Zbornik konference

Jezikovne tehnologije in digitalna humanistika

Proceedings of the Conference on

Language Technologies and Digital Humanities

15.– 16. september 2022

Ljubljana, Slovenija

September 15h – 16th 2022

Ljubljana, Slovenia

Uredila / Edited by:

Darja Fišer, Tomaž Erjavec

ZBORNIK KONFERENCE
JEZIKOVNE TEHNOLOGIJE IN DIGITALNA HUMANISTIKA

PROCEEDINGS OF THE CONFERENCE ON
LANGUAGE TECHNOLOGIES & DIGITAL HUMANITIES

Uredila / *Edited by*: Darja Fišer, Tomaž Erjavec
Tehnični uredniki / *Technical editors*: Jakob Lenardič, Katja Meden, 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, 2022
First edition

Spletno mesto konference