrstni način pridobivanja označenih korpusnih tekstov ni izkazal za učinkovitega. V besedilih, ki so jih označili, je bilo zaradi njihove tendence k pretiranemu (stilističnemu) popravljanju besedil, slabega znanja slovenskega pravopisa in slovnice ter nepozornosti in naglice pri delu toliko neustreznih oznak, da jih ni bilo mogoče vključiti neposredno v označeni KOST, pač pa jih je bilo treba pozorno pregledati in/ali na novo označiti (prim. Stritar Kučuk, 2023b).

Nazadnje je jezikovne napake v besedilih za KOST označevalo 12 sodelavcev: profesorjev slovenistike, ki poučujejo slovenščino kot drugi jezik, in študentov slovenistike, ki so se med seminarskim delom pokazali kot uspešnejši. Pred začetkom dela so bili vsi deležni individualnega usposabljanja, med označevanjem pa so si pomagali s priročnikom za označevanje napak (Stritar Kučuk, 2023a), v katerem so natančnejša navodila za označevanje, odločanje v primeru dvoumnosti in primeri neustreznih oznak napak. Pokazalo se je, da je pri tovrstnem delu nujna določena mera izkušenj s slovenščino kot neprvim jezikom. Zaradi konsistentnosti oznak je bilo nujno tudi, da sem vsa označena besedila še enkrat pregledala urednica KOST-a in poenotila ter popravila oznake.

4.3 Potek dela z orodjem za označevanje napak

Označevanje jezikovnih napak v tujih korpusih poteka s pomočjo različnih orodij oz. aplikacij, ki so bodisi razvite za potrebe točno določenega korpusa (za češki korpus CzeSL so razvili program Feat, Rosen in sod., 2020, str. 173–174) bodisi gre za prilagojene aplikacije obstoječih korpusov. Napake v hrvaškem korpusu CrolTec so denimo označene v okolju TEITOK, ki je uporabljen tudi v portugalskem korpusu učečih se COPLE2 (Mikelić Preradović, 2020, str. 910). Za ta način smo se odločili tudi pri KOST-u. Napake označujemo v aplikaciji Svala,[15] ki je bila razvita za potrebe švedščine kot tujega jezika v okviru projekta SweLL (prim. Wirén in sod., 2018), nato pa v okviru projekta Razvoj slovenščine v digitalnem okolju prilagojena slovenskim specifikam (prim. Arhar Holdt in sod., 2022). Svala omogoča jasno vizualizacijo, dodajanje metajezikovnih oznak ter vzporeden prikaz izvornih in popravljenih besedil, poleg tega pa je odprto dostopna in uporabna za več projektov.[16]

Označevanje besedil za KOST s pomočjo orodja Svala poteka po naslednjem postopku:

- Označevalec dobi besedila, ki jih mora označiti, vsako v posebni tekstovni datoteki (format *txt*), v anonimizirani obliki in brez kakršnih koli metapodatkov. Ime datoteke je enako kodi besedila.
- V Svali označevalec označi vsak odstavek besedila posebej. Če gre za

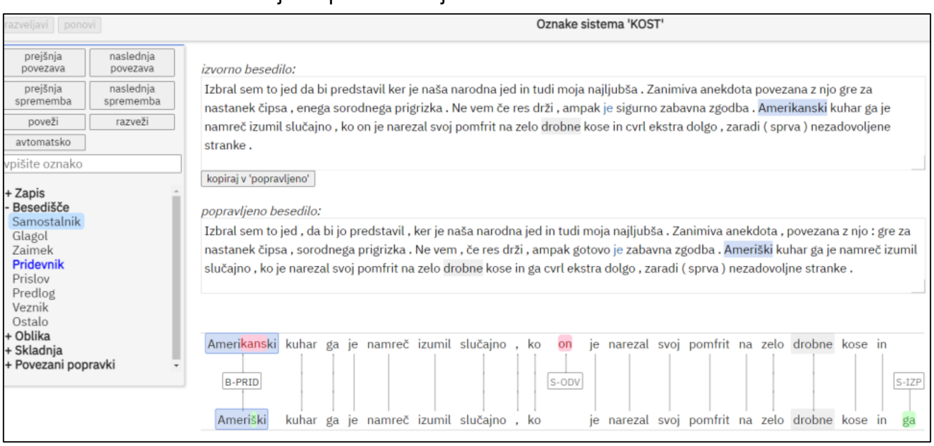
[15] Trenutno v različici 1.1, prim. https://orodja.cjvt.si/svala/.

[16] Poleg KOST-a je bila zaenkrat uporabljena pri označevanju korpusa Šolar, uporablja pa se tudi pri nastajajočem korpusu revidiranih besedil z območja jezikovnega stika med slovenščino in italijanščino STIKit.

daljše besedilo, v katerem tvorec ni naredil odstavkov, ga zaradi preglednejšega prikaza v konkordančniku segmentira na odstavke po smislu.

- Vsakega od odstavkov označevalec posebej vnese v aplikacijo Svala. V zgornje okence (»izvorno besedilo«) skopira izvorno besedilo, v spodnje okence (»popravljeno besedilo«) pa vnese popravljeno verzijo besedila.
- Kjer je mogoče, orodje Svala avtomatsko poveže izvirno in popravljeno besedilo. Povezave med obema, t. i. špageti, so vidne v najnižjem okencu (Slika 2). Povezave je mogoče ročno spreminjati, dodajati, združevati ali razdruževati, označevalec pa vsaki povezavi določi eno ali več oznak napak iz izbirnika na levi.
- Označeno besedilo oz. posamezni odstavek označevalec izvozi v formatu *json*, pri čemer ohrani izvorno ime datoteke, in ga pošlje uredniku korpusa.

Slika 2: Primer označevanja napak v orodju Svala.



Kot je bilo že omenjeno, označevalec v Svali vidi celotni odstavek, pri določanju tipov napak pa celo samo eno poved. To je tudi eden od razlogov, zakaj v KOST ne vključujemo besedil, v katerih tvorec ne postavlja ločil. Zelo dolg odstavek lahko za Svalo predstavlja preveliko obremenitev. Tak je primer naslednjega kratkega besedila z le dvema končnima ločiloma, ki ni vključeno v označeni KOST:

> Včeraj smo imeli naš naslednj profesorica slovenčina svoj ime je [XImeX], ona prihaja iz [XKrajX] ima 32 let v svoj prostem času druži se z svoj pes, svoj najljbši potovanje je bilo na Praga, pošluša glasbo Rock ona tudi priporočela Batista Cadillac jaz osebno sem ga šlišala in mi je bilo všeč (dobro okus za glasba ima). Svenina nimara gavori slovensko, hrvasko in malo španščina in svoj najljbši film je bilo Wolf Street Film.

5 ZAKLJUČEK

Korpus slovenščine kot tujega jezika KOST 2.0 je sodoben jezikovni vir, ki je s svojimi parametri primerljiv z obstoječimi korpusi za druge svetovne jezike (Stritar Kučuk, 2022), zaradi bogatih metaoznak in tudi zaradi novorazvitega, uporabniku prijaznega konkordančnika pa je na voljo za širok spekter jezikoslovnih analiz in aplikacij. Ob tem se morajo uporabniki podatkov zavedati, da označevanje jezikovnih napak v besedilih neprvih govorcev slovenščine odpira številna jezikoslovna vprašanja – jezikovnotehnološka so v tem prispevku seveda puščena ob strani. Odgovori na ta vprašanja niso samoumevni ali enoznačni. Sama se z njimi srečujem že od snovanja poskusnega korpusa slovenščine kot tujega jezika (prim. Stritar, 2012) naprej. Ali gre pri rabi neustreznega glagolskega vida, npr. glagola *počakati* nam. *čakati* v primeru *Nisem mogla veliko s njim in sem ga veliko počakala*, za napako oblike ali besedišča? Izbrati bi bilo mogoče oboje, a ker je nujno, da so oznake konsistentne, se je treba odločiti za eno možnost, v tem primeru je bilo izbrano besedišče. Tovrstne dileme se ob označevanju napak pojavljajo tako rekoč na dnevni ravni. O njihovih rešitvah je nemalokrat mogoče dvomiti, a celo v primerih, ko niso najboljše, je treba v zvezi z njimi vztrajati (prim. Lüdeling, Hirschmann, 2015, str. 144). Le konsistentno označevanje omogoča razmeroma enostavno izvedljive spremembe in izboljšave korpusnega gradiva v prihodnosti. Vsakič znova pa se tudi potrjuje, da je označevanje jezikovnih napak kompleksno delo, ki zahteva označevalca z izkušnjami s področja slovenščine kot neprvega jezika in s sposobnostjo sprejemanja pragmatičnih rešitev.

LITERATURA

Arhar Holdt, Š., Kosem, I., Stritar Kučuk, M. (2022). Metode in orodja za lažjo pripravo korpusov usvajanja jezika. V N. Pirih Svetina, I. Ferbežar (ur.), *Na stičišču svetov* (str. 23–30). Založba Univerze. https://doi.org/10.4312/Obdobja.41.23-30

Gilquin, G., (2015). From design to collection of learner corpora. V S. Granger, G. Gilquin, F. Meunier (ur.), *The Cambridge Handbook of Learner Corpus Research* (str. 9–34). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy *CALICO Journal* 20(3), 465–479.

Granger, S., Swallow, H., Thewissen, J. (2022). *The Louvain Error Tagging Manual Version 2.0*. Centre for English Corpus Linguistics, Université catholique de Louvain. Pridobljeno januarja 2024, https://cdn.uclouvain.be/groups/cms-editors-cecl/Granger%20et%20al._Error%20tagging%20manual_v2.0_2022.pdf

Hammarberg, B. (2010). *Introduction to the ASU Corpus: A longitudinal oral and written text corpus of adult learner Swedish with a corresponding part from native Swedes.* http://www.diva-portal.org/smash/get/diva2:778204/FULLTEXT01.pdf

Klinar, M., Pisek, S., Stritar Kučuk, M., Šter, H. (2022). Poučevanje slovenščine za redno vpisane tuje študente na Univerzi v Ljubljani. V N. Pirih Svetina, I. Ferbežar (ur.), *Na stičišču svetov* (str. 185–194). Založba Univerze v Ljubljani. https://doi.org/10.4312/Obdobja.41.185-194

Lüdeling, A., Hirschmann, H. (2015). Error annotation systems. V S. Granger, G. Gilquin, F. Meunier (ur.), *The Cambridge Handbook of Learner Corpus Research* (str. 133–157). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Mikelić Preradović, N. (2020). Označavanje pogrešaka u CroLTeC-u (računalnom učeničkom korpusu hrvatskog kao stranog jezika). *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje* 46(2), 899–920.

Razpotnik, B. (2024). Vsak sedmi prebivalec Slovenije se je v Slovenijo priselil. Statistični urad Republike Slovenije. https://www.stat.si/StatWeb/news/Index/9999

Rosen, A., Hana, J., Hladká, B., Jelínek, T., Škodová, S., Štindlová, B. (2020). *Compiling and annotating a learner corpus for a morphologically rich language: CzeSL, a corpus of non-native Czech*. Praga: Karlova univerza.

Stritar, M. (2012). *Korpusi usvajanja tujega jezika*. ZDSD.

Stritar Kučuk, M. (2022). KOST med korpusi usvajanja tujega jezika. V N. Pirih Svetina, I. Ferbežar (ur.), *Na stičišču svetov* (str. 323–334). Založba Univerze v Ljubljani. https://doi.org/10.4312/Obdobja.41.323-334

Stritar Kučuk, M. (2023a). *KOST, Korpus slovenščine kot tujega jezika: Priročnik za označevanje napak*. https://www.cjvt.si/korpus-kost/wp-content/uploads/sites/24/2022/04/Prirocnik-za-oznacevanje-napak-v-KOST-u-2022-04-13.pdf

Stritar Kučuk, M. (2023b). Error annotation in Slovene learner corpus KOST – why L1 students can(not) do the job. *CLARC 2023: Jezik i jezični podaci*. https://uniri-my.sharepoint.com/:w:/g/personal/bperak_uniri_hr/EdBOkvsg4vJOrVeHTkQw3uYB16acgdyFh2g5S5fpdXqhYA?rtime=XEOZp-9w3Eg

Stritar Kučuk, M. (2023c). Neizbežno spodbudna ovira: vpliv strojnega prevajanja na pisno produkcijo v slovenščini kot drugem jeziku. V M. Leskovec, I. Samide (ur.), *Z jeziki danes za jutri: Aktualni izzivi poučevanja jezikov, literatur in kultur. Zbornik povzetkov prispevkov konference* (str. 66). Založba Univerze v Ljubljani.

Stritar Kučuk, M. (2024). Prvi korpus slovenščine kot tujega jezika KOST 1.0. V Š. Arhar Holdt, S. Krek (ur.), *Razvoj slovenščine v digitalnem okolju* (str. 93–117). Založba Univerze. https://doi.org/10.4312/9789612972561

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C., Sundberg, G., Wirén, M. (2019). The SweLL Language Learner Corpus: From Design to Annotation. *Northern European Journal of Language Technology* 6, 67–104.

Wirén, M., Matsson, A., Rosén, D., Volodina, E., 2018. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. *Selected papers from the CLARIN Annual Conference 2018* (str. 227–239).

TEST POZNAVANJA SPLOŠNIH BESED V SLOVENŠČINI MED UDELEŽENCI MLADINSKE POLETNE ŠOLE SLOVENŠČINE

Matej KLEMEN

Center za slovenščino kot drugi in tuji jezik, Filozofska fakulteta, Univerza v Ljubljani

V prispevku predstavljamo pilotno izvedbo testa poznavanja splošnih besed v slovenščini. S testom smo želeli ugotoviti, v kolikšni meri govorci slovenščine kot drugega in tujega jezika poznajo splošne besede v slovenščini in kakšen je pri testiranih napredek v poznavanju splošnega besedišča po udeležbi na poletni šoli slovenščine. Zanimalo nas je, ali bi bilo tak test mogoče uporabiti kot alternativo časovno potratnejšim uvrstitvenim testom, da bi z njim udeležence tečajev slovenščine kot drugega in tujega jezika razvrstili v skupine, ki so homogene po jezikovnem znanju. Po vzoru testov za angleščino in grščino kot neprva jezika smo pripravili t. i. da/ne test poznavanja najfrekventnejših 4000 splošnih besed v slovenščini, ki smo ga v dveh ponovitvah izvedli med udeleženci 17. Mladinske poletne šole slovenščine.

Ključne besede: besedišče, test, slovenščina kot drugi in tuji jezik, študentski prispevek

1 UVOD[1]

Pri organizaciji tečajev slovenščine se srečujemo z izzivom, kako udeležence hitro in učinkovito razporediti v skupine, ki so po jezikovnem znanju čim bolj homogene. Tradicionalno jih na Centru za slovenščino kot drugi in tuji jezik Filozofske fakultete Univerze v Ljubljani razvrščamo na podlagi uvrstitvenega testa in uvrstitvenega intervjuja, s katerima ocenimo sporazumevalno jezikovno zmožnost udeležencev tečajev. Ob velikem številu udeležencev, zlasti ob začetku večjih prireditev (npr. poletnih šol, ki se jih udeležuje po več kot 100 tečajnikov), tak način uvrščanja v skupine pomeni veliko dela in terja veliko časa.

Kot odgovor na podobne težave je v osemdesetih letih 20. stoletja Meara s sodelavci za angleščino razvil Eurocentres Vocabulary Size Test, ki ga je kasneje še dopolnjeval (Meara in Jones, 1988; Meara in Miralpeix, 2016). Test

[1] Prispevek je nastal v sklopu doktorskega študija digitalnega jezikoslovja na Univerzi v Ljubljani.

temelji na predpostavki, da je poznavanje besedišča povezano s splošno jezikovno zmožnostjo. Gre za enostaven in hiter test, pri katerem se testirani za posamezne besede v tujem jeziku odločajo, ali jih poznajo ali ne.[2] Kot eno od prednosti testa Meara in Miralpeix (2016, str. 115) omenjata tudi možnost prilagoditve za različne jezike.

V prispevku predstavljamo pilotno izvedbo testa poznavanja splošnega besedišča v slovenščini in njegovo uporabnost pri uvrščanju udeležencev tečajev slovenščine kot drugega in tujega jezika (SDTJ) v skupine, ki so homogene po jezikovnem znanju.

2 TEST POZNAVANJA SPLOŠNIH BESED V SLOVENŠČINI

2.1 Zgledi za pripravo testa besedišča

Test za poznavanje besed v slovenščini je bil pripravljen po vzoru testa V_YesNo (Meara in Miralpeix, 2016), ki izhaja iz omenjenega Eurocentres Vocabulary Size Test, in testa obsega slovarja, kot sta ga pripravila Milton in Alexiou (2010).

Test V_YesNo preverja poznavanje 10.000 najpogostejših besed v angleščini. Gre za test v digitalni obliki, v katerem so testiranemu predstavljene posamezne besede brez konteksta. Testirani za besedo označi, ali njen pomen pozna ali ne.[3] Ob pravih besedah so v test vključene tudi v angleščini neobstoječe besede (nebesede), kar naj bi preprečevalo goljufanje oz. ugibanje. Test obsega 100 pravih besed (po 10 besed za vsakih 1000) in 100 nebesed. Ob koncu testa 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 za posamezno napačno prepoznavo nebesede 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 (Meara in Miralpeix 2016).

[2] V angleščini se je takih testov prijelo poimenovanje *yes/no test*.

[3] Test je dostopen na: https://www.lognostics.co.uk/tools/V_YesNo/V_YesNo.htm.

Test, s katerim sta Milton in Alexiou (2010) želela oceniti obseg besedišča pri učencih grščine, je podoben testu V_YesNo. Prav tako gre za *da/ne test*, besede pa zajema izmed prvih 5000 najpogostejših lem v grškem nacionalnem korpusu (Hellenic National Corpus). Za vsakih 1000 besed je izbranih po 20 besed, 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« (Milton in Alexiou, 2010, str. 318). 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 (Milton in Alexiou, 2010).

Predstavljena testa sta bila kot model za pripravo testa poznavanja splošnih besed v slovenščini izbrana predvsem, ker omogočata enostavno distribucijo in hitro odgovarjanje. Tako smo lahko novo pripravljeni test izvedli med udeleženci Mladinske poletne šole slovenščine (MPŠ), ne da bi z raziskavo pretirano posegli v pouk.

2.2 Priprava testa poznavanja splošnih besed v slovenščini

Pri pripravi testa poznavanja besed v slovenščini smo sledili principom omenjenih testov. Pripravili smo *da/ne test*, ki izhaja iz Referenčnega seznama pogostih splošnih besed za slovenščino (Pollak in sod., 2020). 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 (Arhar Holdt in sod., 2020). Za pripravo testa je bilo upoštevanih prvih 4000 lem, saj zadnja tisočica lem v seznamu ni popolna. Za vsakih zaporednih 1000 lem je bilo naključno izbranih 20 lem, skupaj torej 80. Ker smo želeli, da je izbor čim bolj enakomeren, je bila izmed vsakih zaporednih 50 lem naključno izbrana ena. 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 ranga. V nasprotju z nekaterimi testi besedišča, ki vključujejo samo samostalnike, glagole in pridevnike (gl. Kremmel in Schmitt, 2017, str. 1–2), so bile kot pri Meari (1992) vključene tudi druge besedne vrste (npr. veznik *kajti*, prislovi *tako*, *večinoma* in *podobno*, ki so v izboru za prvih 1000 najpogostejših splošnih besed). Poleg izbranih besed smo v test vključili tudi v slovenščini neobstoječe besede, ki smo si jih izmislili in na videz delujejo kot slovenske (npr. *posminati*, *čembrita*, *deptanjski*). Razmerje med besedami in nebesedami je bilo kot v testu Miltona in Alexiou (2010) 5 : 1. Z izbranimi 80 obstoječimi besedami in 16 nebesedami test tako vsebuje 96 vprašanj.

Zaradi enostavnejše izvedbe testiranja, zbiranja in analize odgovorov smo se odločili za digitalno obliko testa. Test poznavanja besed v slovenščini smo naložili v didaktično orodje za izvedbo različnih kvizov Socrative.[4] Pri vsakem vprašanju se je testiranemu prikazalo enako navodilo (*Če veste, kaj beseda pomeni, izberite TRUE. Če ne veste, kaj beseda pomeni, izberite FALSE. Če niste prepričani, izberite FALSE.*), ob njem pa beseda in gumbi *true*, *false* in *submit answer* (Slika 1).[5] Besede so se vsakemu testiranemu prikazovale v naključnem vrstnem redu, kar je pomenilo, da so bile bolj in manj frekventne besede med seboj pomešane, kot za enega od testov besedišča priporoča Nation (2012).

Pred izvedbo je bil test preizkušen na 5 govorcih slovenščine kot prvega jezika, ki so vseh 16 besed ustrezno prepoznali kot nebesede. Dve verziji testa z različnimi 80 besedami, na enak način izbranimi iz Referenčnega seznam pogostih splošnih besed, in 16 nebesedami sta bili preizkušeni na skupini 5 govorcev SDTJ, ki so na obeh testiranjih dosegli podoben rezultat.

[4] https://www.socrative.com

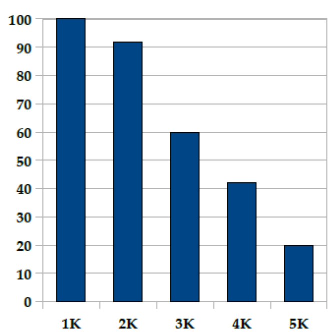
[5] Ker testa nismo izvajali pri popolnih začetnikih v slovenščini (gl. razdelek 4), je bilo navodilo samo v slovenščini, saj smo predvidevali, da ga bodo testirani razumeli. Zaradi prednastavljenega formata in jezika kviza gumbov v angleščini nismo mogli preimenovati. Prav tako ni bilo mogoče dodati dodatnega gumba *Ne vem.* ali *Nisem prepričan/a.*, s katerim bi bilo mogoče znižati stopnjo ugibanja pri odgovarjanju (Zhang, 2013).

Slika 1: Prikaz vprašanja v orodju Socrative.

3 HIPOTEZE

S testom poznavanja najpogostejših 4000 splošnih besed v slovenščini smo želeli preveriti, ali tako kot za govorce drugih tujih jezikov tudi za govorce SDTJ velja, da v procesu učenja prej in v večji meri poznajo bolj frekventne besede kot manj frekventne. Predvidevamo, da bomo za udeležence MPŠ, med katerimi smo test izvedli, prišli do padajočega profila poznavanja besedišča, kot ga predstavlja Meara (1992; Slika 2) (hipoteza 1).

Slika 2: Tipični profil poznavanja besedišča (vir: Meara, 1992, str. 6).



Želeli smo preveriti, ali bi test poznavanja najpogostejših 4000 splošnih besed v slovenščini lahko uporabili kot alternativo uvrstitvenim testom in/ali uvrstitvenim intervjujem. Predvidevamo, da se bo pokazala razlika v skupnem številu točk med posameznimi skupinami udeležencev MPŠ glede na jezikovno znanje, in sicer da bodo začetniki dosegli nižje rezultate kot nadaljevalci in izpopolnjevalci (hipoteza 2). Predvidevamo, da bi na podlagi rezultatov testa poznavanja besedišča lahko večji del učencev razvrstili v skupine, ki bi bile sorazmerno homogene po jezikovnem znanju, primerljivo ustrezno kot na podlagi rezultatov uvrstitvenega testa (hipoteza 2.1).

Z dvema ponovitvama testa smo želeli poleg tega ugotoviti, kakšen je napredek v poznavanju splošnega besedišča po 40 urah pouka in dveh tednih življenja v slovenskem okolju. Predvidevamo, da testirani pri drugem testiranju dosežejo pomembno višji rezultat (hipoteza 3) in da bodo spremembe večje predvsem pri testiranih, ki so pri prvem testiranju dosegli nižji rezultat (hipoteza 3.1).

4 TESTIRANJE IN PODATKI

Za sodelovanje smo prosili udeležence Mladinske poletne šole slovenščine, ki jo obiskujejo mladostniki, stari od 13 do 18 let, za katere je slovenščina najpogosteje eden od prvih jezikov (zamejci, izseljenci), dediščinski ali tuji jezik.

Na 17. MPŠ je leta 2022 pouk potekal v devetih skupinah (v nadaljevanju so označene s številkami, npr. 4, 7, 9). Udeleženci MPŠ so bili vanje razdeljeni na podlagi rezultatov uvrstitvenega testa, deloma pa tudi na podlagi drugih dejavnikov (npr. prvega jezika). Skupine so bile po stopnji jezikovnega znanja torej različne, sosednje skupine (npr. skupini 8 in 9) pa se v jezikovnem znanju med seboj niso vedno nujno močno razlikovale.

Testiranje je bilo izvedeno dvakrat: 5. julija 2022 in 15. julija 2022, prvi in zadnji dan pouka na MPŠ. Učitelji so udeležencem MPŠ dali navodila, udeleženci pa so test izpolnjevali na svojih pametnih telefonih. Podatkov o času reševanja program ni beležil, ocenjujemo pa, da so testirani test zaključili v približno 5 do 12 minutah. Ker je bilo za test V_YesNo ugotovljeno, da ni zanesljiv za začetnike (Meara in Miralpeix, 2016, str. 118), smo se odločili, da popolnih začetnikov v slovenščini v testiranje ne bomo vključili, in tako na MPŠ 2022 test izvedli le v skupinah od 3 do 9. Iz različnih razlogov udeleženci v nekaterih skupinah testa niso reševali v obeh testiranjih.

V prvem testiranju (T1) je sodelovalo 56 udeležencev MPŠ: 3 od teh testa niso dokončali, za 5 pa nismo dobili podatka, v katero skupino so bili razporejeni na podlagi uvrstitvenih testov. V drugem testiranju (T2) je sodelovalo 54 udeležencev MPŠ: 2 testa nista izpolnila do konca, za 8 pa nismo dobili podatka o skupini. V analizo posameznega testiranja so bili vključeni samo odgovori oseb, ki so na test odgovorile v celoti in za katere smo imeli podatek o skupini (za T1 je bilo takih 48, za T2 pa 44), za primerjalno analizo pa so bili upoštevani samo odgovori oseb, ki so sodelovale v obeh testiranjih in bile v obeh v isti skupini (takih je bilo 25).[6]

Poleg rezultatov testiranja smo za testirane, ki so sodelovali v T1, pridobili število točk, ki so jih dosegli na uvrstitvenem testu pri nalogah za preverjanje bralnega razumevanja in oblikoskladenjske zmožnosti. Za večino smo dobili tudi podatek, kako je bilo njihovo jezikovno znanje ocenjeno na podlagi omenjenih nalog uvrstitvenega testa in dodatne naloge za pisanje z oznakami Z, Z+, Z/N, N–, N, N+, N+/I, I-, I, ki pomenijo (boljši ali slabši) začetnik, nadaljevalec in izpopolnjevalec.[7] Ker je naš vzorec majhen in je posamezno oznako lahko dobil tudi le en testirani, smo oznake razdelili v tri glavne: Z (začetnik), N (nadaljevalec) in I (izpopolnjevalec).[8]

5 REZULTATI

V podatkih je bilo pri obstoječih besedah vsakemu odgovoru, pri katerem je testirani označil, da besedo pozna (*true*), pripisanih 50 točk, vsak napačen odgovor pa je dobil 0 točk. Posebej je bilo točkovanih tudi 16 nebesed. Vsaki, ki je bila prepoznana kot slovenska beseda (testirani je označil *true*), je bilo pripisanih –250 točk. Vsak »pravilen« odgovor pa je dobil 0 točk. Izračunano je bilo skupno število točk za vsakih tisoč besed, skupno število točk za prave besede, skupno število točk za neprave besede in skupno korigirano število točk (od točk za obstoječe besede so bile odštete točke za nebesede, ki so bile prepoznane kot obstoječe besede).[9]

[6] Nekateri udeleženci MPŠ so bili lahko po prvih dneh glede na svoje želje in/ali jezikovne zmožnosti prerazporejeni v druge skupine.

[7] Te oznake niso usklajene s stopnjami jezikovnega znanja po *Skupnem evropskem jezikovnem okviru* (SEJO).

[8] Oznaka Z/N je bila prevedena v Z, N+/I pa v N.

[9] Pri analizi rezultatov se ukvarjamo s točkami, ki so jih testirani dosegli pri odgovorih za besede, (opomba se nadaljuje na naslednji strani)

Pri T1 in T2 so testirani v povprečju v večji meri poznali frekventnejše besede in za vsakih naslednjih tisoč dosegli nekoliko nižji rezultat (Tabela 1). V obeh testiranjih se je pokazal padajoči profil poznavanja besedišča (Graf 1 in 2),[10] ki se začne stopničasto spuščati šele pri poznavanju tretjih in četrtih tisoč besed. To nam potrdi tudi analiza variance (ANOVA) za odvisne vzorce, ki je pokazala, da se pri T1 in T2 rezultati za posameznih tisoč besed med seboj statistično pomembno razlikujejo (ob upoštevanju Greenhouse-Geissserjevega popravka T1: $F(2{,}48) = 29{,}9$, $p < 0{,}001$; T2: $F(2{,}37) = 31{,}9$, $p < 0{,}001$). Če s post hoc testom za T1 in T2 primerjamo povprečne dosežke za tisoč besed in upoštevamo Bonferronijev popravek, edino razlika med T1 1–1000 in T1 1001–2000 ($p_{\text{bonferroni}} = 1{,}0$) ter T2 1–1000 in T2 1001–2000 ($p_{\text{bonferroni}} = 0{,}8$) ni statistično značilna, vse druge razlike pa so statistično značilne (Tabela 2).

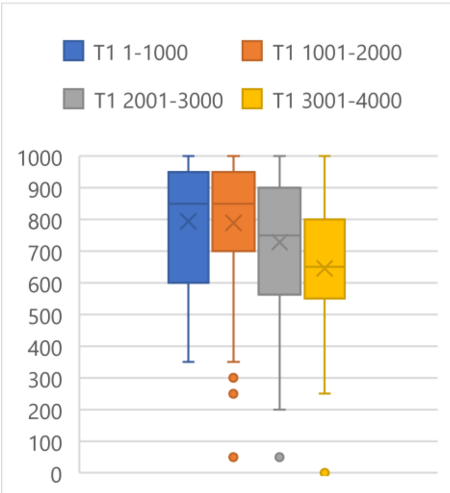
Tabela 1: Število točk, doseženih pri T1 in T2 za posameznih tisoč besed.

	N	*M*	*Mdn*	*SD*	*Minimum*	*Maksimum*
T1 1–1000	48	794	850	189	350	1000
T1 1001–2000	48	789	850	221	50	1000
T1 2001–3000	48	728	750	234	50	1000
T1 3001–4000	48	646	650	206	0	1000
T2 1–1000	44	842	875	166	400	1000
T2 1001–2000	44	816	900	225	250	1000
T2 2001–3000	44	738	800	248	100	1000
T2 3001–4000	44	684	700	252	50	1000

odgovore za nebesede večinoma puščamo ob strani. Kako pri rezultatu testa upoštevati nebesede, je med raziskovalci pogosto predmet razprave (Durrant in sod., 2022, str. 178). V predstavljeni test smo jih vključili, ker smo sledili formatu testa za angleščino oz. grščino in z namenom, da bi testirane odvrnile od ugibanja. Zdi se, da so jih testirani prepoznali kot take. Pri T1 in T2 bi odbitek točk za nebesede verjetno le malo vplival na povprečje (T1: N = 48, *M* = –313, *Mdn* = –125; T2: N = 44, *M* = –455, *Mdn* = –125).

[10] Ker sta se skupini, ki sta sodelovali v T1 in T2, deloma razlikovali, rezultate prikazujemo v ločenih grafih. Rezultate za 25 testiranih, ki so sodelovali tako pri T1 kot pri T2, primerjalno prikazujemo kasneje.

Graf 1: Število doseženih točk za posameznih tisoč besed pri T1.



Graf 2: Število doseženih točk za posameznih tisoč besed pri T2.

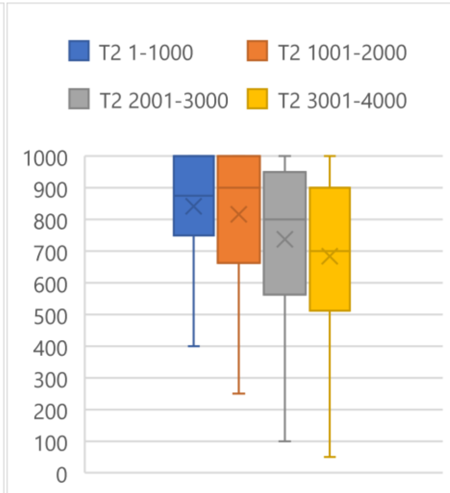


Tabela 2: Primerjava povprečnih dosežkov za posameznih tisoč besed.

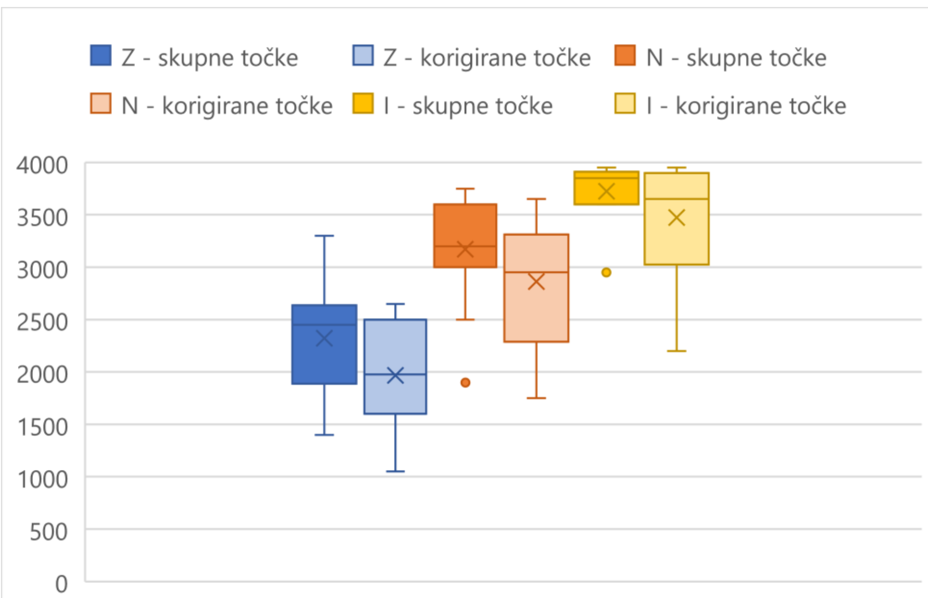
primerjava			*razlika v povprečju*	*SE*	*df*	*t*	$p_{bonferroni}$
T1 1–1000	-	T1 1001–2000	5,21	19,9	47,0	0,261	1,000
	-	T1 2001–3000	65,63	20,8	47,0	3,155	0,017
	-	T1 3001–4000	147,92	20,6	47,0	7,190	< 0,001
T1 1001–2000	-	T1 2001–3000	60,42	13,6	47,0	4,456	< 0,001
	-	T1 3001–4000	142,71	15,2	47,0	9,400	< 0,001
T1 2001–3000	-	T1 3001–4000	82,29	15,4	47,0	5,357	< 0,001
T2 1–1000	-	T2 1001–2000	26,1	17,1	43,0	1,53	0,801
	-	T2 2001–3000	104,5	21,6	43,0	4,84	< 0,001
	-	T2 3001–4000	158,0	21,7	43,0	7,26	< 0,001
T2 1001–2000	-	T2 2001–3000	78,4	15,0	43,0	5,22	< 0,001
	-	T2 3001–4000	131,8	17,4	43,0	7,56	< 0,001
T2 2001–3000	-	T2 3001–4000	53,4	14,5	43,0	3,69	0,004

Kot kažeta Tabela 3 in Graf 3 se glede na stopnjo jezikovnega znanja pri doseženem številu točk na T1 kaže razlika med testiranimi, ki so na različnih stopnjah jezikovnega znanja. Analiza variance (ANOVA) za neodvisne vzorce je pokazala, da se rezultati testa poznavanja splošnih besed v slovenščini, ki so jih testirani na različnih stopnjah jezikovnega znanja (Z, N, I) dosegli, med seboj statistično pomembno razlikujejo (T1 skupne točke: $F(2, 41) = 28{,}8$, $p < 0{,}001$; T1 korigirane točke: $F(2, 41) = 20{,}5$, $p < 0{,}001$).

Tabela 3: Število doseženih točk pri T1 za testirane glede na stopnjo jezikovnega znanja.

	poenostavljena oznaka stopnje	*N*	*M*	*Mdn*	*SD*	*Minimum*	*Maksimum*
T1	Z	12	2321	2450	549	1400	3300
skupne	N	22	3170	3200	449	1900	3750
točke	I	10	3725	3850	304	2950	3950
T1	Z	12	1967	1975	539	1050	2650
korigirane	N	22	2864	2950	566	1750	3650
točke	I	10	3475	3650	571	2200	3950

Graf 3: Skupno število točk in korigirane točke pri T1 glede na stopnjo jezikovnega znanja.



Izračunali smo povezanost med tem, v katero od devetih skupin na MPŠ so bili testirani uvrščeni, in rezultatom, ki so ga dosegli pri T1. Ugotovili smo, da je povezanost visoka in statistično značilna (T1 skupne točke: r = 0,737, df = 46, p = < 0,001; T1 korigirane točke: r = 0,732, df = 46, p = < 0,001). Ta povezanost je le nekoliko manjša, kot je povezanost med številom točk na uvrstitvenem testu in razvrstitvijo v skupino (r = 0,842, df = 46, p = < 0,001).

Če opazujemo odgovore 25 udeležencev MPŠ, ki so sodelovali v obeh testiranjih, vidimo, da so pri drugem testiranju nekoliko bolje odgovarjali glede poznavanja v slovenščini obstoječih besed (*M* = 3108 : 3296, *Mdn* = 3450 : 3750), pri prepoznavanju nebesed pa so v povprečju dosegli malenkost slabši rezultat (*M* = –320 : –440, *Mdn* = 0 : –250) (Tabela 4). S t-testom za odvisne vzorce smo pokazali, da za te testirane obstaja statistično pomembna razlika med T1 in T2 pri skupnih točkah ($t(24) = 2{,}744$, $p = 0{,}011$) in rezultatih za besede 1–1000 ($t(24) = 2{,}113$ $p = 0{,}045$), 1001–2000 ($t(24) = 2{,}511$, $p = 0{,}019$), 3001–4000 ($t(24) = 2{,}343$, $p = 0{,}028$), za besede iz frekvenčnega ranga 2001–3000 pa razlika ni bila statistično značilna ($t(24) = 1{,}67$, $p = 0{,}108$). Prav tako razlika ni bila statistično značilna za korigirane točke ($t(24) = 0{,}583$, $p = 0{,}565$).

Tabela 4: Rezultati testiranih, ki so sodelovali pri T1 in T2.

	T1 skupaj točk	*T1 nebesede (odbitek)*	*T1 korigirane točke*	*T2 skupaj točk*	*T2 nebesede (odbitek)*	*T2 korigirane točke*
N	25	25	25	25	25	25
M	3108	–320	2788	3296	–440	2856
Mdn	3450	0	3050	3750	–250	3050
SD	883	547	998	828	634	917
Minimum	450	–2250	450	800	–2500	800
Maksimum	3950	0	3950	4000	0	3850

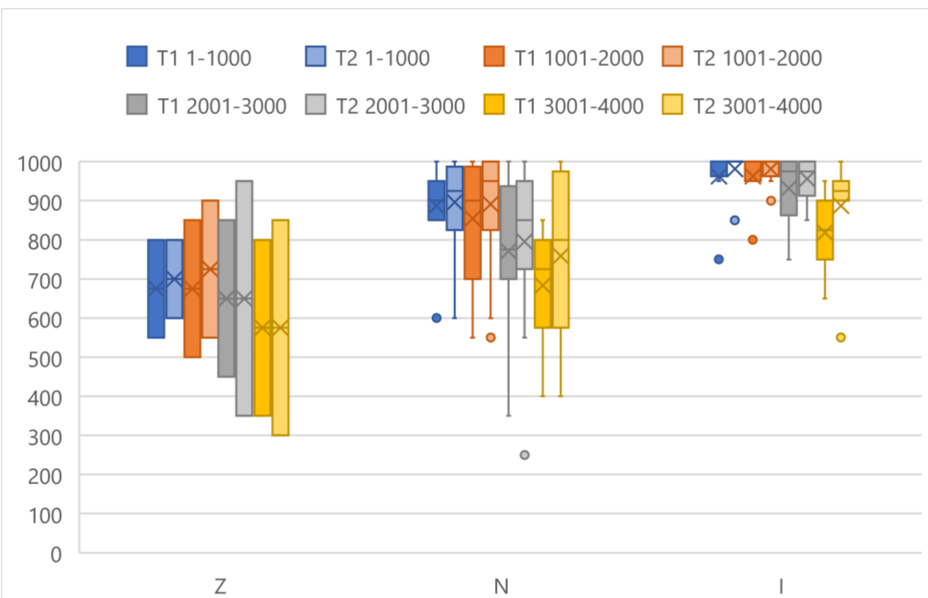
Ko smo pri testiranih, ki so sodelovali tako pri T1 kot pri T2, opazovali rezultate glede na njihovo splošno jezikovno znanje (Z, N, I), smo za posameznih tisoč besed opazili, da so rezultati višji ali podobni (Tabela 5, Graf 4). Za skupini N in I smo izvedli t-test za odvisne vzorce.[11] Niti za korigirane točke niti za posameznih tisoč besed nismo ugotovili, da bi bile vrednosti statistično značilno drugačne (*p* je povsod > 0,05). Izjema je bilo skupno število točk pri skupini I, ki se je statistično značilno spremenilo ($p = 0{,}013$).

[11] Za skupino Z t-testa nismo izvedli, saj sta test reševali le 2 osebi, katerih jezikovno znanje je bilo ob T1 označeno kot začetno.

Tabela 5: Rezultati za testirane, ki so sodelovali tako pri T1 kot pri T2, glede na splošno jezikovno znanje.

	poenostavljena oznaka stopnje	*T1 skupne točke*	*T2 skupne točke*	*T1 korigirane točke*	*T2 korigirane točke*
N	Z	2	2	2	2
	N	12	12	12	12
	I	8	8	8	8
M	Z	2575	2650	1450	1900
	N	3196	3342	2904	2904
	I	3675	3806	3581	3556
SD	Z	1025	1202	566	141
	N	539	672	621	741
	I	323	272	384	309

Graf 4: Število doseženih točk za posameznih tisoč besed za testirane, ki so sodelovali tako pri T1 kot pri T2, glede na splošno jezikovno znanje.



6 DISKUSIJA

V obeh testiranjih se je potrdilo, da v splošnem govorci SDTJ prej v procesu učenja in v večji meri poznajo frekventnejše besede. Pri udeležencih MPŠ, ki smo jih opazovali, z uporabljenim testom nismo ugotovili večje razlike v poznavanju prvih in drugih tisoč besed, za nadaljnjih dvakrat tisoč besed pa se je pokazal padajoči profil. Tako se je hipoteza 1 le delno potrdila.

Podobni profili za posameznih tisoč besed pri T1 in T2 so se pokazali za vse testirane skupaj (Grafa 1 in 2), prav tako pa tudi za skupine z različnim jezikovnim znanjem – Z, N in I (Graf 4), ki pa so se po pričakovanjih med seboj razlikovale v doseženem rezultatu: začetniki so dosegli nižji rezultat kot nadaljevalci in ti nižjega kot izpopolnjevalci (hipoteza 2). Ob primerjavi rezultatov testa poznavanja besedišča pri T1 in rezultatov, doseženih na uvrstitvenem testu, se je pokazalo, da je tudi v naši raziskavi korelacija med številom točk in rezultatom, doseženim pri T1, visoka (r = 0,78 za skupne točke; r = 0,732 za korigirane točke), kar je skladno z navedbami Meare in Miralpeix (2016, str. 129) za druge podobne raziskave. To pomeni, da bi lahko le na podlagi testa poznavanja najpogostejših 4000 splošnih besed v slovenščini podobno ustrezno razvrstili udeležence MPŠ v skupine, ki bi bile po jezikovnem znanju homogene (hipoteza 2.1).

Če bi za razvrščanje tečajnikov uporabili predstavljeni test poznavanja splošnih besed, bi ga bilo za natančnejše razporejanje v skupine verjetno smiselno dopolniti z nalogo za samostojno produkcijo, ki bi dala vpogled v posameznikovo jezikovno zmožnost bolj celostno (in s tem učitelju omogočila načrtovanje pouka, ki bo po meri učencev), tako kot so bile naloge za preverjanje bralnega razumevanja in oblikoskladenjske zmožnosti v uvrstitvenem testu za MPŠ 2022 dopolnjene z nalogo za pisanje daljšega besedila. Iz predstavljenega testa poznavanja besedišča namreč dobimo informacijo samo o omejenem delu jezikovne zmožnosti (prim. Read, 2000, str. 3–7).

Rezultati testiranj kažejo, da so testirani na različnih stopnjah jezikovnega znanja pri T2 poznali več besed, vendar razlike z izjemo skupnega števila točk pri skupini I niso bile statistično značilne. Tako hipotezo 3 lahko potrdimo le deloma. Zdi se, da je 40 ur pouka premalo, da bi lahko govorili o velikem napredku pri poznavanju besedišča. Če pogledamo razlike med povprečji za skupne točke za skupine Z, N in I pa se nam hipoteza, da bo do največjega napredka prišlo pri testiranih z nižjim rezultatom na T1 (hipoteza 3.1), ni potrdila: večji napredek se kaže pri testiranih, ki so v slovenščini nadaljevalci in izpopolnjevalci (razlike med povprečji skupnih točk za T1 in T2: Z – 75 točk, N – 146 točk in I – 131 točk). Verjetno bi bila slika drugačna, če bi napredek v poznavanju besedišča merili pri popolnih začetnikih.

7 SKLEP

Predstavili smo test poznavanja splošnega besedišča, ki poznavanje besed preverja diskretno, izbirno in neodvisno od konteksta (gl. Read, 2000, str. 7–13). S testom smo med udeleženci 17. Mladinske poletne šole slovenščine preverjali, koliko besed od 4000 najpogostejših splošnih besed v slovenščini poznajo. Rezultati kažejo, da testirani v večji meri poznajo bolj frekventne besede kot manj frekventne: po drugih tisoč besedah se pokaže tipičen stopničasto spuščajoč se profil. Pilotna izvedba testa je pokazala, da bi ga lahko na MPŠ uporabili kot alternativo uvrstitvenemu testu za razvrščanje tečajnikov v skupine, ki bi bile homogene po jezikovnem znanju. Ali bi ga lahko uporabili tudi na drugih tečajih SDTJ, bi bilo treba še preveriti.

Pri interpretaciji rezultatov glede obsega in globine poznavanja besedišča moramo biti previdni. Test preverja le najosnovnejše poznavanje besede, in sicer ali določeno kombinacijo črk testirani prepozna kot besedo (prim. Meara, 1994, str. 6). O tem, v kolikšni meri besedo receptivno ali produktivno zna (na ravni oblike, pomena in rabe, prim. Nation, 2022, str. 54), tako ne moremo sklepati. Na podlagi rezultatov izvedenega testa tudi ne moremo z gotovostjo trditi, da testirani – tudi zaradi možnosti ugibanja in subjektivnosti pri odločanju, kaj testiranim pomeni, da *vejo*, kaj beseda pomeni – dejansko poznajo tolikšen del testiranih besed, kot jih pokažejo dosežene točke, t. i. *da/ne testi* namreč v večini primerov precenjujejo poznavanje besedišča pri testiranih (Durrant in sod., 2022, str. 178), na kar sta opozorila že Meara in Jones (1988). Opozoriti velja tudi, da so nekateri testirani kot govorci slovenščini sorodnih jezikov lahko zaradi podobnosti besed (napačno) sklepali, da poznajo pomen besed v slovenščini.[12] Uporabljeni test zajema v primerjavi z drugimi testi besedišča, ki segajo tudi do 10.000 najfrekventnejših besed ali dlje, iz sorazmerno majhnega seznama 4000 lem. Domnevamo, da so testirani poznali tudi manj frekventne besede, ki se niso uvrstile na

[12] Na podoben fenomen opozarja Meara (1992, str. 11) pri govorcih romanskih in germanskih jezikov, ki si z angleščino delijo del besedišča.

Referenčni seznam pogostih splošnih besed za slovenščino, in tudi zato bi o obsegu slovarja testiranih lahko postavili le zelo približne ocene.[13]

Če bi želeli dobiti natančnejši vpogled v obseg besedišča za posamezne stopnje jezikovnega znanja po SEJO, bi morali pripraviti test drugačnega formata in vanj zajeti tudi manj frekventne besede. Za večjo zanesljivost testa pa bi morali vključiti tudi večje število besed.[14]

ZAHVALA

Zahvaljujem se vodjema Mladinske poletne šole slovenščine dr. Mihaeli Knez in dr. Damjani Kern Andoljšek ter vsem učiteljicam in učencem, ki so mi omogočili izvedbo testa. Za nasvete pri obdelavi rezultatov se zahvaljujem prof. dr. Anji Podlesek. Za pomoč se zahvaljujem mentorici znan. sod. dr. Špeli Arhar Holdt. Zahvaljujem se tudi recenzentom za koristne pripombe.

LITERATURA

Arhar Holdt, Š., Pollak, S., Robnik Šikonja, M. in Krek, S. (2020). Referenčni seznam pogostih splošnih besed za slovenščino. V D. Fišer in T. Erjavec (Ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika* (str. 10–15). Inštitut za novejšo zgodovino. Pridobljeno 12. januarja 2021, https://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Arhar-Holdt-et-al_Referencni-seznam-pogostih-splosnih-besed-za-slovenscino.pdf

Durrant, P., Siyanova-Chanturia, A., Kremmel, B. in Sonbul, S. (2022). *Research Methods in Vocabulary Studies*. John Benjamins Publishing Company.

Kremmel, B. in Schmitt, N. (2017). Vocabulary Levels Test. V *The TESOL Encyclopedia of English Language Teaching*. John Wiley & Sons. https://doi.org/10.1002/9781118784235.eelt0499

Meara, P. ([1]1992, 2010). *EFL vocabulary tests*. University College Swansea, Centre for Applied Language Studies. Pridobljeno 22. maja 2022, https://www.lognostics.co.uk/vlibrary/meara1992z.pdf

[13] Na podlagi rezultatov tega testa lahko zelo grobo ocenimo, da začetniki (ki pa niso popolni začetniki v slovenščini), nadaljevalci in izpolnjevalci, ki smo jih opazovali, v slovenščini poznajo okoli 2000 najpogostejših splošnih besed, okoli 3000 oz. okoli 3500 besed od 4000 najpogostejših splošnih besed v slovenščini.

[14] Po 30 besed oz. 3 % odstotke besed za posameznih tisoč (prim. Durrant in sod., 2022, str. 158).

Meara, P. (1994). The complexities of simple vocabulary tests. V F. G. Brinkman, J. A. van der Schee in M. C. Schouten-van Parreren (Ur.), *Curriculum research: different disciplines and common goals* (str. 15–28). Vrije Universiteit. Pridobljeno 4. maja 2024, https://www.lognostics.co.uk/vlibrary/meara1994a.pdf

Meara, P. in Jones, G. (1988): Vocabulary size as a placement indicator. V P. Grunwell (Ur.): *Applied Linguistics in Society* (str. 80–87). CILT.

Meara, P. in Miralpeix, I. (2016). *Tools for Researching Vocabulary*. Multilingual Matters. https://doi.org/10.21832/9781783096473

Milton, J. in Alexiou, T. (2010). Developing a vocabulary size test in Greek as a foreign language. V A. Psaltou – Joycey in M. Mattheoudakis (Ur.), *Advances in Research on Language Acquisition and Teaching: Selected Papers* (307–318). Greek Applied Linguistics Association. Pridobljeno 22. maja 2022, https://www.enl.auth.gr/gala/14th/Papers/English%20papers/Milton&Alexiou.pdf

Nation, P. (2012). *The Vocabulary Size Test*. Pridobljeno 31. januarja 2024, https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I. S. P. (2022). *Learning Vocabulary in Another Language*. *Third edition*. Cambridge University Press.

Pollak, S., Arhar Holdt, Š., Krek, S. in Robnik-Šikonja, M. (2020). *Reference List of Slovene Frequent Common Words*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1346

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press.

Zhang, X. (2013). The I don't know option in the vocabulary size test. *TESOL Quarterly*, 47(4), 790–811. doi:10.1002/tesq.98

TEST OF KNOWLEDGE OF COMMON WORDS IN SLOVENE AMONG PARTICIPANTS OF THE YOUTH SUMMER SCHOOL OF SLOVENE

This paper presents a pilot test assessing the knowledge of the common general words in Slovene. The aim of the test was to find out to what extent speakers of Slovene as a second and foreign language know general words in Slovene and what progress the test participants made in their knowledge of general vocabulary after attending a summer school of Slovene. We were interested in whether such a test could be used as an alternative to the more time-consuming placement tests, in order to classify participants of Slovene as a second and foreign language courses into groups homogeneous in terms of language proficiency. Following the example of the tests for English and Greek as FL, we developed a yes/no test of the knowledge of the most frequent 4000 common words in Slovene, which was administered in two iterations to the participants of the 17th Youth Summer School of Slovene.

Keywords: vocabulary, test, Slovene as a second and foreign language, student paper

PARLAY: A METHOD FOR CONSTRUCTING A PARAGRAPH-LEVEL NLI DATASET BASED ON MULTI-CATEGORY SCENARIOS

Meta KOKALJ

Student of Digital Linguistics, University of Ljubljana; mk10260@student.uni-lj.si

This paper presents a novel methodology for constructing a diverse and comprehensive paragraph-level dataset tailored for Natural Language Inference tasks. By incorporating five scenario categories – historical events, scientific explanations, everyday situations, news reports, and fictional stories – our approach ensures broad coverage of various reasoning patterns, including transitive reasoning. The methodology employs iterative and Chain-of-Thought prompting to elicit detailed, context-rich outcomes by generating intermediate reasoning steps. Our dataset is notable for adding detailed explanations for each hypothesis, providing a valuable overview of the thought process behind each categorization. To assess the quality of these outputs, we implemented the OQACM metrics, a comprehensive framework for evaluating large language model outputs across fluency, coherence, and inference quality. This approach improves the interpretability of language models and offers detailed insights into their inferential processes, advancing NLI dataset development.

Keywords: natural language inference (NLI), paragraph-level inference, Chain-of-Thought, large language models (LLMs), LLM evaluation metrics, student paper

1 INTRODUCTION

Natural Language Inference (NLI) tasks are crucial benchmarks for evaluating Natural Language Understanding (NLU) models, which are fundamental to various NLP applications such as question answering, translation, and dialogue (Williams et al., 2018). However, most NLI datasets and research have focused on sentence-level inference, overlooking the importance of paragraphs in discourse and text generation. Paragraphs, often described as "units of thought", go beyond the grammatical structure of individual sentences (syntax) (Zadrozny & Jensen, 1991). Analyzing how sentences within a paragraph connect through cohesion (referential links) and coherence (logical flow of ideas) unlocks a deeper understanding of the meaning (Alyousef, 2021; Bublitz, 2011; Siddharthan, 2006). This analysis of sentence relationships is essential for tasks such as discourse, where sentences need to be arranged in a way that creates a clear and cohesive flow of information. It is essential for tasks that involve verifying the factual accuracy of summaries of documents (Yin et al., 2021). Incorporating paragraph-level semantics into NLP models could significantly improve their ability to handle real-world language tasks.

Many crowdsourced NLI datasets often fail to address real-world NLP challenges due to their creation in isolation from specific tasks and inherent annotator biases (Khot et al., 2018; Hu et al., 2020). For example, in MultiNLI, crowd workers sometimes introduced bias using specific strategies, like adding negators for contradictions, which oversimplifies the complexity of human reasoning (Hu et al., 2020). This creates an unrealistically easy task that doesn't capture the depth needed for robust NLP models. Analyses of SNLI reveal that models often rely on simple heuristics like lexical overlap, lacking compositionality in their representations (McCoy et al., 2019). High word overlap usually predicts entailment, while contradictions show minimal overlap or negation. This highlights the need for models that integrate external knowledge for more nuanced NLI tasks (Camburu et al., 2018; Dasgupta et al., 2018).

What is emerging as NLI's most pressing problem is the propensity of models to overfit on datasets. While these models perform well on training data, their generalization to unseen data is limited, a well-known issue in machine learning, and is referred to as the generalization problem (Gubelmann, 2024). An underlying cause might be the focus on deductive reasoning in training data creation. Careful examination of crowd worker instructions for popular NLI datasets reveals an emphasis on logically guaranteed inferences, neglecting inductive reasoning.

To address some of these challenges and generate more nuanced data, we propose a novel paragraph scope scenario-inference pair elicitation method in which each scenario is accompanied by multiple potential inferences. By integrating diverse reasoning modes during prompt creation, this approach aims to train an NLI model that can handle a wider range of reasoning patterns, enhancing its robustness and generalizability.

The paper is structured as follows: Section 2 reviews related work in NLI, with a particular focus on Slovenian research, as the Parlay dataset is designed for Slovenian. Section 3 describes the Parlay dataset and its construction process. Section 4 provides a detailed analysis of the prompting techniques and demonstrates structured prompting and transitive reasoning. Section 5 evaluates the performance of GPT-4o and Gemini models using the Outcome Quality Assessment Composed Metrics (OQACM). Finally, Section 6 outlines future research directions.

2 RELATED WORK

This section reviews influential NLI datasets, focusing on cornerstone contributions and those involving reasoning beyond sentence-level, to provide context for current NLI research and lay the groundwork for subsequent sections. In addition, the development of NLI resources for underrepresented languages, such as Slovenian, is also discussed to highlight ongoing efforts in multilingual NLI.

Pioneering datasets like The Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and the Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018) set benchmarks in NLI with large-scale, balanced data and clear labels. However, over time, their limitations have become apparent.

2.1 Cornerstone and Challenging Datasets

The SNLI dataset, a cornerstone of NLI, provides a framework for modeling diverse logical relationships but is limited by sentence generation from simple static scenes, restricting reasoning diversity, and by relying on affirmative image captions, which limits the assessment of negation (Bowman et al., 2015; Hossain et al., 2020). The MNLI dataset builds on SNLI by incorporating diverse genres and domains (e.g., fiction, telephone conversations), improving robustness and supporting transfer learning (Williams et al., 2018). However, MNLI inherits SNLI's reliance on human-generated sentences and suffers from genre imbalance. Furthermore, Gururangan et al. (2018) raised concerns that SNLI-trained models might simply exploit coincidental patterns rather than achieving genuine language comprehension.

For reasoning that goes beyond sentences, ContractNLI (Koreeda & Manning, 2021) and ConTRoL (Liu et al., 2020) are notable. ContractNLI evaluates hypotheses against entire documents, particularly contracts. The system is provided with a contract and a set of hypotheses, which could be statements about the contract's obligations, rights, or other key points, requiring models to grasp the broader document context. ConTRoL pushes boundaries in the field with reasoning at a passage level presenting challenges, beyond previous benchmarks.

2.2 Advancements in Slovenian NLI Datasets

Significant progress has been made in developing NLI resources for Slovenian, notably with the Slovene SuperGLUE benchmark and the SI-NLI dataset. The Slovene SuperGLUE benchmark (Žagar in Robnik-Šikonja, 2022) adapts the original SuperGLUE tasks to Slovenian, advancing NLU for this morphologically complex language, but it focuses on broad NLU evaluation rather than NLI specifically. The SI-NLI dataset, the first comprehensive NLI resource for Slovenian, addresses a critical gap in multilingual NLI studies by providing high-quality annotations from the Slovene reference corpus ccKres, with linguist students editing hypotheses to fit NLI relations (Klemen et al., 2024). Despite its contributions, the SI-NLI dataset faces challenges related to the manual construction of examples, which can be time-consuming and introduce biases.

The Parlay dataset differs from SI-NLI by using diverse prompting strategies to automatically generate complex examples, reducing the reliance on human-crafted data. It aims to address the challenge of higher-level reasoning beyond sentence-level NLI, allowing for deeper contextual understanding and enhancing the adaptability of the dataset to various languages and tasks. What's more, Parlay's automated methodology seeks to provide a scalable and efficient solution for NLI research, potentially complementing existing resources like SI-NLI by broadening their scope and applicability.

3 PARLAY DATASET FOR NLI

The section introduces the "Parlay" dataset, which is created to support NLI tasks across various contexts. It explains how the dataset is organized, the categories it covers and the methods used to generate scenarios and collect data.

The Parlay dataset promotes generalizability by encompassing five distinct categories for NLI tasks. To ensure further diversity, each category includes scenarios drawn from various contexts within its topic. These scenarios give brief descriptions of events or situations providing background information for the model to effectively perform different inference tasks.

The scenario categories – historical events, scientific explanations, everyday situations, news reports and fictional stories – were carefully chosen to offer a training platform for various types of reasoning. Historical events enable examination of cause and effect relationships, temporal reasoning, counterfactuals ("what if" scenarios) and analogies (Bottou et al., 2013; Pearl, 2009; Tatu & Srikanth, 2008). Scientific explanations require understanding reasoning such as modus ponens/tollens (Brachman & Levesque, 2004). Everyday scenarios provide a fertile environment for practical reasoning. The model may encounter incomplete information that necessitates inductive reasoning based on patterns, experiences, and common sense (Mitchell, 1997). In addition, everyday scenarios require analogical reasoning, where the model finds similarities to apply knowledge from one scenario to make sense of another (Gentner, 1983). This method helps the language model develop reasoning skills that can be useful in unfamiliar situations. News reports naturally require the model to exercise critical reasoning skills (Chen et al., 2023). Fictional stories provide a platform for exploring narrative reasoning, including understanding characters' motivations and emotions (theory of mind) (Kahneman, 2011; Reniers et al., 2011). Furthermore, because stories often involve social interactions and dynamics, the language model can learn to reason about how characters behave in social contexts and predict their actions based on social norms (social reasoning) (Paul & Frank, 2020; Wu et al., 2018).

We opted for an alternative approach to traditional methods that use natural text-based datasets, using prompt-driven scenario generation instead. Inspired by natural texts and ideas as prompt inputs, the scenarios themselves were created based on those prompts. While this approach reduces citation and authorship challenges in natural text datasets, it has limitations, including potential LLM biases similar to those introduced by human annotators in previous NLI datasets. Recognizing this, we aim to guide future research towards solutions that proactively reduce bias in generated datasets.

The dataset, constructed in Slovenian, contributes to the linguistic diversity of NLI resources, offering unique language-specific challenges and opportunities for advancing multilingual NLI models.

3.1 Dataset structure

For our NLP course project, we constructed a dataset of 100 data points, with 20 examples per category. Each category has 10 examples generated using Gemini LLM ((Google DeepMind, 2024) and 10 with GPT-4o (OpenAI, 2024). This dataset, referred to as the "Parlay dataset", serves as a starting point for familiarizing ourselves with the data collection process. However, to achieve robust results, we plan to significantly expand the dataset and develop more sophisticated prompting techniques to enhance data quality and relevance.

Our dataset, structured in JSON format, organizes relationships across five categories, with each scenario uniquely identified and classified into "Entailment", "Contradiction", and "Neutral" labels. An "OQACM Score" is provided for each scenario to evaluate its quality and relevance, supporting the selection process for both research and practical applications (refer to Section 5 for detailed methodology).

For demonstration purposes, we have selected a scenario from the historical category. The full JSON structure for this scenario can be found in our repository at the following link: https://github.com/UL-FRI-NLP-2023-2024/ul-fri-nlp-course-project-parlay/tree/main/Parlay_dataset/Results_Final_dataset/Sample_from_json

Adding explanations directly to scenario-hypothesis pairs in our NLI dataset is a novel contribution, unlike traditional NLI datasets like SNLI and MultiNLI, which don't include explanations. Explanations, when present, are often presented in separate corpora or in studies specifically focused on explainable AI. For example, Camburu et al. (2018) introduced e-SNLI, which is an extension of SNLI with explanations added in a subsequent step rather than integrated from the outset. Similarly, other studies on explainability in NLP, such as DeYoung et al. (2020), treat explanations as an additional layer of information applied to existing datasets rather than integrating them into the core data from the beginning.

Our approach, which integrates explanations from the start, allows deeper analysis of the large language model (LLM) logic and supports "few-shot learning" by providing explicit reasoning patterns. Analyzing the explanations allows us to gain a deeper understanding of the underlying logic LLM used to establish the relationship between paragraphs. By providing explicit reasoning, the dataset supports "few-shot learning" (Parnami et al., 2022), allowing models to learn effectively even from fewer starting examples. The explanations act as a training aid, helping the LLM apply similar reasoning patterns to novel scenarios. Additionally, by analyzing the explanations, we can identify broader reasoning principles that can be applied beyond specific scenarios.

In our initial, anecdotal experiments, we noticed that prompts including explanations seemed to produce better results. While our observations were informal and not systematically tested, it appeared that the explicit reasoning in the explanations provided a clearer guide for the model, potentially leading to more accurate and coherent outputs. It seems that understanding the rationale behind each decision could help refine the model’s learning process and may improve its ability to generalize from limited data.

3.2 Data collection

We developed scenarios by following specific guidelines and refining them iteratively to establish a well-controlled foundation for diverse inference pairs through prompt-driven development. For scenarios related to historical events, we drew inspiration from well-documented breakthrough events, utilizing resources such as the Wikipedia Timeline of World History and the List of Historical Anniversaries (Wikipedia, 2024). When creating everyday life situations, we combined our creativity with generally known proverbs to develop realistic and relatable scenarios. For fictional scenarios, we relied on our imagination to craft unique and engaging narratives. For news reports, we simply provided the language model with a basic idea, and it generated the news story. Lastly, for scientific explanations, we sourced ideas from Wikipedia's List of Geological Phenomena and Timeline of Historic Inventions (Wikipedia, 2024) ensuring accuracy and relevance in the scenarios.

We employed reasoning rule-guided prompt engineering, incorporating rules like modus ponens, modus tollens, causal reasoning, textual entailment, and predictive inference (Liu et al., 2023). Prompts were crafted to integrate factual hints, guiding the model toward specific inferences (entailment, contradiction, neutral). Factual hints, pieces of information embedded in prompts, provided essential scenario knowledge and anchored the model's reasoning. While foundational, these hints alone were insufficient for real-world inferences, which required broader reasoning. To address this, we incorporated heuristics to capture general reasoning patterns, making prompts adaptable to diverse scenarios and inferences. This adaptability was achieved by exploring templates or conditional statements aligned with targeted reasoning rules.

Prompts were created and refined in English, then only the final, suitable examples were translated into Slovenian using the DeepL model (DeepL SE, 2024), followed by manual corrections. This decision stemmed from the fact that languages with fewer resources require more tokens to represent the same information. This increased tokenization could have compromised the LLM's ability to effectively simulate understanding of the concepts. English was also preferred for its extensive use in machine learning.

Another observation was that LLMs often favor short, two-sentence outputs. To address this, we implemented techniques to generate longer, more coherent paragraphs. This involved providing an example with the main idea condensed into one or two sentences, followed by explicit instructions, additional context, or question-aided expansion encouraging the model to think step-by-step, compare concepts, and elaborate in more detail, with the final option being to instruct the model to output at least 5 sentences, helping it understand the expectation for more coherent, extended output.

4 METHODOLOGY

This section outlines the methodologies used to create the Parlay dataset, focusing on refining prompting techniques, applying few-shot learning, and structured prompting with transitive reasoning. The approaches were divided into two main workflows: one utilizing GPT-3.5 (OpenAI, 2023) and GPT-4o for structured prompting, and the other using the Gemini LLM with curated few-shot learning. Below is a brief overview of each workflow, followed by detailed expansions on their key technical aspects.

4.1 Refining prompting techniques using GPT-3.5

Overview

The first workflow involved refining prompting techniques with GPT-3.5 and later GPT-4o. The process included:

- Initial tests to develop an effective strategy for generating consistent NLI examples.
- Transitioning to GPT-4o for large-scale dataset creation, emphasizing structural consistency and content clarity.
- Iteratively improving prompts to address issues like correctly generating "neutral" hypotheses and ensuring the required output lengths.

Detailed Steps

Preliminary Testing with GPT-3.5:

The process began with a familiarization phase, where GPT-3.5 was introduced to the concept of NLI datasets. Initial attempts exhibited issues, particularly with generating appropriate "neutral" labels. To address this, the project member refined the prompting strategy, giving an example from the SI-NLI dataset, to guide the LLM in following the structure and reasoning required to create diverse and consistent examples across various categories. This served as a foundation for prompting GPT-3.5 to generate its own examples. The generated examples were refined to achieve a consistent format that included the premise, hypothesis, and explanations, with a primary focus on prompting GPT-3.5 to create diverse NLI examples across various categories. On a few occasions, the project member provided prompts based on specific quotes or ideas but primarily relied on GPT-3.5's generation capabilities. A major challenge involved manually correcting examples where GPT-3.5 misinterpreted neutrality. However, some initially unclear connections between premises and hypotheses were retained when found to be logically sound upon deeper analysis.

Dataset Creation Using GPT-4o:

Following the success with GPT-3.5, the project transitioned to using GPT-4o for generating the dataset. The established structure was maintained throughout this phase by giving the LLM one of the examples from the familiarization process and asking it to create a similar example with a different topic. Prompts were refined to ensure that the generated paragraphs met the length, focusing on maintaining a five-sentence structure. Consistent output was achieved by repeatedly referencing the initial structure in subsequent prompts, allowing GPT-4o to generate coherent and structured NLI examples across various topics.

The project member chose the topic and gave it to the LLM, which then had the freedom to generate paragraphs without other limitations. An exception was made for the "News Reports" category, where the title and a source article link were provided, guiding the LLM in generating examples tied to real-world content.

4.2 Curated Few-Shot Learning with Gemini

Overview

The second workflow explored curated few-shot learning with the Gemini LLM. This approach focused on:

- Selecting well-defined data points across all five established categories.
- Employing Chain-of-Thought prompting to guide the LLM through a structured reasoning process.
- Iteratively refining prompts to improve the LLM's ability to generate logical, coherent, and insightful responses, with a particular focus on

neutral hypotheses and ensuring the required length.

Detailed Steps

Few-Shot Learning with High-Fidelity Data Points:

The process began with the careful selection of core themes for the dataset. Each theme was chosen for its potential to generate meaningful and varied NLI examples. Initial prompts were designed to lead the LLM through logical reasoning steps, helping to clarify the relationships between different elements of each data point.

Application of Chain-of-Thought Prompting and Transitive Reasoning

In this study, we enhanced inference across various contexts by employing Chain-of-Thought prompting (Wei et al., 2022) and structured prompting techniques, drawing on the dual-process theory of transitive reasoning (Wright, 2012). Chain-of-Thought prompting guided the LLM through logical steps, encouraging it to reference underlying factors and relationships in the data. This approach conditioned the LLM to infer related concepts or situations while explicitly explaining its reasoning, thereby clarifying the connections between original and inferred concepts. The method included a crucial feedback loop where initial responses were analyzed, and prompts were refined based on these analyses to improve the clarity and insightfulness of the generated explanations. This back-and-forth process ensured that the prompts became progressively more effective.

We structured the prompts to initiate transitive reasoning, facilitating the generation of intermediate steps that deepened the LLM's comprehension of the subject matter. These steps clarified relationships, connected different scenario elements, and resulted in a richer and more comprehensive set of hypotheses. This method aligns with findings by Nye et al. (2021) and Zellers et al. (2019) on breaking down complex problems into smaller parts.

Our approach was applied across five categories, each with a specific focus:

- News Reports: Extracting the core message or central theme, developing logical implications, and drawing conclusions about likely outcomes.
- Fictional Stories: Delving into character psychology by analyzing

emotions, motives, and expectations, and concluding the likely emotional or behavioral outcome.

- Historical Events: Pinpointing pivotal events, identifying cause-and-effect relationships, and concluding the broader historical impact.
- Scientific Explanations: Elucidating the main idea or phenomenon, explaining the process or mechanism involved, and concluding the significance or byproducts of the phenomenon.
- Everyday Life Situations: Focusing on common-sense reasoning by identifying the main situational factors, considering social cues, past experiences, and likely responses, then assessing the logical implications of these factors to conclude the likely outcome based on common-sense reasoning.

The prompts used for each of these categories are detailed in the provided repository: https://github.com/UL-FRI-NLP-2023-2024/ul-fri-nlp-course-project-parlay/tree/main/Working_process/second_submission/Prompt_techniques_CoT_demonstration

Addressing Challenges in Neutrality:

A significant challenge identified during this process was the LLM's interpretation of neutrality. The models often generated "neutral" hypotheses by presenting balanced opinions, juxtaposing positive and negative statements, or offering moderate viewpoints that neither supported nor opposed the main message. While this approach might seem neutral in a general sense, it did not align with the specific rationale of the NLI task, which required neutrality to involve contextually relevant statements that were not directly related to the main message. To address this, prompts were adjusted to clarify this specific interpretation of neutrality.

This refinement resulted in the inclusion of a specific prompt instruction: "The neutral hypothesis should be something mentioned in the context but not directly related to the main message and should not be made more or less likely because of the premise."

The methodology in this study shows great promise, but applying it on a larger scale brings up a few important concerns. One of the primary challenges is the identification of high-quality, referential data points for training models using few-shot learning. This process is particularly time-intensive and requires iterative prompt refinement to optimize performance, making it labor-intensive. Furthermore, the creation of prompts for generating diverse scenarios demands careful crafting to avoid repetition, a common issue with LLMs, which tend to revert to familiar themes encountered in previous outputs. In terms of scalability, the time and cost implications of extending this methodology are not insignificant. While the initial development of prompts and LLM outputs is manageable with moderate resources, large-scale dataset creation requires considerably more effort.

4.3.1 DEMONSTRATION OF STRUCTURED PROMPTING AND TRANSITIVE INFERENCING

This section illustrates the use of structured prompting and transitive reasoning in the context of an everyday life scenario and its corresponding entailment.

1. Scenario Premise Creation:

- Identify the main idea: Organize a child's birthday party with a promised ice cream bar, but face a challenge.
- Expansion: Ice cream is unavailable due to a power outage, forcing the parent to find an alternative.

2. Entailment Hypothesis Creation:

- Identify key elements of the scenario: Main problem (no ice cream) and solution (finding alternative ingredients).
- Hypothesis and its Expansion: The parent improvises with whipping cream and cookies to create a homemade ice cream substitute.
- Reasoning: The hypothesis logically follows from the scenario.

3. Transitive Reasoning Application:

- Premises: 1. Ice cream is needed. 2. Frozen foods are unavailable. 3. The parent looks for alternatives. 4. Finds whipping cream, cookies, and milk. 5. These ingredients can substitute for ice cream.
- Connecting Premises and Conclusion: Given that the promised ice cream is not available (Premise 1 + Premise 2), the parent must find a

solution (Premise 3). They find alternative ingredients (Premise 4), which can be used to make a substitute (Premise 5).

- Conclusion: The parent ensures the party's success by improvising with a homemade ice cream substitute.

For detailed analysis, refer to the provided link: https://github.com/UL-FRI-NLP-2023-2024/ul-fri-nlp-course-project-parlay/blob/main/report/Appendices/Full_demontration_of_transitive_inferencing.md

The integration of transitive reasoning and structured prompting significantly improved the relevance and coherence of hypotheses generated in the Parlay dataset. By providing a clear central idea and expanding it into detailed scenarios, the model is guided to focus on key aspects, ensuring logically connected and coherent hypotheses. For entailment, the main message drives the creation of hypotheses that logically follow the scenario, reinforcing the theme. In generating contradictions, the model produces texts that logically counter the main message, illustrating potential pitfalls. For neutral hypotheses, a strong grasp of the main message ensures the creation of responses that maintain neutrality without influencing the core idea.

5 LLM PERFORMANCE QUALITY EVALUATION

This section evaluates the performance of the LLMs used in this project, GPT-4o and Gemini, using the Outcome Quality Assessment Composed Metrics (OQACM).

To enhance the value of our dataset, we have incorporated a diverse array of output examples generated by two LLMs, specifically Gemini and GPT-4o. Instead of manually refining these examples to the level of grammatical correctness and clarity of inference that human editors might achieve, we opted for a different approach. Gemini outputs were more heavily refined through advanced prompting techniques, while GPT-4o outputs were selected based on their pre-existing structure, maintaining the fixed structure and length of the paragraphs that form each example. Several iterations of manual improvement were applied to enhance the quality of these outputs, though they were not taken to the standard of full human evaluation. These enhancements were aligned with what could be achieved through advanced prompting techniques, balancing diversity and classification utility. Thus, we aimed for a dataset with higher diversity that still conveys classification value. To assess the quality of these outputs, we developed OQACM.

5.1 Outcome Quality Assessment Composed Metrics

As presented in Table 1, the OQACM outlines the metrics used to evaluate the Parlay dataset quality. This comprehensive framework evaluates LLM outputs across various categories, with a particular focus on fluency and coherence.

Table 1: OQACM Dataset quality metrics.

Dataset quality	*Weight*	*Human evaluation*
Fluency	2.5	
*Readability	2.5/3	1-5 scale
*Grammatical accuracy	2.5/3	1-5 scale
*Style	2.5/3	1-5 scale
Coherence	2.5	
*Completeness	0.5	1-5 scale
*Conciseness	0.5	1-5 scale
*Clarity	0.5	1-5 scale
*Organization	0.5	1-5 scale
*Focus	0.5	1-5 scale
NLI success	5	
*Entailment	1.5	1-5 scale
*Contradiction	1.5	1-5 scale
*Neutral	1.5	1-5 scale
*Explanation	0.5	1-5 scale
TOTAL OQACM (sum of grade*weight)		

Meeting these metric elements is critical for generating text that is both meaningful and engaging, ensuring high-quality LLM output. Thus, the OQACM serves as a tool to evaluate which outputs are more valuable and have greater learning potential.

The OQACM is grounded in extensive literature on the critical aspects of text quality. Coherence is key to text comprehension, as coherent texts help form accurate mental representations (Kintsch & van Dijk, 1978). Halliday and Hasan (1976) analyze linguistic mechanisms that contribute to coherence, and McNamara and Kintsch (1996) show that coherence significantly enhances comprehension and retention. Additionally, Bamberg (1983) discusses practical criteria for evaluating textual coherence, offering a useful framework for our metrics. Fluency is crucial for discourse comprehension. Graesser et al. (2011) highlight the impact of grammatical accuracy, readability, and style on understanding. Reynolds and Schwartz (1983) show that clear examples and analogies enhance comprehension, underscoring the importance of style and readability. Snow (2010) emphasizes the importance of readability and grammatical accuracy in reading comprehension, especially within educational texts. Fluency ensures that text is grammatically correct, easy to read, and engaging, while coherence guarantees logical structure and comprehension.

"NLI success" refers to the core objective of an NLI dataset: determining whether a hypothesis is true (entailment), false (contradiction), or neutral relative to the premise. Human evaluators assess how well the paragraphs express these relations and the clarity of the reasoning ("Explanation"), with inference quality given more weight than the explanation. Inference quality ("NLI success") carries the highest weight in the OQACM, reflecting its importance relative to fluency and coherence. Fluency is weighted at 2.5/3 per category, and coherence at 0.5 per category, while each category within "NLI success" weights 1.5.

These weights play a crucial role in calculating the total score of the evaluation metrics. For each category, the evaluator's score (ranging from 1 to 5) is multiplied by the category's weight. The resulting products are then summed to produce the total score. Thus, the maximum achievable OQACM score is 12.5 + 12.5 + 25 = 50. Conversely, the minimum possible OQACM score, calculated using the same weighted components but with the lowest scale value of 1, is 2.5 + 0.5 + 5 = 8.

In our evaluation process, project team members score each element of the metrics on a scale from 1 to 5, guided by specific questions designed to determine the quality of the outputs.

5.2 Analysis of GPT-4O and Gemini outcomes

Our main focus was to ensure a range of examples to enhance the datasets classification effectiveness. The differences were noticeable in the complexity and diversity of inference elements, such as entailment, contradiction and neutrality across categories. Some categories featured examples with inference relationships that added depth to the dataset while others had simpler and more uniform examples following patterns like summarization or negation.

As reflected in the OQACM score, more intricate and creative examples received higher scores compared to simpler ones which although still valuable, scored lower. This intentional scoring difference highlights the nature of the dataset.

In summary, deliberate variations in complexity and diversity seen in the paragraphs generated by both models were intentionally incorporated to develop a dataset that is comprehensive, varied and effective, for classification purposes. Despite these intentional differences, our preliminary and anecdotal observations suggest that the performance of GPT-4o and Gemini in content generation is largely comparable. Both models consistently produced high-quality paragraphs. The performance of GPT-4o and Gemini was deliberately not directly compared due to the different methodologies applied in prompt refinement.

6 CONCLUSIONS

Our proposed dataset, Parlay, covers diverse situations and genres, providing a comprehensive training ground for various reasoning modes in NLI models. This diversity aims to equip models with transferable skills for novel situations and handling complex inference tasks. By providing explanations for each hypothesis, we ensure transparency in the reasoning process and enhance the dataset's value, setting a new standard for future datasets.

In the next step, we will use the dataset to train a model and see how well our approach works in real-world situations.

Manual evaluation, represented by OQACM scores, ensures high-quality training data for LLMs and provides a standardized metric that encourages consistency in evaluation across the NLI research community. Although variations in output processing between model Gemini and GPT-4o models prevent direct comparison, we plan to standardize these processes for a more meaningful evaluation.

Future research will explore transitive reasoning in greater detail, offering intermediate steps to enhance the logical progression in hypothesis derivation.

The Parlay dataset is publicly available under the Apache License Version 2.0: https://github.com/UL-FRI-NLP-2023-2024/ul-fri-nlp-course-project-parlay

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to our mentor, Aleš Žagar, whose guidance throughout the project phases was crucial to bringing this endeavor to fruition. As the faculty assistant for our NLP course lab work, Aleš's expertise and support significantly enhanced our understanding and implementation of the concepts learned during the course.

I also want to recognize the dedication and hard work of my teammate, Blaž Lipar, whose collaboration and commitment played a vital role in the success of this project.

A special thank you goes to Hasan Çağatay, Professor of Philosophy of AI at the Social Sciences University of Ankara. Our stimulating discussions on the logical processes involved in creating inference pairs, combined with his tutorship in developing more effective prompting techniques, were truly enlightening. Furthermore, Hasan reviewed both of my conference paper drafts and provided insightful feedback.

REFERENCES

Alyousef, H. S. (2021). Text cohesion in English scientific texts written by Saudi undergraduate dentistry students: A multimodal discourse analysis of textual and logical relations in oral biology texts. *Sage Open, 11*(3). Retrieved March 27, 2024, from https://doi.org/10.1177/21582440211032194

Bamberg, B. (1983). What makes a text coherent? *College Composition and Communication, 34*(4), 417–429.

Bottou, L., Peters, J., Quinonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., & Snelson, E. (2013). Counterfactual reasoning and learning systems. *arXiv preprint arXiv:1209.2355*. Retrieved March 29, 2024, from https://arxiv.org/abs/1209.2355

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326.* Retrieved March 27, 2024, from https://arxiv.org/abs/1508.05326

Brachman, R. J. & Levesque, H. J. (2004). *Knowledge Representation and Reasoning*. Elsevier.

Bublitz, W. (2011). Discursive pragmatics. In J. Zienkowski, J.-O. Östman, & J. Verschueren (Eds.), *Handbook of Pragmatics Highlights 8* (pp. 37–49).

Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., & Blunsom, P. (2018). e-SNLI: Natural language inference with natural language explanations. *arXiv preprint arXiv:1812.01193.* Retrieved March 25, 2024, from https://arxiv.org/abs/1812.01193

Chen, D., Zheng, Y., Ma, L., & Qin, F. (2023). An ontology-based knowledge representation method for typhoon events from Chinese news reports. ISPRS International *Journal of Geo-Information, 12*(9), 383. Retrieved April 3, 2024, from https://doi.org/10.3390/ijgi12090383

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439–477.

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., & Goodman, N. D. (2018). Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302.* Retrieved March 27, 2024, from https://arxiv.org/abs/1802.04302

DeepL SE. (2024). DeepL Translate. Retrieved May 27, 2024, from https://www.deepl.com/translator

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2020). ERASER: A Benchmark to Evaluate Rationalized NLP Models. *arXiv preprint* arXiv:1911.03429. Retrieved March 28, 2024, from https://arxiv.org/abs/1911.03429

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155–170.

Google DeepMind. (2024). Gemini 1.5. Retrieved May 28, 2024, from https://www.deepmind.com

Graesser, A. C. & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3*(2), 371–398.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In R. F. Lorch & E. J. O'Brien (Eds.), *Sources of coherence in reading* (pp. 132–154). Lawrence Erlbaum Associates.

Gubelmann, R., Katis, I., Niklaus, C., & Handschuh, S. (2024). Capturing the varieties of natural language inference: A systematic survey of existing datasets and two novel benchmarks. *Journal of Logic, Language and Information, 33*, 21–48.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 107112). Retrieved April 5, 2024, from https://doi.org/10.18653/v1/N18-2017

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. English Language Series. Longman.

Hossain, M. M., Kovatchev, V., Dutta, P., Kao, T., Wei, E., & Blanco, E. (2020). An analysis of natural language inference benchmarks through the lens of negation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9106–9118). Retrieved April 3, 2024, from https://doi.org/10.18653/v1/2020.emnlp-main.732

Hu, H., Richardson, K., Xu, L., Li, L., Kuebler, S., & Moss, L. S. (2020). OCNLI: Original Chinese natural language inference*. arXiv preprint arXiv:2010.05444*. Retrieved March 26, 2024, from https://arxiv.org/abs/2010.05444

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394.

Klemen, M., Žagar, A., Čibej, J., & Robnik-Šikonja, M. (2024). SI-NLI: A Slovene Natural Language Inference Dataset and Its Evaluation. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 14859–14870). Retrieved July 10, 2024, https://aclanthology.org/2022.lrec-1.221.pdf

Koreeda, Y. & Manning, C. (2021). ContractNLI: A dataset for document-level natural language inference for contracts. *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 1907–1919). Retrieved March 28, 2024, from https://doi.org/10.18653/v1/2021.findings-emnlp.164

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics* (pp. 105–122). Mouton de Gruyter.

List of geological phenomena. (2024, June 14). In *Wikipedia*. Retrieved May 22, 2024, from https://en.wikipedia.org/wiki/List_of_geological_phenomena

List of historical anniversaries. (2024, August 14). In *Wikipedia*. Retrieved May 20, 2024, from https://en.wikipedia.org/wiki/List_of_historical_anniversaries

Liu, H., Cui, L., Liu, J., & Zhang, Y. (2020). Natural language inference in context – Investigating contextual reasoning over long texts. *arXiv preprint arXiv:2011.04864.* Retrieved March 24, 2024, from https://arxiv.org/abs/2011.04864

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007.* Retrieved April 24, 2024, from https://arxiv.org/abs/1902.01007

McNamara, D. S. & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*(1), 247–287.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., & Odena, A. (2021). Show Your Work: Scratchpads for Intermediate Computation with Language Models. *arXiv preprint arXiv:2112.00114*. Retrieved March 24, 2024, from https://doi.org/10.48550/arXiv.2112.00114

OpenAI. (2023). ChatGPT version 3.5. Retrieved April 16, 2024, from https://openai.com/chat

OpenAI. (2024). ChatGPT version 4.0. Retrieved May 29, 2024, from https://openai.com/chat

Parnami, A. & Lee, M. (2022). Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291.* Retrieved March 29, 2024, from https://arxiv.org/abs/2203.04291

Paul, D. & Frank, A. (2020). Social commonsense reasoning with multi-head knowledge attention. *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2969–2980). Retrieved March 22, 2024, from https://doi.org/10.18653/v1/2020.findings-emnlp.267

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Reniers, R. L. E. P., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A questionnaire of cognitive and affective empathy. *Journal of Personality Assessment, 93(*1), 84–95.

Reynolds, R. E. & Schwartz, R. M. (1983). Relation of metaphoric processing to comprehension and memory. *Journal of Educational Psychology, 75*(3), 450–459.

Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation, 4*(1), 77–109.

Snow, C. E. (2010). *Reading for understanding: Toward an R&D program in reading comprehension*. RAND Corporation.

Tatu, M. & Srikanth, M. (2008). Experiments with reasoning for temporal relations between events. *22nd International Conference on Computational Linguistics, Proceedings of the Conference* (pp. 857–864). Manchester, UK.

Timeline of historic inventions. (2024, June 15). In *Wikipedia*. Retrieved June 8, 2024, from https://en.wikipedia.org/wiki/Timeline_of_historic_inventions

Timeline of world history. (2024, June 15). In *Wikipedia*. Retrieved June 10, 2024, from https://en.wikipedia.org/wiki/Timeline_of_world_history

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E. H., Guu, K., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*. Retrieved March 28, 2024, from https://arxiv.org/abs/2201.11903

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426.* Retrieved April 2, 2024, from https://arxiv.org/abs/1704.05426

Wright, A. G. (2012). Dual-Process Theory and Logical Reasoning: Transitive Inference as a Case Study. *Journal of Cognitive Psychology, 24*(6), 682–695.

Wu, Y., Gardner, M., Stenetorp, P., & Dasigi, P. (2022). Generating data to mitigate spurious correlations in natural language inference datasets. *arXiv preprint arXiv:2203.12942.* Retrieved March 24, 2024, from https://arxiv.org/abs/2203.12942

Wu, P., Li, X., Ling, C., Ding, S., & Shen, S. (2018). Sentiment classification using attention mechanism and bidirectional long short-term memory network. *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 3181–3187).

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601.* Retrieved March 28, 2024, from https://arxiv.org/abs/2305.10601

Yin, W., Radev, D. & Xiong, C. (2021). DocNLI: A large-scale dataset for document-level natural language inference. *arXiv preprint arXiv:2106.09449.* Retrieved April 24, 2024, from https://arxiv.org/abs/2106.09449

Zadrozny, W. & Jensen, K. (1991). Semantics of paragraphs. *Computational Linguistics, 17*(2), 171–210. Retrieved April 5, 2024, from https://aclanthology.org/J91-2003

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4791–4800).

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., & Dolan, B. (2020). DIALOGPT: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536.* Retrieved April 2, 2024, from https://doi.org/10.48550/arXiv.1911.00536

Zhao, Y., Ni, X., Ding, Y., Ke, Q., Zhao, Z., & Ji, Y. (2022). Robust low-resource natural language inference models with cross-lingual transfer. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3165–3174). Retrieved March 26, 2024, from https://doi.org/10.18653/v1/2022.naacl-main.235

Žagar, A. & Robnik-Šikonja, M. (2022). Slovene SuperGLUE Benchmark: Translation and Evaluation. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)* (pp. 2058–2065). Retrieved June 11, 2024, from https://aclanthology.org/2022.lrec-1.221.pdf

PARLAY: NA SCENARIJIH OSNOVANA PODATKOVNA MNOŽICA ZA SKLEPANJE V NARAVNEM JEZIKU

S pričujočim prispevkom uvajamo inovativno metodologijo za ustvarjanje raznolike in celovite podatkovne množice, prilagojene za naloge sklepanja v naravnem jeziku (NLI). Vključitev petih različnih kategorij scenarijev – zgodovinski dogodki, znanstvene razlage, vsakodnevne situacije, novice in fikcijske zgodbe — zagotavlja široko pokritost različnih vzorcev sklepanja, vključno s tranzitivnim sklepanjem. Metodologija uporablja iterativna in na veriženju misli osnovana navodila za pridobivanje podrobnih, kontekstno bogatih rezultatov z vmesnimi koraki sklepanja. Naša podatkovna množica izstopa po vključitvi podrobnih razlag za vsako hipotezo, kar nudi dragocen vpogled v miselni proces za vsako kategorizacijo. Izpeljali smo metriko OQACM, celovit okvir za vrednotenje izhodov velikih jezikovnih modelov glede na tekočnost, koherenco in kakovost sklepanja. Ta strukturiran pristop ne le izboljšuje interpretabilnost jezikovnih modelov, ampak tudi omogoča podroben vpogled v njihove procese sklepanja, s čimer postavlja nov standard za prihodnji razvoj podatkovnih množic za naloge sklepanja v naravnem jeziku.

Ključne besede: sklepanje v naravnem jeziku, sklepanje na ravni odstavkov, veriženje misli, veliki jezikovni modeli, ocenjevanje uspešnosti LLM, študentski prispevek

EFFICIENT FINE-TUNING TECHNIQUES FOR SLOVENIAN LANGUAGE MODELS

Camile LENDERING,[1] Manfred GONZÁLEZ,[1] Joaquín FIGUEIRA[1]

[1]Faculty of Computer and Information Science, University of Ljubljana

*{camile.lendering01, joaquin.figueira01, manfred.gonzalez01}@estudiant.upf.edu

This work adapts and evaluates several parameter-efficient fine-tuning techniques (PEFT) for two large language models: the Slovenian-specific *SloBERTa* and the cross-lingual BERT-Base-Multilingual-Uncased. The study examines methods based on Low-Ranking Adaptation (LoRA), bias update, Prefix Tuning, and Infused Adapter for Attention Approximation (IA3). Performance is assessed using benchmarks from SloBENCH, a Slovenian language evaluation framework, focusing on the following tasks: natural language understanding, named entity recognition, dependency parsing, and textual entailment recognition. The techniques and models are compared in terms of performance, memory consumption, and runtime efficiency, providing a comprehensive evaluation of PEFT methods in diverse linguistic contexts.

Keywords: LLM, PEFT, NLP, Slovenian, LoRA, BitFit, p-tuning, student paper

1 INTRODUCTION

Large language models and, by extension, language models (LM) in general are becoming ubiquitous in the current technological landscape. As such, the amount of tasks and applications that are being performed by, or with the help of these models is continuously growing. Until recently, LMs were built for specific tasks, but with the advent of Large Language Models (LLM) this has become both unnecessary, as models are complex enough now that they can perform multiple tasks well out-of-the-box, and unfeasible, as building these models takes an enormous amount of computational resources. However, if one is to achieve the optimal performance on a given task or expand the capabilities of a LM, some adaptations of the model are in order.

These adaptations usually come in the form of a process called fine-tuning, in which the parameters of a pre-trained language model (PLM) are fitted for the specific task we want to achieve. Nonetheless, given the vast number of parameters in current LLMs, full fine-tuning (FFT) of the weights of a model is usually too computationally prohibitive for most applications. As a result, in the recent years many techniques have been developed to efficiently tune LMs, focusing specifically in reducing the number of trainable parameters (refer to (L. Xu et al., 2023) for a comprehensive review of these techniques). These are usually referred to as parameter efficient fine-tuning techniques (PEFT) and are mainly based on the Lottery Ticket Hypothesis, which states that "large models are needed in pre-training only to induce (in high probability) the existence of sub-networks initialized with the correct inductive bias for learning.". In other words, that PLMs already contain the capacity and linguistic "knowledge" to perform a wide variety of tasks, but these abilities need to be accentuated through the proper learning stimulus that induces the use of the correct sub-networks when performing a particular task.

The purpose of this work is precisely to adapt and evaluate several PEFT techniques with a focus on the Slovenian language. To cover a wide range of approaches the analyzed techniques will span different families: Low-Rank Adaptation (LoRA) (Hu et al., 2022) and its derivatives, bias update (Zaken et al., 2021) and soft prompt based techniques. As a base model for analysis the SloBERTa LM will be used, which is a BERT model optimized for Slovenian. For performance assessment, the techniques will be evaluated across multiple NLP tasks, including Natural Language Inference (NLI), Named Entity Recognition (NER), Dependency Parsing (DP), and Recognizing Textual Entailment (RTE), using the benchmarks provided by SloBENCH Slovenian evaluation framework (Žitnik & Dragar, 2021).

2 LITERATURE REVIEW

2.1 LoRA

The field of adapting large language models (LLMs) for specific tasks has seen significant advancements with the introduction of LoRA. Hu et al. (Hu et al., 2022) pioneered this approach by freezing the pre-trained model weights and integrating trainable rank decomposition matrices into the Transformer architecture. This method not only substantially reduces the number of trainable parameters required for downstream tasks but also lessens the GPU memory requirements, thereby enabling comparable or superior model quality with notable efficiency. The provision of a package to facilitate the integration of LoRA with PyTorch models, including implementations for popular models like RoBERTa, DeBERTa, and GPT-2, marks a significant contribution to the field.

Further enhancing the parameter efficiency of fine-tuning PLMs, Zhang et al. (Zhang et al., 2023) introduced IncreLoRA, an incremental parameter allocation method that adds trainable parameters based on the importance scores of each module. This approach, distinguished from structured pruning methods, not only improves parameter efficiency but also incorporates early learning and restart warmup techniques to improve training effectiveness and stability. The method demonstrated superior parameter efficiency and model performance, particularly in low-resource settings, through experiments on the GLUE benchmark.

On the deployment front, Xu et al. (Y. Xu et al., 2023) proposed the Quantization-Aware Low-Rank Adaptation (QA-LoRA) algorithm, aimed at the efficient deployment of LLMs on edge devices. By introducing group-wise operators, QA-LoRA enhances the quantization flexibility while streamlining adaptation, enabling the integration of quantized LLM and auxiliary weights without compromising accuracy. This method stands out by allowing for low-bit inference directly, overcoming the limitations of previous approaches like QLoRA and enabling faster model deployment on resource-constrained devices.

Lastly, the scalability and efficiency of serving multiple LoRA adapters derived from a base model have been addressed by Sheng et al. (Sheng et al., 2023) through the introduction of S-LoRA. This system, designed for scalable serving, significantly improves throughput and the capacity to serve numerous task-specific fine-tuned models by employing a unified memory management approach and optimized computation strategies. Complementing these efforts, Gao et al. (Gao et al., 2024) introduced MoE-LoRA with Layer-wise Expert Allocation (MoLA), which optimizes the allocation of LoRA experts across different layers of the Transformer model, thereby enhancing model efficiency and performance across various NLP tasks. These advancements collectively signify a leap forward in the efficient adaptation, deployment, and serving of LLMs, paving the way for broader application and innovation in the domain.

2.2 Bias update

BitFit (Zaken et al., 2021) is the first recorded technique to have implemented a sparse fine tuning of LMs using only the bias parameters and it falls under the category of partial fine-tuning according to Xu et al.'s taxonomy (L. Xu et al., 2023). The main mechanism of the technique is simple: during fine-tuning an LM on a particular task update only the bias parameters of the model's encoder layers. These parameters account for only a small fraction of all the parameters of the model (0.1% in the case of BERT). Additionally, the authors found that fitting only a small subset of these bias parameters (mainly the bias of the query encoders of the attention heads and the biases of one of the layers of the MLP inside of the encoder layer) leads to almost no performance drop and modifies only 0.04% of the parameters. According to Xu et al. (L. Xu et al., 2023) the technique achieves good results with only a fraction of the memory footprint of other PEFTs. These promising results, combined with the existence of multiple pre-packaged implementations of the technique, led us to choose it as one of the subjects of analysis of this work.

The findings of the initial BitFit paper were further expanded in the works of Lawton et al. (Lawton et al., 2023) using neural architecture search (NAS), more specifically, iterative network pruning. The core idea of the method is to iteratively fine-tune the model using BitFit[1] and then prune it's bias parameters according to a criteria based on the first order approximation of the loss that results from eliminating certain parameters from the network. The authors found that the resulting network architectures could maintain good performance with a large portion of their bias parameters pruned, further solidifying the findings in the initial BitFit paper that only a relatively small number of bias parameters are responsible for the fine-tuned performance. Unfortunately, there is no code or implementation freely available to replicate the results of this technique in our work.

[1] The authors also fine-tuned the models using LoRA, but that falls outside of the scope of this analysis.

2.3 IA3

Infused Adapter for Attention Approximation (IA3) (Liu et al., 2022) is a PEFT technique that focuses on modifying the attention mechanism of a language model by introducing lightweight, learnable adapters that adjust its attention scores. This approach allows for significant reductions in the number of trainable parameters, making fine-tuning more resource-efficient while maintaining or even improving performance on specific downstream tasks. IA3 achieves this by infusing additional parameters only in critical parts of the attention mechanism, thus preserving the core structure and pre-trained knowledge of the original model. The main advantage of IA3 with respect to other PEFT techniques is precisely this infusion mechanism of the learnable parameters, which means that no additional weights have to be added to the model after training and hence inference time remains unchanged.

2.4 Soft prompts

Soft prompts, represent various methods to efficiently adapt LLMs for downstream tasks without altering the underlying model architecture or weights. This technique involves appending a sequence of tunable tokens, or ”soft prompts,” to the input of the model. During fine-tuning, these tokens are optimized to guide the model towards generating the desired task-specific output. Two prominent methods for implementing soft prompts are prompt tuning and Prefix Tuning (P-tuning).

2.4.1 PROMPT TUNING

Prompt tuning introduces a set of ℓ learnable tokens (soft prompts), denoted as $P = \{P_1, P_2, \ldots P_\ell\}$, and concatenates these tokens to beginning of the input to the model $X \in \mathbb{R}^{n\times d}$ to form $\hat{X} \in \mathbb{R}^{(n+\ell)\times d}$. Throughout the fine-tuning process, only the parameters associated with the prompt tokens P are adjusted via gradient descent, while the pre-trained model parameters are kept fixed. Hence, the length of the prompt and the dimensionality of the token embeddings determine the parameter cost for fine-tuning. (Lester et al., 2021)

2.4.2 PREFIX-TUNING

Prefix-tuning introduces the idea of appending a set of soft prompts $P = \{P_1, P_2, \ldots P_\ell\}$, not to the input layer but to the hidden states within the multi-head attention layers of the model. This is different from prompt tuning, which concatenates soft prompts directly to the input. To promote stability during training, a feed-forward network (FFN) is used to parameterize these soft prompts. During fine-tuning, two distinct sets of prefix vectors, $\hat{P}_K$ and $\hat{P_V}$, are concatenated to the attention layer's original key (K) and value (V) vectors, respectively. Hence, the only parameters that require optimization are those of $\hat{P}_K$, $\hat{P_V}$, and the FFN. Once the model is fine-tuned, the FFN is no longer needed, and only the optimized key and value prefix vectors are kept for model inference. (Li & Liang, 2021)

3 PROPOSED METHODOLOGY

This section outlines the methodology used in our experiments, covering data preparation, model fine-tuning, evaluation, and resource measurement. All the code used to perform these tasks can be found in the following Github repository: https://github.com/UL-FRI-NLP-2023-2024/ul-fri-nlp-course-project-pirates-of-the-caribbean.

3.1 Data Preparation

We began by selecting datasets from the SloBENCH framework, focusing on tasks like Natural Language Inference (NLI), Named Entity Recognition (NER), Dependency Parsing (DP), and Recognizing Textual Entailment (RTE).

For NLI, we used the SI-NLI (Slovene Natural Language Inference) dataset, containing 5,937 Slovene sentence pairs labeled as ”entailment,” ”contradiction,” or ”neutral.” These pairs were derived from the Slovenian reference corpus ccKres (Logar et al., 2013) and balanced by annotators who created three variations for each pair. The dataset is split into training (4,392 pairs), validation (547 pairs), and test sets (998 pairs), with the splits designed to evenly distribute difficult and easy instances. The dataset is provided as part of the SloBENCH evaluation framework (Žitnik & Dragar, 2021).

For both NER and DP tasks, we used the SSJ500K corpus, containing around 500,000 tokens annotated for tokenization, sentence segmentation, morphosyntactic tagging, and lemmatization. Half of the corpus includes syntactic dependencies, named entities, and verbal multiword expressions, while a quarter is annotated with semantic role labels. The data was split randomly into 80% for training and 20% for testing, with 10% of the training data further used for validation (Krek et al., 2021).

For the RTE task, we used the Slovene translation of the SuperGLUE benchmark, which includes translations from both machine and human translators. Specifically, the RTE dataset in this benchmark was translated by Google Machine Translation, with some portions also translated by human translators. The dataset was randomly split into 233 examples for training, 30 for validation, and 30 for testing. (Žagar et al., 2020).

3.2 Fine-Tuning Techniques

We applied four PEFT methods: LoRA, BitFit, IA3, and Prefix Tuning. These methods were selected to cover a range of approaches, from modifying attention mechanisms to introducing trainable tokens. Each technique offers a different strategy for reducing the number of trainable parameters, freezing different parts of the language models to gain resource efficiency.

3.3 Experimental Setup

All experiments were conducted on a Tesla V100 GPU on the Arnes HPC cluster. Training parameters such as batch size, number of epochs, and learning rates were standardized across all methods to ensure fair comparisons. For each PEFT technique, we fine-tuned the SloBERTa and BERT Base Multilingual Uncased (BBMU) models on the selected datasets. While BBMU is a multilingual model designed to handle multiple languages, our experiments were conducted exclusively in Slovene, rather than in a cross-lingual setting. Training was performed for a fixed number of epochs.

3.4 Evaluation Metrics

We assessed model performance using standard metrics: precision, recall, and F1-score. These metrics were chosen to provide a comprehensive view of model accuracy and balance between precision and recall. Additionally, memory consumption and runtime during training were recorded for each PEFT method to evaluate the trade-off between performance and resource efficiency. As for the memory consumption, we recorded the allocated Memory which is the amount of GPU memory actively used by the tensors at each step of the process. This metric is essential for understanding the memory footprint of the model and ensuring that the GPU resources are utilized efficiently without exceeding capacity.

4 RESULTS

This section outlines the bench marking results for various Slovene and multilingual LLMs applied to different NLP tasks, including Named Entity Recognition (NER), Dependency Parsing (DP), Natural Language Inference (NLI), and Recognizing Textual Entailment (RTE). We evaluated the performance using popular PEFT methods to determine their effectiveness across these different language processing tasks.

4.1 Natural Language Inference (NLI)

NLI involves determining whether a ”premise” sentence logically supports, contradicts, or is neutral to a ”hypothesis” sentence. Using the SI-NLI dataset, this study uses $5,937$ Slovene sentence pairs from the Slovenian reference corpus ccKres, each manually annotated with labels of ”entailment,” ”*contradiction*,” or ”*neutral*.”

Table 1: **Performance Comparison on SI-NLI** This table shows the results of training SloBERTa and BBMU models using various Parameter-Efficient Fine-Tuning techniques and Full Fine-Tuning on the SI-NLI dataset. Performance is measured by mean Accuracy, F1 Score, and training time (seconds) in 5 executions of the experiment.

Model	PEFT Method	Accuracy	F1 Score	Time (s)
SloBERTa	LoRa	**68.5%**	**68.0**%	260.5
SloBERTa	Prefix Tuning	35.3%	28.9%	235.1
SloBERTa	IA3	54.9%	56.4%	245.1
SloBERTa	BitFit	33.8%	20.3%	**106.1**
SloBERTa	Full	49.5%	47.7%	323.3
BBMU	LoRa	51.7%	50.8%	297.8
BBMU	Prefix Tuning	43.8%	44.0%	237.7
BBMU	IA3	**52.3%**	**51.6**%	247.2
BBMU	BitFit	32.2%	22.8%	**107.2**
BBMU	Full	51.1%	46.6%	321.4

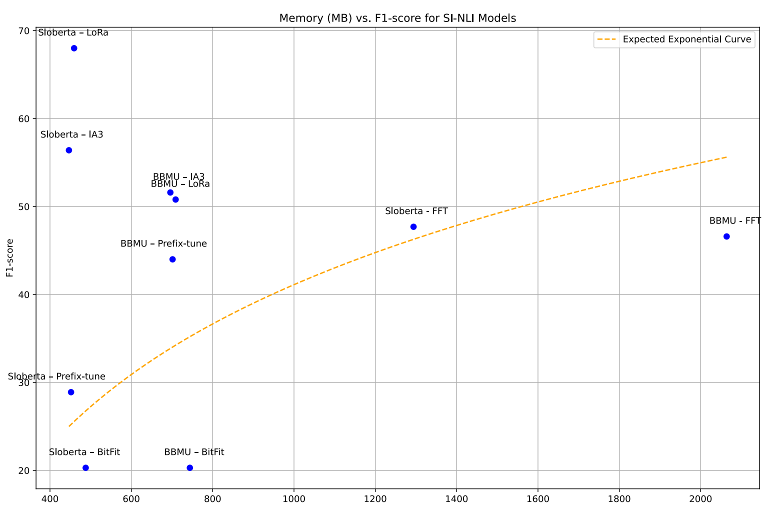


Figure 1: Memory (MB) vs. F1-score for various models and PEFT methods for the SI-NLI task.

The results in the Table 1 indicate that SloBERTa with LoRa achieves the highest accuracy (68.5%) and F1 score (68.0%), while BitFit shows the lowest performance for both models, with accuracies of 33.8% and 32.2%. IA3 offers a moderate performance, and Full fine-tuning results in varied outcomes. Training times range from 235 to 323 seconds, highlighting efficiency differences among methods.

The exponential curve in Figure 1 is included as a general reference to illustrate deviations from the expected trend, where increased memory usage typically correlates with improved performance. However, this relationship is not always straightforward, as demonstrated by our results, which show that some PEFT methods achieve higher efficiency, increasing performance without a proportional increase in memory usage.

Methods such as LoRa and IA3 for the SloBERTa model demonstrate high F1 scores with relatively lower memory usage compared to other methods. This suggests that these PEFT methods are more efficient, providing better performance without the need for many computational resources. However, BitFit has lower performance despite its lower memory requirements, indicating that not all PEFT methods offer the same benefits.

4.2 Name Entity Recognition (NER)

The objective of NER is to identify specific key elements in the words of a sentence. For instance, an NER model may be able to identify when a text mentions or refers to a person, a particular location, an organization, etc. In this experiment we use the SSJ500K training dataset to benchmark several PEFT techniques in the same LM as in the previous task: the Slovene language specific *SloBERTa* and the multilingual BBMU. The corpus contains about 500.000 manually annotated tokens for token-level classification in the Slovene language. Furthermore, it contains several subsets for different tasks; we used the NER subset which contains roughly 9,500 manually annotated sentences using the entities of the CoNLL-2003 shared task.

Table 2: **Performance Comparison on SSJ500K's NER subset** The table shows the result of training the *SloBERTa* and BBMU models using different PEFT techniques and FFT as a baseline in the NER subset of the SSJ500k corpus. All results where generated using three training epochs with batches of 16 samples. All the statistics obtained are averages from 5 execution of the experiments.

Model	Method	Accuracy	F1 score	Training Time (s)
SloBERTa	LoRA	**85.7%**	**88.2%**	85.5
SloBERTa	P-Tuning	80.8%	84.0%	68.3
SloBERTa	IA3	82.3%	85.3%	**40.2**
SloBERTa	BitFit	14.9%	16.9%	72.5
SloBERTa	FFT	85.8%	88.0%	107.5
BBMU	LoRA	80.0%	82.3%	115.2
BBMU	P-Tuning	75.1%	78.5%	**89.8**
BBMU	IA3	65.1%	68.5%	95.3
BBMU	BitFit	17.4%	18.5%	94.2
BBMU	FFT	**80.9%**	**82.8%**	153.2

The results show decent performance of almost all methods with the *SloBERTa* model, with LoRA even surpassing full fine-tuning with the same amount of training epochs and batch size. Furthermore, all methods reduced considerably the training time over full fine tuning. In fact, Prefix Tuning almost halved the training time of FFT for the *SloBERTa* model, achieving a training time of only 63% of that of FFT. Additionally, of all techniques LoRA performs by far the best. This is expected as LoRa has proven efficacy. Regarding the other methods, it's surprising that such a simple method as Prefix Tuning can outperform IA3 by such a wide margin in one of the models, specially considering IA3 is intended to be an improvement on LoRA.

Regarding the performance comparison between the two models, it seems *SloBERTa* performs better overall in all experiments. This is to be expected given *SloBERTa* was specifically trained to perform well in Slovene. Additionally, it seems that it responds better to PEFT techniques, hinting at the fact that these techniques may not be as efficient when some cross-lingual transfer learning is required, such as the one required to fine-tune the BBMU model.

Finally, we can see a graph of the different models and techniques combinations performance (f1) vs. memory consumption in figure 2. We can observe similar results as in the previous NLI task, where the results don't seem to follow the expected logarithmic curve of diminishing returns. Rather, we see a much flatter curve where the same results can be achieved with an extensively reduced memory footprint. The only outlier seems to be BitFit as seen previously, which, while still using limited resources, achieves very poor performance.

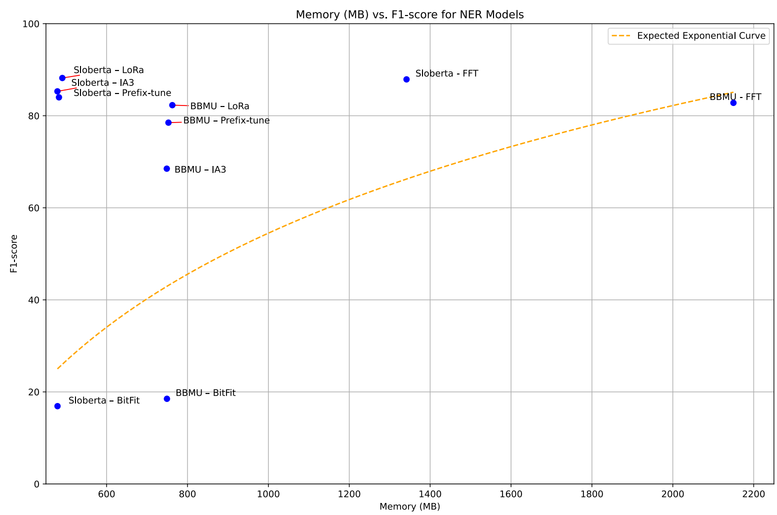


Figure 2: Memory (MB) vs. F1-score for various models and PEFT methods for the NER task.

4.3 Dependency Parsing

Using a different subset from the previous SSJ500k dataset we developed further experiments to evaluate the techniques' performance in Dependency Parsing (DP). In particular, we wanted to predict for each word, the relation to its root. The subset we used contains about 11 400 synthetically annotated sentences with their corresponding dependency tree. Each word has assigned both a relation to its head or root and an index identifying the mentioned root. We used the same experimental setup as in the previous NER task. The results can be seen in table 3.

Table 3: **Performance Comparison on SSJ500K's DP subset** The table shows the result of training the *SloBERTa* and BBMU models using different PEFT techniques and FFT as a baseline in the Dependency Parsing task subset of the SSJ500k corpus. All results where generated using three training epochs with batches of 32 samples, and all the statistics obtained are averages from 5 execution of the experiments.

Model	Method	Accuracy	F1 Score	Time (s)
SloBERTa	LoRA	**94.0%**	**94.1%**	45.6
SloBERTa	Prefix Tuning	91%	90.9%	**37.2**
SloBERTa	IA3	91.9%	91.9%	39.3
SloBERTa	BitFit	83.7%	82.1%	77.8
SloBERTa	FFT	93.1%	93.1%	56.0
BBMU	LoRA	**90.4%**	**90.4%**	58.3
BBMU	Prefix Tuning	85.6%	85.3%	**50.8**
BBMU	IA3	85.9%	85.5%	52.8
BBMU	BitFit	78.3%	77.0%	113.9
BBMU	FFT	90.3%	90.2%	78.7

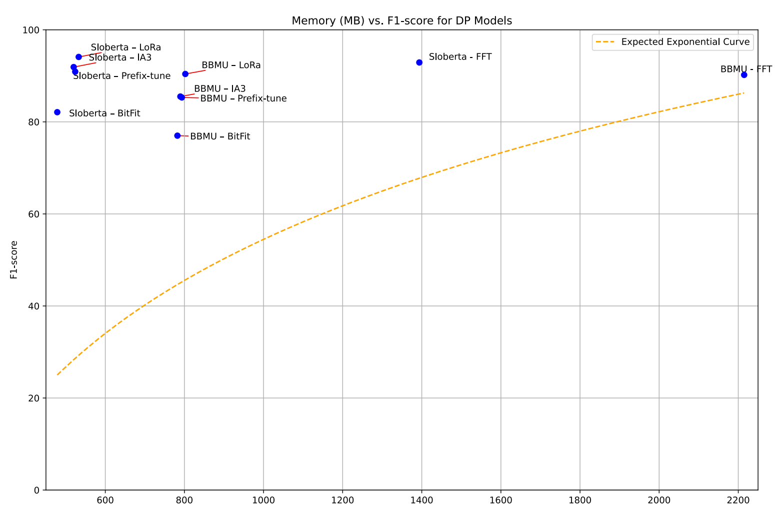


Figure 3: Memory (MB) vs. F1-score for various models and PEFT methods for the DP task.

In this task we observe similar results as in the previous one. LoRA achieves by far the best results in both models, surpassing FFT with significantly lower training times. However, it seems Prefix Tuning never performs better than IA3. The reason for this may be that the task is relatively more complex, requiring the model to discover deeper relationships between the tokens than in NER, so a more complex approach that focuses on enhancing the attention mechanism of the models may be a better strategy to tackle it. Overall, it seems that PEFT techniques perform quite well, specially with the *SloBERTa* model, obtaining similar or even superior results with significantly reduced training times.

Regarding the comparison between the models, similar results are observed as with the NER task: performance is better in the Slovene specific *SloBERTa* and the PEFT methods are less effective in the multilingual model. Finally, just as in the previous experiments, we can see a comparison between memory consumption and performance for this task in figure 3. We see almost the same trend as before, with higher, but similar, performance and slightly lower memory consumption. The only difference is that BitFit performs much better in this task, indicating that BitFit seems to be very sensitive to its particular application: it may achieve good results with reduced memory usage on some very specific tasks, while completely failing in others. Finally, BitFit has by far the highest run time.

4.4 Recognizing Textual Entailment (RTE)

Table 4: **Performance Comparison on RTE** This table summarizes the mean F1 Score and training time (seconds) across five runs for SloBERTa and BBMU models fine-tuned with various Parameter-Efficient Fine-Tuning techniques and Full Fine-Tuning on the SuperGLUE Human-based RTE dataset from SloBENCH.

Model	PEFT Method	F1 Score	Training Time (s)
SloBERTa	LoRa	33.2%	17.53
SloBERTa	Prefix Tuning	**54.9%**	17.01
SloBERTa	IA3	32.4%	87.67
SloBERTa	BitFit	39.2%	97.94
SloBERTa	FFT	37.1%	**10.58**
BBMU	LoRa	65.5%	**24.78**
BBMU	Prefix Tuning	40.6%	26.90
BBMU	IA3	46.9%	84.36
BBMU	BitFit	**47.8%**	81.78
BBMU	FFT	38.1%	51.29

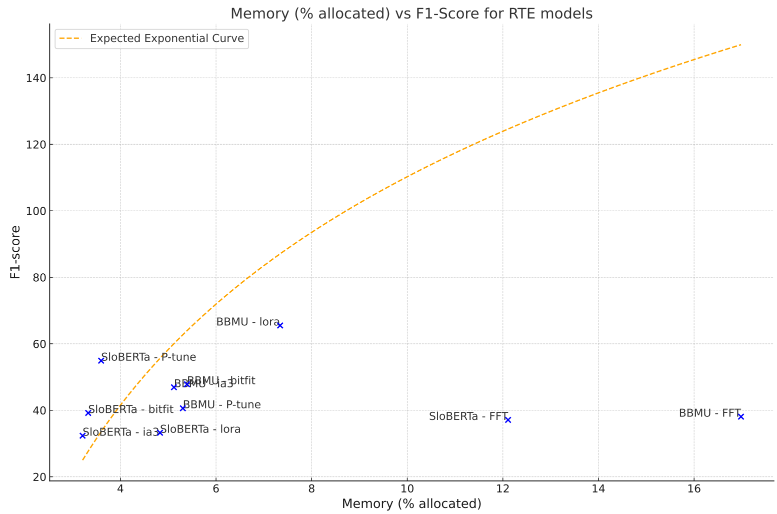


Figure 4: Memory (% allocated) vs. F1-score for various models and PEFT methods for the RTE task.

In this section, we analyze how LoRA fine tuning influences over the F1 Scores over the human-translated SuperGLUE task RTE dataset from SloBENCH.

4.4.1 LORA HYPER-PARAMETER TUNING

In the table 4 we can observe that models like LoRA have bad performance with this small dataset. For this reason we want to explore a bit more this model and see opportunities of improvement by tweaking some of it's hyper parameters. In this experiment, the *SloBERTa* model is fine-tuned on the RTE dataset utilizing the LoRa approach to explore the impact of various hyper parameters on the model's performance, particularly focusing on the F1 score. The experiment iteratively tests combinations of three key hyper parameters: r, α, and *weight decay*.

- r **(rank):** This parameter determines the rank of the adaptation in LoRA (Hu et al., 2022), which affects the model's capacity to learn new patterns without significantly increasing complexity. Values explored were [8, 16, 32], enabling an investigation into how increasing the rank influences learning capacity and generalization.

- α **(scaling factor):** This controls the scaling of the updates in LoRA (Hu et al., 2022), impacting how aggressively the model adapts to the new dataset. Tested values were [16, 32, 64], providing insights into finding a balance between overly subtle and overly aggressive updates.
- **Weight Decay (regularization):** Used to prevent over fitting by penalizing larger weights, with values set at [0.01, 0.1, 0.2]. This parameter explores how varying levels of regularization affect over fitting and model performance on validation data.

The function used for this experiment orchestrates a grid search over these parameters, training the *SloBERTa* model for each configuration and logging the F1 score to identify the optimal setup. The configuration leading to the highest F1 score is noted as the best, reflecting the most effective balance of learning capacity, adaptability, and regularization for this task. Metrics and configurations are saved, and the best model setup is printed, showcasing the impact of tuning LoRA parameters on model efficacy in a sequence classification task on the RTE dataset. The results can be observed in the figures 5, 6, 7 where the best parameters to be used under this configuration are r=16, α=64, weight decay=0.1. The best F1 Score achieved with *SloBERTa* was about 0.52 over this dataset, the future work will be focused on compare more models to see which one performs better in this dataset that represents a challenge since it is small and in Slovene.

4.4.2 RTE BEST MODELS

In this subsection we want to explore how early stopping can help by doing the correct parameter setting of these models to achieve better results than table 4. The aim in the experiments of table 5 was to identify the best combination of patience and threshold values for early stopping that maximizes the F1 score for each fine-tuning approach. We fine-tuned SloBERTa on a specific dataset using four different parameter-efficient methods: LoRA (Low-Rank Adaptation), Prefix Tuning, IA3 (Incremental Adaptation), and BitFit (Bias Fine-Tuning). For each method, we tested a range of patience and threshold values for early stopping.

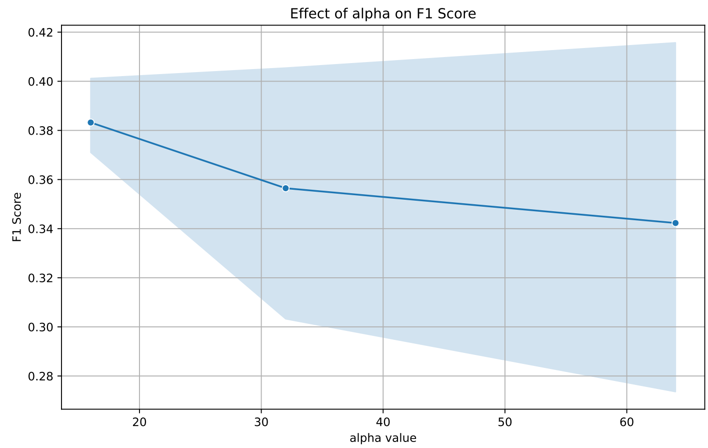


Figure 5: Influence of α on the F1-Score.

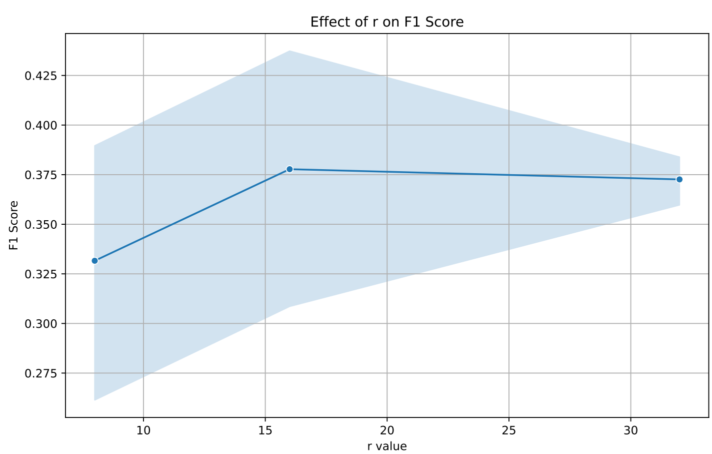


Figure 6: Influence of r on the F1-Score.

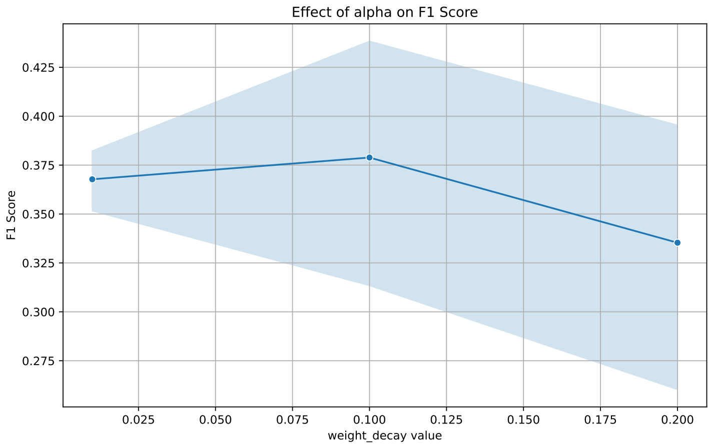


Figure 7: Influence of weight decay on the F1-Score.

Table 5: Best evaluation F1 Scores for each approach with corresponding epoch, patience, and threshold parameters over SloBERTa model.

Method	Patience	Threshold	Eval F1	Epoch
BitFit	1	1×10^{-3}	53%	15
IA3	6	1×10^{-2}	**77%**	9
LoRA	1	1×10^{-2}	54%	3
Prefix Tuning	3	1×10^{-1}	71%	4

The patience values tested in table 5 were [1, 3, 5, 6], and the threshold values were [1×10^{-1}, 1×10^{-2}, 1×10^{-3}, 1×10^{-4}]. Epoch column in the table shows at which epoch the best F1 score was achieved. The patience parameter specifies the number of epochs to wait for an improvement in the validation loss before stopping the training. Threshold parameter defines the minimum change in the monitored quantity to qualify as an improvement.

The results indicate that different early stopping parameters significantly affect the performance of the fine-tuning methods. Notably, the IA3 method achieved the highest F1 score of 0.77 with a patience of 6 and a threshold of 1×10^{-2}. Prefix Tuning also performed well, with an F1 score of 0.71, using a patience of 3 and a threshold of 1×10^{-1}. LoRA and BitFit methods showed moderate performance with the best F1 scores of 0.54 and 0.53, respectively. LoRA achieved its best performance with a patience of 1 and a threshold of 1×10^{-2}, while BitFit required a patience of 1 and a much lower threshold of 1×10^{-3} to achieve its best result.

These findings suggest that more aggressive early stopping (lower patience and higher threshold) may not always yield the best performance. Instead, a more balanced approach, with moderate patience and threshold values, appears to be more effective for certain fine-tuning methods. This experiment underscores the importance of tuning early stopping parameters to optimize model performance and convergence. These insights can guide future experiments and practical applications involving the fine-tuning of pre-trained models, especially in resource-constrained environments where parameter efficiency is crucial.

5 DISCUSSION

From the experiments and results we obtained, several conclusions can be drawn. Firstly, LoRA consistently emerged as the best performing method across most tasks, which aligns with its proven efficacy in the literature and its widespread use in the industry. The only exception to this general observation was the RTE benchmarks, in which it came out last. This last result highlights the fact that while LoRA is often a good choice when task specific context is missing, it's always best to evaluate tasks and datasets individually to determine the best selection of fine tuning techniques.

Additionally, both IA3 and Prefix Tuning demonstrated competitive performance with LoRA, offering slightly reduced memory footprints. In contrast, BitFit showed inconsistent performance overall, with memory consumption similar to other methods. In some instances, BitFit's performance was notably poor, occasionally dropping to single-digit F1 scores. This outcome is not surprising, as BitFit represents an early and relatively simplistic approach to PEFT. Although it was pioneering at the time of its development, it has been surpassed by more advanced and sophisticated techniques in the rapidly evolving field of PEFT.

Furthermore, it is important to note that PEFT techniques outperformed FFT on multiple occasions during our benchmarks. This demonstrates that PEFT methods not only make large language models (LLMs) and advanced NLP techniques more accessible to lower-end hardware but also enhance performance on high-end systems. Our results show little to no correlation between memory consumption and performance across all tasks and benchmarks, with minor exceptions due to BitFit's inconsistent performance.

Given that this work heavily relied on Slovene benchmarks, a few remarks about multilingual learning are pertinent. Notably, SloBERTa outperformed BBMU in all tasks and across all PEFT techniques. This result is expected, as SloBERTa was specifically trained on the Slovene language. However, the significant performance improvements seen in BBMU when using PEFT methods highlight the effectiveness of these techniques even in cross-lingual settings.

Finally, we would like to address the sustainability aspect of our work. The widespread use of large language models (LLMs) has a substantial environmental impact (Jiang et al., 2024) Therefore, adopting PEFT techniques to reduce training time and computational resource requirements represents a significant advancement for the sustainability of the NLP field. However, a limitation of our study is that we have not yet tested these methods on larger language models with over 1 billion parameters, which would be an important direction for future research. The notable reductions in both time and memory usage observed across all models and tasks in our study strongly support this premise. By making NLP more efficient, PEFT techniques not only enhance accessibility but also contribute to more environmentally responsible NLP practices.

ACKNOWLEDGEMENT

We thank our colleagues from the Faculty of Computer and Information Science, University of Ljubljana, especially Boshko Koloski, for their insightful recommendations and feedback.

REFERENCES

Gao, C., Chen, K., Rao, J., Sun, B., Liu, R., Peng, D., ... Subrahmanian, V. (2024). *Higher layers need more lora experts.*

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=nZeVKeeFYf9

Jiang, P., Sonne, C., Li, W., You, F., & You, S. (2024, 04). Preventing the immense increase in the life-cycle energy and carbon footprints of llm-powered intelligent chatbots. *Engineering*. doi: 10.1016/j.eng.2024.04.002

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., ... Zajc, A. (2021). *Training corpus ssj500k 2.3.* Retrieved from http://hdl.handle.net/11356/1434 (Slovenian language resource repository CLARIN.SI)

Lawton, N., Kumar, A., Thattai, G., Galstyan, A., & Steeg, G. V. (2023). *Neural architecture search for parameter-efficient fine-tuning of large pre-trained language models.* arXiv. Retrieved from https://arxiv.org/abs/2305.16597 doi: 10.48550/ARXIV.2305.16597

Lester, B., Al-Rfou, R., & Constant, N. (2021). *The power of scale for parameter-efficient prompt tuning.* arXiv. Retrieved from https://arxiv.org/abs/2104.08691 doi: 10.48550/ARXIV.2104.08691

Li, X. L., & Liang, P. (2021). *Prefix-tuning: Optimizing continuous prompts for generation.* arXiv. Retrieved from https://arxiv.org/abs/2101.00190 doi: 10.48550/ARXIV.2101.00190

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. (2022). *Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.* arXiv. Retrieved from https://arxiv.org/abs/2205.05638 doi: 10.48550/ARXIV.2205.05638

Logar, N., Erjavec, T., Krek, S., Grčar, M., & Holozan, P. (2013). *Written corpus ccKres 1.0.* Retrieved from http://hdl.handle.net/11356/1034 (Slovenian language resource repository CLARIN.SI)

Sheng, Y., Cao, S., Li, D., Hooper, C., Lee, N., Yang, S., ... Stoica, I. (2023). *S-lora: Serving thousands of concurrent lora adapters.*

Xu, L., Xie, H., Qin, S.-Z. J., Tao, X., & Wang, F. L. (2023). *Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment.* arXiv. Retrieved from https://arxiv.org/abs/2312.12148 doi: 10.48550/ARXIV.2312.12148

Xu, Y., Xie, L., Gu, X., Chen, X., Chang, H., Zhang, H., ... Tian, Q. (2023). *Qa-lora:*

Quantization-aware low-rank adaptation of large language models.

Žagar, A., Robnik-Šikonja, M., Goli, T., & Arhar Holdt, Š. (2020). *Slovene translation of SuperGLUE.* Retrieved from http://hdl.handle.net/11356/1380 (Slovenian language resource repository CLARIN.SI)

Zaken, E. B., Ravfogel, S., & Goldberg, Y. (2021). *Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models.* arXiv. Retrieved from https://arxiv.org/abs/2106.10199 doi: 10.48550/ARXIV.2106.10199

Zhang, F., Li, L., Chen, J., Jiang, Z., Wang, B., & Qian, Y. (2023). *Increlora: Incremental parameter allocation method for parameter-efficient fine-tuning.*

Žitnik, S., & Dragar, F. (2021). *SloBENCH evaluation framework.* Retrieved from http://hdl.handle.net/11356/1469 (Slovenian language resource repository CLARIN.SI)

UČINKOVITE TEHNIKE PRILAGAJANJA ZA SLOVENSKE JEZIKOVNE MODELE

To delo prilagaja in ocenjuje več tehnik učinkovitega prilagajanja parametrov (PEFT) za dva velika jezikovna modela: specifično slovenski *SloBERTa* in večjezični BERT-Base-Multilingual-Uncased. Študija preučuje metode, ki temeljijo na nizko rangirani prilagoditvi (LoRA), posodobitvi pristranskosti, prilagajanju s predponami (Prefix Tuning) in vstavljenem adapterju za približek pozornosti (IA3). Učinkovitost je ocenjena z uporabo meril SloBENCH, okvira za ocenjevanje slovenskega jezika, s poudarkom na naslednjih nalogah: razumevanje naravnega jezika, prepoznavanje poimenovanih entitet, odvisnostno razčlenjevanje in prepoznavanje besedilnih implikacij. Tehnike in modeli so primerjani glede na učinkovitost, porabo pomnilnika in čas izvajanja, kar zagotavlja celovito oceno PEFT metod v različnih jezikovnih kontekstih.

Ključne besede: prilagajanje parametrov, PEFT, slovenski jezikovni modeli, SloBERTa, SloBENCH

UPORABA ŠESTIH MER SKLADENJSKE KOMPLEKSNOSTI ZA PRIMERJAVO JEZIKA V GOVORNEM IN PISNEM KORPUSU

Luka TERČON

Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva cesta 2, Ljubljana, Slovenija

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Večna pot 113, Ljubljana, Slovenija

Obstajajo številne metode za merjenje skladenjske kompleksnosti v digitaliziranih bazah jezika. Jezikovni korpusi, posebej takšni, ki vsebujejo skladenjske oznake, nam omogočajo, da analize in primerjave skladenjske kompleksnosti izvedemo avtomatsko in učinkovito. V tem prispevku predstavljam metodo za avtomatsko primerjavo dveh korpusov – korpusa pisnih besedil in korpusa govorjenih besedil – s pomočjo šestih uveljavljenih mer skladenjske kompleksnosti. Rezultati kažejo, da je skladenjska sestava jezika v pisnem korpusu nekoliko kompleksnejša kot v govornem korpusu. Razlike so najbolj izrazite predvsem pri dolžini povedi in globini skladenjskih dreves. Analiza korelacije med različnimi merami nakazuje na to, da nekatere od uporabljenih mer podajo precej drugačno informacijo o skladenjski sestavi neke povedi kot druge.

Ključne besede: skladenjska kompleksnost, pisni korpus, govorni korpus, mere kompleksnosti, študentski prispevek

1 UVOD

S čedalje večjo digitalizacijo jezikovnih virov se pojavlja vedno več možnosti za avtomatsko analizo besedil. Pojav skladenjsko označenih korpusov omogoča hitro analizo skladenjskih vzorcev, ki se pojavljajo v slovenščini, tako v govorjenih kot tudi pisnih besedilih. Prav jezikovna zvrst v smislu prenosnega medija (i. e. ali imamo opravka z govorjenim ali pisnim jezikom) je eden od glavnih dejavnikov, ki v veliki meri vpliva na skladenjsko kompleksnost jezika (Ehret in sod., 2023). Lintunen in Mäkilä (2014) v svoji študiji pokažeta, da obstaja neka razlika v skladenjski kompleksnosti med govorjenim in pisnim jezikom, vendar ta ni nujno višja v pisnih besedilih, kot je pogosto predpostavljeno, pač pa lahko na končni rezultat vplivajo številni dejavniki, kot so vrsta izbrane mere in način segmentiranja besedil na povedi oz. druge osnovne enote. Razlika v skladenjski kompleksnosti med govorjenim in pisnim jezikom torej ni tako jasno opredeljena.

Slovenska korpusa SSJ-UD (Dobrovoljc in sod., 2017) in SST-UD (Dobrovoljc in Nivre, 2016) vsebujeta jezikoslovne oznake, med katerimi so tudi skladenjske odvisnostne relacije po sistemu Universal Dependencies,[1] zato sta dobra vira za raziskovanje skladenjskih značilnosti besedil v slovenščini. SST-UD vsebuje transkripte govora, SSJ-UD pa pisna besedila, zato lahko SST-UD obravnavamo kot reprezentativen korpus govorjene slovenščine (Dobrovoljc in Nivre, 2016, str. 1566), SSJ-UD pa kot reprezentativen korpus pisne slovenščine (Dobrovoljc in sod., 2017, str. 34).

V svoji raziskavi preučujem razlike na ravni skladenjske kompleksnosti med tema dvema korpusoma s pomočjo šestih uveljavljenih mer skladenjske kompleksnosti. Predmet raziskave so tako potencialne razlike v skladenjski kompleksnosti kot tudi razlike med samimi merami, ki jih uporabljam za primerjanje. Številne študije izpostavljajo večplastnost pojma kompleksnost v jezikoslovju in poudarjajo, da tudi znotraj iste domene (skladnja, oblikoslovje, itd.) ene prave stabilne definicije kompleksnosti ni (Ehret in sod., 2023; Bentz in sod., 2023; Berdicevskis in sod., 2018; Jagaiah in sod., 2020). Posledično različne uveljavljene mere, ki se uporabljajo za merjenje skladenjske kompleksnosti, lahko podajo precej različne informacije o obravnavanem jeziku. V luči tega sem oblikoval dve raziskovalni vprašanji, ki naslavljata tako razlike med obema jezikovnima zvrstema kot tudi razlike med uporabljenimi merami, ko jih apliciramo na slovenski jezik:

1. Ali se skladenjska kompleksnost v govorjenem in pisnem jeziku na kakšen način razlikuje?

2. Se izbrane mere skladenjske kompleksnosti med seboj kako razlikujejo?

Sistematičnih raziskav skladenjske kompleksnosti v slovenskih besedilih je relativno malo. Gaberšček (2018) v svoji raziskavi za analizo pisnih besedil, ki so jih ustvarili osnovnošolci, uporabi tako novonastale metode za merjenje skladenjske kompleksnosti, ki jih oblikuje na podlagi štirih identificiranih meril kompleksnosti (merilo dolžine, strukturne kompleksnosti, pogostnosti in ra-

[1] https://universaldependencies.org/

znolikosti), kot tudi že uveljavljene, starejše metode. Martinc in sod. (2021) se v svoji študiji ukvarjajo predvsem z načini merjenja berljivosti v angleških in slovenskih besedilih, vendar med tradicionalnimi pristopi omenjajo tudi mere skladenjske kompleksnosti, kar nakazuje na določeno mero prekrivanja med obema konceptoma. Kolikor se zavedam, je moja študija prva, ki preučuje razlike v skladenjski kompleksnosti med govorjeno in pisno slovenščino z uporabo skladenjsko razčlenjenih korpusov.

V drugem razdelku najprej predstavim oba uporabljena korpusa in shemo skladenjskih oznak, ki jo korpusa uporabljata, ter vsako od šestih izbranih mer skladenjske kompleksnosti. V tem razdelku opišem tudi statistične teste in metode, ki jih uporabljam za preverjanje rezultatov. Tretji razdelek predstavlja glavne rezultate raziskave in ugotovitve statističnih preizkusov. V četrtem razdelku podam svojo interpretacijo rezultatov in odgovorim na zastavljeni raziskovalni vprašanji.

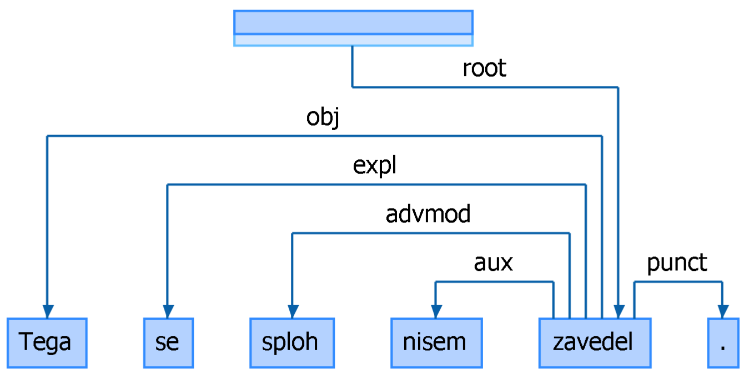
2 RAZISKOVALNI PODATKI IN METODE

2.1 Označevalna shema Universal Dependencies

Korpusa, ki ju v tej študiji uporabljam, vsebujeta oznake, ki sledijo mednarodno uveljavljeni shemi za skladenjsko označevanje Universal Dependencies (de Marneffe in sod., 2021). Označevanje po tej shemi sledi principom odvisnostne skladnje, kar pomeni, da vsaka poved tvori aciklični graf, sestavljen iz usmerjenih odvisnostnih relacij, ki potekajo od nadrejenih besed k podrejenim. Tak skupek razmerij med besedami imenujemo tudi skladenjsko drevo in ga lahko grafično ponazorimo, kot je prikazano na Sliki 1

Vsaka odvisnostna relacija ima po tem sistemu svoj tip (npr. *obj*, *expl*, *advmod*,...), nadrejeni element (oz. glavo) in podrejeni element (oz. cilj). Pri primeru na Sliki 1 na primer med besedama »zavedel« in »Tega« velja relacija *obj* (po slovenskih skladenjskih sistemih to ustreza predmetu, enemu od jedrnih argumentov povedka). Glagol »zavedel« je glava relacije, cilj pa je »Tega«. Vsaka poved mora vsebovati tudi relacijo *root*, katere glava je vedno prazni element, ki ga imenujemo tudi korenski element povedi, cilj pa je jedro povedi – običajno glavni del povedka glavnega stavka.

Slika 1: Primer skladenjskega drevesa po sistemu Universal Dependencies, kot ga zgenerira orodje Q-CAT (Brank, 2023).



2.2 Uporabljena korpusa

Prvi od korpusov, ki sem ju uporabil za primerjavo, je korpus SSJ-UD (Dobrovoljc in sod., 2017; Dobrovoljc in Ljubešić, 2022; Dobrovoljc in sod., 2023), ki vsebuje tako leposlovna kot tudi neleposlovna pisna besedila iz korpusov ssj500k (Krek in sod., 2020) in ELEXIS-WSD (Martelli in sod., 2023) in zato predstavlja uravnotežen vzorec pisne slovenščine. V tej študiji uporabljam različico 2.14[2], ki šteje približno 260.000 pojavnic oz. približno 13.000 povedi. Del oznak je rezultat avtomatske pretvorbe iz skladenjske označevalne sheme JOS-SYN (Erjavec in sod., 2010), preostanek pa je ročno pregledan.

Drugi korpus je SST-UD (Dobrovoljc in Nivre, 2016), ki vsebuje transkripcije govora, vzete iz referenčnega govornega korpusa Gos (Verdonik in sod., 2023). Korpus vsebuje transkripcije govora, ki so ga sproducirali govorci raznolikih ozadij in je nastal v različnih situacijah, zato ga obravnavam kot reprezentativen vzorec govorjene slovenščine. V svoji raziskavi uporabljam še neobjavljeno različico, ki za razliko od trenutno dostopne različice 2.14[3] vsebuje izboljšane oznake, usklajene z najnovejšimi dodatki v splošne označevalne smernice sistema Universal Dependencies. Ta novejša različica korpusa bo vključena v eno od prihodnjih skupnih izdaj, ki redno izhajajo v okviru projekta Universal Depen-

[2] https://github.com/UniversalDependencies/UD_Slovenian-SSJ
[3] https://github.com/UniversalDependencies/UD_Slovenian-SST

dencies. Po velikosti šteje približno 70.000 pojavnic oz. približno 6.000 povedi. Vse oznake v SST-UD so ročno pregledane.

2.3 Mere skladenjske kompleksnosti

Izbral sem šest mer, ki se v literaturi uporabljajo kot indikator skladenjske kompleksnosti. V grobem jih lahko razdelimo v dve skupini: prva, v kateri so tri mere, ki temeljijo na štetju in razmerjih med osnovnimi skladenjskimi elementi (število besed v povedi, število stavkov v povedi, povprečno število stavkov na T-enoto), in druga, v kateri so tri mere, ki so osnovane na principih odvisnostne skladnje in na sestavi skladenjskih dreves (povprečna dolžina odvisnostnih relacij, normalizirana dolžina odvisnostnih relacij, maksimalna globina skladenjskega drevesa). Za večjo primerljivost sem izbral izključno mere, ki podajo vrednosti za kompleksnost na ravni povedi. Vsaka od mer je podrobneje predstavljena v nadaljevanju.

2.3.1 ŠTEVILO BESED V POVEDI (B/P)

To mero bi lahko imenovali tudi enostavna dolžina povedi. Število besed v povedi (v nadaljevanju *B/P*) je ena najbolj razširjenih mer skladenjske kompleksnosti, ki se uporablja v številnih raziskavah in je pogosto predstavljena kot najosnovnejša (Jagaiah in sod., 2020; Wang, 2022; Lintunen in Mäkilä, 2014; Mylläri, 2020a). V svoji analizi kot »besede« upoštevam pojavnice – osnovne podenote, na katere se deli vsaka poved v obeh obravnavanih korpusih.

2.3.2 ŠTEVILO STAVKOV V POVEDI (S/P)

Število stavkov v povedi (v nadaljevanju *S/P*) je pogosto uporabljena mera, ki je osnovana na razumevanju stavka kot osnovne enote skladenjske strukture neke povedi. Ker v korpusnih raziskavah pogosto naletimo na razne fragmente in druge posebne jezikovne pojave, je lahko dokaj težavno na jasen način zastaviti enoznačno definicijo stavka. Konkretne zamejitve stavkov se lahko zato med seboj precej razlikujejo (Mylläri, 2020b). V tem prispevku se pri določanju mej stavkov opiram na različne tipe skladenjskih relacij, ki obstajajo znotraj sistema Universal Dependencies. Kot število stavkov v povedi tako upoštevam število besed, ki jih uvrščamo med glagole in so cilj ene od relacij *csubj* (osebkovi odvisniki), *ccomp* (predmetni odvisniki), *xcomp* (odprta stavčna dopolnila), *advcl* (prislovnodoločilni odvisniki), *acl* (prilastkovi odvisniki), *conj* (stavčna priredja), *parataxis* (stavčna soredja), oziroma če so glave vsaj ene relacije *cop* (vezni glagol).

2.3.3 POVPREČNO ŠTEVILO STAVKOV NA T-ENOTO (S/T)

Pri povprečnem številu stavkov na T-enoto (v nadaljevanju *S/T*) se sklicujem na pojem T-enote, ki se pogosto definira kot en neodvisen stavek skupaj z vsemi odvisnimi stavki, ki se vežejo nanj (Jagaiah in sod., 2020).[4] Ta mera v ospredje postavlja podredne strukture, ki razkrivajo drugačno plat skladenjske kompleksnosti kot priredne strukture (Mylläri, 2020b). Vrednost S/T za neko poved sem poračunal po sledeči formuli:

$$S/T = \frac{\text{število stavkov v povedi}}{\text{število T-enot v povedi}} \tag{1}$$

2.3.4 POVPREČNA DOLŽINA ODVISNOSTNIH RELACIJ (PDO)

V številnih študijah se je uveljavila mera povprečne dolžine vseh odvisnostnih relacij v neki povedi (angl. *Mean Dependency Distance*, v nadaljevanju *PDO*). Daljše razdalje med glavo in ciljem neke relacije se povezuje z oteženim procesiranjem jezika, zato se je ta mera pričela uporabljati kot pokazatelj skladenjske kompleksnosti (Lei in Jockers, 2020; Futrell in sod., 2015). Izračunamo jo po spodnji formuli:

$$PDO = \frac{\sum_{i=1}^{n} DO_i}{n} \tag{2}$$

Pri tem je *DO* (dolžina odvisnostne relacije, angl. *Dependency Distance*) za neko besedo absolutna vrednost razlike med indeksom (pozicijo v stavku) te besede in indeksom njenega nadrejenega elementa. V stavku »Oče je šel v trgovino« je torej DO za besedo »Oče« enak 2, ker je ta beseda cilj odvisnostne relacije, ki izvira iz glagola »šel«. DO je enak 2, ker je indeks te besede 1, indeks nadrejene besede (glagol »šel«) pa 3. *n* v zgornji formuli predstavlja skupno število besed

[4]V povedi *Tisti, ki razumejo, so tu, ostali pa so odšli* imamo torej dve T-enoti ((1) *Tisti, ki razumejo, so tu* in (2) *ostali pa so odšli*)

v neki povedi, *i* pa indeks besede. Relaciji *punct* in *root* sta v skladu s pogosto prakso izvzeti iz obravnave (Lei in Jockers, 2020).

2.3.5 NORMALIZIRANA DOLŽINA ODVISNOSTNIH RELACIJ (NDO)

Normalizirana dolžina odvisnostnih relacij (angl. *Normalized Dependency Distance*, v nadaljevanju *NDO*) je podobna mera kot PDO, le da se pri tej v izračunu upošteva še dolžino povedi in pozicijo korena v povedi (Lei in Jockers, 2020). Poračuna se jo po sledeči formuli:

$$NDO = \text{abs}\left(\ln\left(\frac{PDO}{\sqrt{\text{indeksKorena} \times \text{n}}}\right)\right) \tag{3}$$

Pri tem *PDO* predstavlja povprečno dolžino odvisnostnih relacij, *abs* absolutno vrednost, *ln* naravni logaritem, *indeksKorena* je indeks besede, ki je cilj relacije *root*, *n* pa je skupno število besed v povedi.

2.3.6 MAKSIMALNA GLOBINA SKLADENJSKEGA DREVESA (MGD)

Maksimalna globina skladenjskega drevesa (v nadaljevanju *MGD*) je mera, ki nam pove, kaj je največja globina skladenjskega drevesa v neki povedi. Izračunamo jo tako, da začnemo pri neki besedi in se pomaknemo do izvora relacije, ki ima to besedo za cilj. To ponavljamo dokler ne dosežemo korenskega elementa povedi. Število premikov, ki smo jih morali narediti, da smo dosegli korenski element, predstavlja globino skladenjskega drevesa za to besedo. Globine skladenjskih dreves izračunamo za vse besede v povedi in jih primerjamo. MGD je najvišja dobljena vrednost v povedi. V literaturi se ta mera uporablja nekoliko manj pogosteje za merjenje skladenjske kompleksnosti kot ostale predstavljene mere. Xu in Reitter (2016) ugotavljata, da MGD v angleških povedih v veliki meri korelira z dolžino povedi (B/P).

2.4 Potek statistične analize

V sledečem razdelku na podlagi poračunanih vrednosti za vsako od šestih zgoraj opisanih mer najprej poročam osnovne vrednosti opisne statistike (povprečje in standardni odklon) in rezultate ponazorim s pomočjo grafikonov kvartilov.

Po pregledu histogramov frekvenčne porazdelitve in t. i. grafov Q-Q (angl. *Q-Q plot*) sem ugotovil, da podatki niso normalno porazdeljeni, zato za preverjanje statistične pomembnosti v raziskavi uporabljam neparametrični test Mann-Whitney U. Za korigiranje pri številnih primerjavah za statistično značilne razlike uporabim metodo Holm-Bonferroni. Za merjenje velikosti efekta uporabljam biserialni korelacijski koeficient.

Da bi raziskal raven ujemanja različnih mer skladenjske kompleksnosti med seboj, poračunam tudi Pearsonov korelacijski koeficient za vsak par mer kompleksnosti. Rezultate prikažem v obliki korelacijske matrike. Tudi tu preverim statistično pomembnost rezultatov in uporabim metodo Holm-Bonferroni za korigiranje.

Ker je korpus SSJ-UD po številu pojavnic precej večji kot korpus SST-UD, sem analizo izvedel hkrati še z manjšim vzorcem naključno izbranih povedi iz korpusa SSJ-UD in celotnim korpusom SST-UD. Število povedi za ta manjši vzorec je bilo enako številu v SST-UD (6.104 povedi), tako da sta bila vzorca po velikosti primerljiva. Vse izvedene analize in statistični testi so pokazali povsem enake tendence kot pri analizi s korpusom SSJ-UD v polni velikosti, zato v naslednjem razdelku poročam le rezultate primerjave obeh korpusov v polnem obsegu.

Za izvedbo statističnih analiz sem uporabil program Jamovi (The jamovi project, 2024), za izris grafov in korelacijske matrike pa Pythonovo knjižnico Seaborn (Waskom, 2021).

3 REZULTATI

3.1 Opisna statistika

Za vsako od mer v Tabeli 1 prikazujem primerjavo aritmetične sredine in standardnega odklona med obema korpusoma. Kaže se jasen vzorec: pri vseh merah skladenjske kompleksnosti je aritmetična sredina višja v pisnem korpusu SSJ-UD kot v govornem SST-UD. Slika 2 prikazuje primerjavo grafikonov kvartilov obeh korpusov za vsako mero. Ker se je izkazalo, da je v podatkih pri večini mer zelo veliko osamelcev, sem se odločil za postavitev končnih ročajev grafa na minimalno in maksimalno vrednost. Pri večini mer je frekvenčna porazdelitev zgoščena pri nižjih vrednostih, hkrati pa je prisotnih tudi veliko višjih vrednosti s postopoma padajočo frekvenco. Ta pojav imenujemo tudi »dolgi rep« (angl. *long tail*) in ga pri delu z raznimi jezikovnimi podatki pogosto srečamo (Ryland Williams in sod., 2015). Drugačno frekvenčno porazdelitev opazimo predvsem pri meri NDO, ki pa se vseeno ne ujema povsem z normalno porazdelitvijo. Pri tej meri je večina vrednosti zgoščena bolj proti sredini variacijskega razmika kot pri ostalih.

Tabela 1: Pregled aritmetične sredine in standardnega odklona za vse mere kompleksnosti.

		B/P	S/P	S/T	PDO	NDO	MGD
SSJ	Aritm. sredina	19,9	2,28	1,59	2,57	1,14	5,09
	Stand. odklon	12,8	1,45	0,80	0,92	0,51	1,81
SST	Aritm. sredina	12,5	2,11	1,34	2,31	0,90	3,70
	Stand. odklon	14,5	1,95	0,63	0,89	0,48	2,13

3.2 Statistični testi

Tabela 2 prikazuje rezultate statističnega testa Mann-Whitney U za preverjanje razlik med korpusoma SSJ-UD in SST-UD za vsako od obravnavanih mer kompleksnosti. Rezultati kažejo, da so p-vrednosti za vse mere manjše od 0,01. Tudi po apliciranju metode Holm-Bonferroni so razlike med korpusoma pri vseh merah statistično pomembne. Velikosti efekta so večinoma nizke (< 0,3) z izjemo mer B/P in MGD, ki dosegata nekoliko višje vrednosti (> 0,4). To nakazuje, da so pri teh dveh merah razlike med korpusoma precej bolj očitne kot pri ostalih. V splošnem so nizke velikosti efekta pri ostalih merah lahko tudi posledica razmeroma majhne velikosti obravnavanih korpusov. Ena od možnih razlag je, da so te mere v prvi vrsti povezane s skladenjskimi pojavi, ki so v jeziku redkejši, zato pri manjši količini podatkov te razlike ne pridejo do izraza v tako veliki meri.

3.3 Korelacija med merami

Poleg preverjanja razlik med obema korpusoma sem preveril še stopnjo korelacije med različnimi merami kompleksnosti. Slika 3 prikazuje korelacijsko matriko za vse pare mer kompleksnosti. Vse obravnavane mere vsaj do neke mere pozitivno korelirajo, pri čemer je precej visoka korelacija med merama B/P in MGD ter B/P in S/P (> 0,7), nekoliko nižja, a vseeno znatna, pa je med MGD in S/P ter B/P in PDO (> 0,6). Najnižjo stopnjo korelacije dosega par NDO

Slika 2: Grafikoni kvartilov za primerjavo obeh korpusov pri vsaki meri.

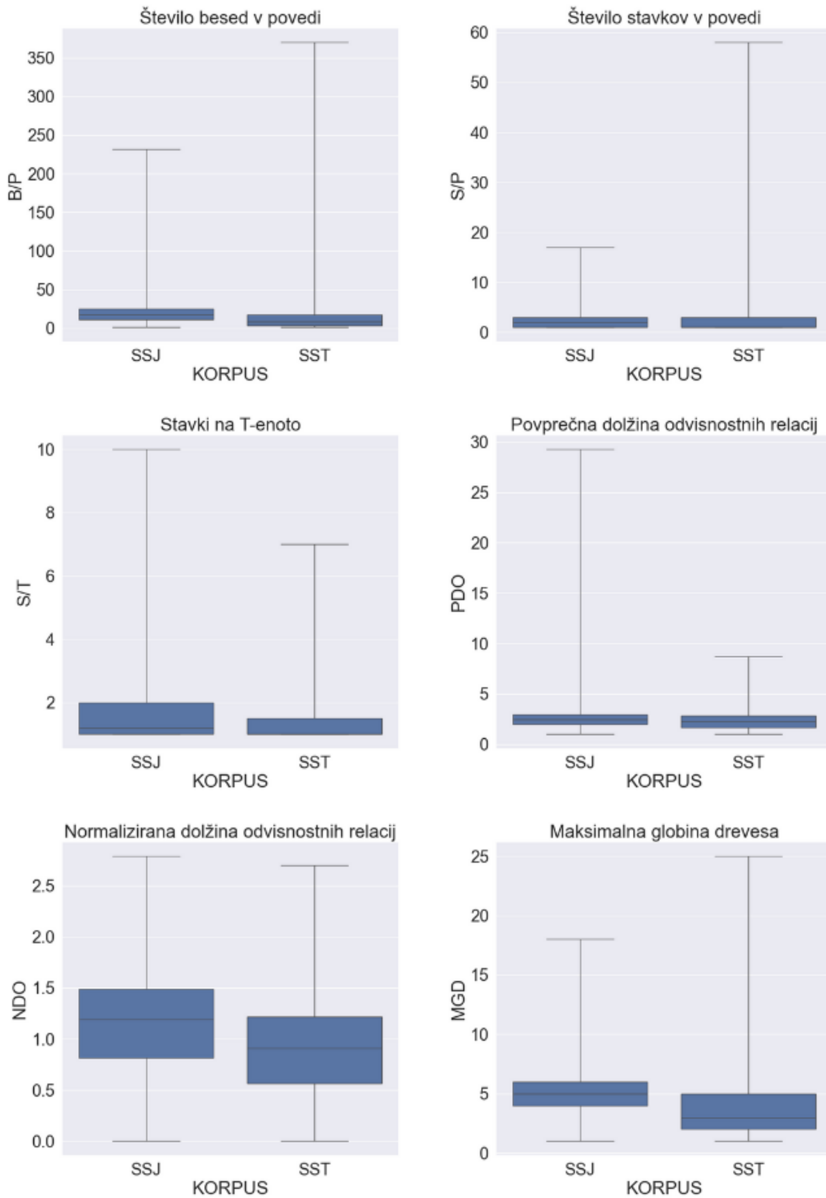
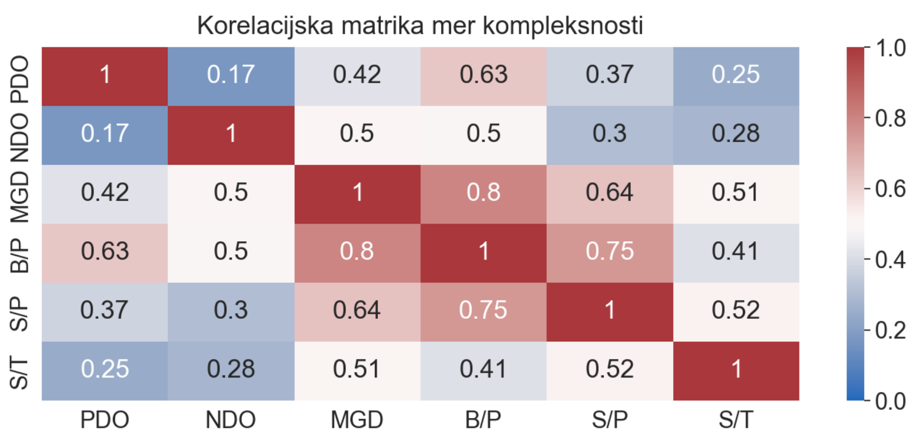


Tabela 2: P-vrednosti za statistični test Mann-Whitney U in velikost efekta (biserialni korelacijski koeficient) za vsako mero skladenjske kompleksnosti.

Mera kompleksnosti	p	Velikost efekta
B/P	<0,01	0,45
S/P	<0,01	0,15
S/T	<0,01	0,19
PDO	<0,01	0,17
NDO	<0,01	0,29
MGD	<0,01	0,43

Slika 3: Korelacijska matrika, ki prikazuje Pearsonov korelacijski koeficient za vsak par mer skladenjske kompleksnosti. Večja mera korelacije je prikazana z rdečo barvo, manjša pa z modro.



in PDO (< 0,2), kar ni presenetljivo, saj je bila mera NDO ustvarjena z namenom zmanjšanja vpliva dolžine povedi in pozicije korena na vrednosti, ki jih vrne mera PDO. Očitno je vpliv teh dveh dejavnikov na PDO tako velik, da je med merama še najmanj linearne korelacije. Največjo skupno korelacijo z ostalimi merami dosega B/P, najmanjšo pa NDO. P-vrednosti za vse pare obravnavanih mer so manjše od 0,01.

4 DISKUSIJA

Vse obravnavane mere kompleksnosti imajo v povprečju višje vrednosti v korpusu SSJ-UD kot v SST-UD. To nakazuje, da je skladenjska sestava pisnega jezika nekoliko kompleksnejša kot pri govorjenem jeziku, vsaj kar se tiče aspektov skladenjske kompleksnosti, ki jih pokriva izbranih šest mer. Višjo razpršenost vrednosti pri nekaterih merah – B/P, S/P in MGD – v korpusu SST-UD v primerjavi s SSJ-UD lahko pripišemo večjemu variacijskemu razmiku za dolžine povedi v govornem korpusu. To namreč odraža zapletenost segmentacije govorjenega jezika, ki jo izpostavljata že Dobrovoljc in Nivre (2016) v opisu govornega korpusa.

Analiza korelacijske matrike razkriva, da, četudi vse mere nakazujejo na precejšnje razlike v skladenjski kompleksnosti med govornim in pisnim korpusom, korelacija med samimi merami ni vedno visoka. To pomeni, da nekatere od mer zaznavajo precej drugačen vidik skladenjske kompleksnosti kot druge. Največjo korelacijo dosega par B/P in MGD, kar se ujema z ugotovitvijo Xu in Reitter (2016), ki opažata, da ti dve vrednosti v jeziku po navadi izkazujeta veliko mero linearne korelacije. Nasploh se zdi, da lahko mere razdelimo v dve skupini: tiste, ki izkazujejo precejšno mero linearne povezanosti s preprosto dolžino povedi B/P (S/P, MGD in PDO), in tiste, ki imajo z dolžino povedi manj linearne povezanosti (S/T in NDO). Meri S/T in NDO nam torej razkrivata nek aspekt skladenjske kompleksnosti, ki je še najmanj odvisen od preproste dolžine povedi, in nam pove več o notranji skladenjski sestavi neke povedi.

To razliko med obema vidikoma skladenjske kompleksnosti lahko ponazorimo s spodnjima dvema povedma iz korpusa SSJ-UD:

(a) *Spomnimo se samo pred leti prenosa žiro računa elektrogospodarstva v Ljubljano, spomnimo se "svežega" sedeža holdinga elektrogospodarstva, v to kategorijo pa spada tudi privatizacija Nove KBM.*

(b) *Kljub majskemu odprtju meja trga tisočerih priložnosti velikega povpraševanja delodajalcev nismo doživeli.*

Pri primeru (a) je vrednost B/P precej visoka (31) glede na povprečje v tem korpusu (19,9), vrednost za NDO pa je precej nizka (0,42) glede na povprečje

(1,14). Prav nasprotno je pri primeru (b), kjer je B/P nižji (13), NDO pa je višji (1,70). Primer (a) torej bolj odraža osnovnejši vidik skladenjske kompleksnosti, ki se v glavnem povezuje z dolžino povedi. V povedi (a) je razvidno, da tak jezik pogosto vsebuje veliko priredij in enostavnih sosledij zaporedno vezanih stavkov. Po drugi strani pa je (b) primer povedi, v kateri se bolj kaže zapletena notranja sestava skladenjskih elementov. Primer (b) predvsem zaznamuje dolg uvajalni element na začetku povedi (*Kljub majskemu odprtju...*), ki sproducira precej visoko dolžino odvisnostne relacije. Posledično ob prvem branju povedi bralec precej težko razbere, kje se začne naslednji skladenjski element in kako se skladenjska struktura povedi nadaljuje.

Ena od pomembnih razlik med obema vidikoma skladenjske kompleksnosti je tudi to, da je mera B/P že po svoji definiciji močno odvisna od prvotne segmentacije besedil v korpusu na povedi, kar posledično pomeni, da so od segmentacije precej odvisne tudi vrednosti MGD, S/P in PDO, ki izkazujejo visoko korelacijo z B/P. V nasprotju sta NDO in B/P, ki ne izkazujeta tako močne linearne povezanosti z B/P, precej manj odvisni od prvotne segmentacije. Povsem verjetno je, da so razlike med korpusi ob uporabi te prve skupine mer predvsem posledica različnih načel segmentacije ob nastanku obeh korpusov in le v manjši meri razkrivajo razlike v notranji skladenjski strukturi povedi. Če se vrnemo k Tabeli 2, lahko razberemo, da je velikost efekta najvišja prav pri B/P in MGD, kar lahko nakazuje na različna načela segmentacije v obeh obravnavanih virih. Tudi prej že omenjen višji variacijski razmik za dolžine povedi v SST-UD govori v prid različnim principom segmentacije v obeh korpusih.

Četudi sem zgoraj uporabljene mere skladenjske kompleksnosti razvrstil v dve grobi skupini, pa to ne pomeni, da razlik znotraj teh skupin ni. NDO in S/T, ki sta edini s precej manjšo korelacijo z mero B/P, hkrati med seboj linearno nista tako močno povezani (r = 0,28). Očitno torej vsaka od obravnavanih mer poda neko svojo posebno informacijo o skladenjski strukturi neke povedi. Skladenjsko kompleksnost moramo tako razumeti kot večplasten koncept, ki se ga ne da povzeti le z eno pravo mero, pač pa nam različni načini merjenja lahko povejo precej različne stvari.

5 ZAKLJUČEK

V študiji na podlagi analize skladenjsko razčlenjenih korpusov odkrivam, da je pisni jezik v slovenščini z vidika skladenjske sestave kompleksnejši kot govorjeni jezik. Vendar pa se ta razlika ne kaže le v enem vidiku skladenjske kompleksnosti, pač pa v večih.

Iz rezultatov je razvidno, da so razlike najbolj očitne pri dolžini povedi in pri globini skladenjskih dreves. Odkrivam tudi, da je med uporabljenimi merami veliko razlik, kar potrjuje opažanje številnih študij (Ehret in sod., 2023; Bentz in sod., 2023; Berdicevskis in sod., 2018; Jagaiah in sod., 2020), ki trdijo, da skladenjska kompleksnosti ni enoten pojem, ki bi se ga dalo povzeti oz. izmeriti z eno samo vseobsegajočo mero kompleksnosti, pač pa gre za večplasten koncept, ki ga lahko v jeziku analiziramo na več različnih načinov.

Rezultati kažejo, da se nekatere uporabljene mere skladenjske kompleksnosti v veliki meri ujemajo z dolžino povedi, so pa tudi take, ki tega ujemanja ne izkazujejo in nam zato razkrivajo nek drug vidik skladenjske kompleksnosti. Pri raziskavah skladenjske kompleksnosti je torej potreben poseben premislek o tem, kaj točno želimo izmeriti. Pri raziskavah, ki se osredotočajo na analizo uporabe prirednih struktur, je na primer smiselno osredotočanje na mere, kot je število stavkov v povedi. Če pa bi želeli preučevati skladenjske vzorce, ki ljudem najbolj otežujejo razumevanje besedil, pa bi morali v razmislek vključiti tudi druge mere. Martínez in sod. (2022) na primer ugotavljajo, zakaj pravna besedila v angleščini bralcem pogosto povzročajo težave v razumevanju v primerjavi z ostalimi vrstami besedil. Pri tej analizi uporabijo mere kompleksnosti, ki temeljijo na merjenju dolžine odvisnostnih relacij.

Doprinos te študije je večplasten: (1) v prispevku sem predstavil novo metodo za avtomatsko primerjanje razlik v skladenjski kompleksnosti med različnimi skladenjsko razčlenjenimi korpusi z uporabo šestih različnih mer, (2) preučil sem razlike v skladenjski kompleksnosti med reprezentativnima korpusoma govorjene in pisne slovenščine, (3) preučil sem razlike med šestimi uporabljenimi merami, ki se v literaturi uporabljajo kot merilo skladenjske kompleksnosti.

Pri tem je treba opozoriti na pomembno vlogo skladenjsko razčlenjenih korpusov s kvalitetnimi oznakami po principu odvisnostne skladnje. Takšni korpusi nam namreč omogočajo uporabo mer, kot so PDO, NDO in MGD, ki sem jih uporabil tudi jaz v tej raziskavi. Stalno izboljševanje, vzdrževanje in nadgrajevanje tovrstnih skladenjsko razčlenjenih virov je torej izjemnega pomena za jezikoslovne raziskave v slovenščini.

Seveda moja študija ni brez pomanjkljivosti. Čeprav sta obravnavana korpusa v primerjavi s številnimi drugimi, ki so vključeni v zbirko Universal Dependencies, kar obsežna, se obenem ne moreta primerjati z velikostjo nekaterih drugih skladenjsko označenih korpusov, kot je npr. PDT (Hajič in sod., 2020). Pri preučevanju skladenjskih pojavov je še posebej pomembna količina preučenih podatkov, zato bo v prihodnosti, ko bo na voljo večja količina podatkov s skladenjskimi oznakami, potrebna ponovna preverba tokratnih ugotovitev. V študiji tudi izhajam iz predpostavke, da je stavčna segmentacija v pisnem korpusu primerljiva s segmentacijo v govornem korpusu, kar pa lahko v določenih primerih vodi do velikih razlik pri rezultatih, ki jih vrnejo določene mere kompleksnosti, posebej pri številu besed na poved. Vsekakor problem določanja ustrezne segmentacije v govornih korpusih ni enostavno rešljiv (Dobrovoljc in Nivre, 2016), zato sem se odločil obdržati izvorno označenost v vseh virih in tako obdržati preprosto zasnovo raziskave.

Rezultati te študije hkrati odpirajo tudi veliko vprašanj za nadaljnje študije. Podobno raziskavo bi bilo zanimivo izvesti z uporabo tujih skladenjsko označenih korpusov in ugotovitve primerjati s temi, ki sem jih predstavil v tem prispevku. V svoji raziskavi sem različne mere skladenjske kompleksnosti primerjal predvsem z vidika linearne korelacije, naslednji korak pa je še natančneje raziskati, na kakšen način se mere kompleksnosti med seboj razlikujejo. Vredno bi bilo na primer na bolj sistematičen in razširjen način identificirati skupine primerov, ki dosegajo visoke vrednosti pri vsaki od mer. Prav tako bi bilo zanimivo podobno raziskavo ponoviti še z drugimi merami, ki se v literaturi povezujejo s konceptom skladenjske kompleksnosti in jih v tem prispevku ne uporabljam.

6 DOSTOPNOST PODATKOV

Vse Pythonove skripte, s pomočjo katerih sem naredil izvoze in poračunal vrednosti za mere, so dostopne prek GitHub repozitorija.[5]

[5] https://github.com/lukatercon/SyntComplex

ZAHVALA

Delo, predstavljeno v tem prispevku sta financirala program mladi raziskovalci Javne agencije za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije ter projekt SPOT – Na drevesnici temelječ pristop k raziskavam govorjene slovenščine, ARIS št. Z6-4617.

LITERATURA

Bentz, C., Gutierrez-Vasques, X., Sozinova, O. in Samardžić, T. (2023). Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. *Linguistics Vanguard*, *9*(s1), 9–25. Pridobljeno 2024-05-31, https://doi.org/10.1515/lingvan-2021-0054 doi: 10.1515/lingvan-2021-0054

Berdicevskis, A., Çöltekin, Ç., Ehret, K., von Prince, K., Ross, D., Thompson, B., ... Bentz, C. (2018, November). Using Universal Dependencies in cross-linguistic complexity research. V M.-C. de Marneffe, T. Lynn in S. Schuster (Ur.), *Proceedings of the second workshop on universal dependencies (UDW 2018)* (str. 8–17). Brussels, Belgium: Association for Computational Linguistics. https://aclanthology.org/W18-6002 doi: 10.18653/v1/W18-6002

Brank, J. (2023). *Q-CAT corpus annotation tool 1.5.* http://hdl.handle.net/11356/1844 (Slovenian language resource repository CLARIN.SI)

de Marneffe, M.-C., Manning, C. D., Nivre, J. in Zeman, D. (2021, June). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://aclanthology.org/2021.cl-2.11 doi: 10.1162/coli_a_00402

Dobrovoljc, K., Erjavec, T. in Krek, S. (2017, April). The Universal Dependencies treebank for Slovenian. V T. Erjavec, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger in R. Yangarber (Ur.), *Proceedings of the 6th workshop on Balto-Slavic natural language processing* (str. 33–38). Valencia, Spain: Association for Computational Linguistics. https://aclanthology.org/W17-1406 doi: 10.18653/v1/W17-1406

Dobrovoljc, K. in Ljubešić, N. (2022, June). Extending the SSJ Universal Dependencies treebank for Slovenian: Was it worth it? V S. Pradhan in S. Kuebler (Ur.), *Proceedings of the 16th linguistic annotation workshop (law-xvi) within lrec2022* (str. 15–22). Marseille, France: European Language Resources Association. https://aclanthology.org/2022.law-1.3

Dobrovoljc, K. in Nivre, J. (2016, May). The Universal Dependencies treebank of spoken Slovenian. V N. Calzolari in sod. (Ur.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (str. 1566–1573).

Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1248

Dobrovoljc, K., Terčon, L. in Ljubešić, N. (2023, Sep.). Universal dependencies za slovenščino: Nove smernice, ročno označeni podatki in razčlenjevalni model. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, *11*(1), 218–246. https://journals.uni-lj.si/slovenscina2/article/view/12031 doi: 10.4312/slo2.0.2023.1.218-246

Ehret, K., Berdicevskis, A., Bentz, C. in Blumenthal-Dramé, A. (2023). Measuring language complexity: challenges and opportunities. *Linguistics Vanguard*, *9*(s1), 1–8. Pridobljeno 2024-05-31, https://doi.org/10.1515/lingvan-2022-0133 doi: 10.1515/lingvan-2022-0133

Erjavec, T., Fišer, D., Krek, S. in Ledinek, N. (2010, May). The JOS linguistically tagged corpus of Slovene. V N. Calzolari in sod. (Ur.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf

Futrell, R., Mahowald, K. in Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

Gaberšček, N. (2018). Merjenje določenih vidikov skladenjske kompleksnosti v pisnih besedilih slovenskih osnovnošolcev. *Jezik in slovstvo*, *63*(2/3).

Hajič, J., Bejček, E., Hlavacova, J., Mikulová, M., Straka, M., Štěpánek, J. in Štěpánková, B. (2020, May). Prague dependency treebank - consolidated 1.0. V N. Calzolari in sod. (Ur.), *Proceedings of the twelfth language resources and evaluation conference* (str. 5208–5218). Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.641

Jagaiah, T., Olinghouse, N. G. in Kearns, D. M. (2020, Dec 01). Syntactic complexity measures: variation by genre, grade-level, students' writing abilities, and writing quality. *Reading and Writing*, *33*(10), 2577-2638. https://doi.org/10.1007/s11145-020-10057-x doi: 10.1007/s11145-020-10057-x

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J. in Brank, J. (2020). The ssj500k training corpus for slovene language processing. *Jezikovne Tehnologije in Digitalna Humanistika*, 24–33.

Lei, L. in Jockers, M. L. (2020). Normalized dependency distance: Proposing a new measure. *Journal of Quantitative Linguistics*, *27*(1), 62–79. https://doi.org/10.1080/09296174.2018.1504615 doi: 10.1080/09296174.2018.1504615

Lintunen, P. in Mäkilä, M. (2014, 12). Measuring syntactic complexity in spoken and written learner language: Comparing the incomparable? *Research in Language*, *12*. doi: 10.1515/rela-2015-0005

Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., ... Kolkovska, S. (2023). *Parallel sense-annotated corpus ELEXIS-WSD 1.1.* http://hdl.handle.net/11356/1842 (Slovenian language resource repository CLARIN.SI)

Martinc, M., Pollak, S. in Robnik-Šikonja, M. (2021, 04). Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, *47*(1), 141-179. https://doi.org/10.1162/coli_a_00398 doi: 10.1162/coli_a_00398

Martínez, E., Mollica, F. in Gibson, E. (2022). Poor writing, not specialized concepts, drives processing difficulty in legal language. *Cognition*, *224*, 105070. https://www.sciencedirect.com/science/article/pii/S0010027722000580 doi: https://doi.org/10.1016/j.cognition.2022.105070

Mylläri, T. (2020a). Measuring syntactic complexity in learner finnish. *Apples: Journal of Applied Language Studies*, *14*(2).

Mylläri, T. (2020b). Words, clauses, sentences, and t-units in learner language: Precise and objective units of measure? *Journal of the European Second Language Association*, *4*(1).

Ryland Williams, J., Lessard, P. R., Desu, S., Clark, E. M., Bagrow, J. P., Danforth, C. M. in Sheridan Dodds, P. (2015). Zipf's law holds for phrases, not words. *Scientific reports*, *5*(1).

The jamovi project. (2024). *jamovi (version 2.5).* https://www.jamovi.org (Retrieved from https://www.jamovi.org)

Verdonik, D., Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., ... Rupnik, P. (2023). *Spoken corpus gos 2.1 (transcriptions).* http://hdl.handle.net/11356/1863 (Slovenian language resource repository CLARIN.SI)

Wang, Z. (2022). Dynamic development of syntactic complexity in second language writing: A longitudinal case study of a young chinese efl learner. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.974481 doi: 10.3389/fpsyg.2022.974481

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021 doi: 10.21105/joss.03021

Xu, Y. in Reitter, D. (2016). Convergence of syntactic complexity in conversation. V *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (str. 443–448).

THE USE OF SIX SYNTACTIC COMPLEXITY MEASURES FOR LINGUISTIC COMPARISONS BETWEEN A SPOKEN AND A WRITTEN CORPUS

There are a number of methods for measuring syntactic complexity in digital language databases. Linguistic corpora, especially those containing syntactic annotations, enable researchers to automatically and efficiently conduct analyses and comparisons of syntactic complexity. In this paper, I present a method with which I automatically compare two corpora – one containing written texts and the other containing spoken texts – using six established measures of syntactic complexity. The results of this comparison indicate that the syntactic makeup of the language contained in the written corpus is slightly more complex than in the spoken corpus. The differences are most pronounced in sentence length and in syntactic tree depth. Additionally, an analysis of the correlation between the different measures suggests that some provide quite different information about the syntactic structure of a sentence compared to others.

Keywords: syntactic complexity, written corpus, spoken corpus, complexity measures, student paper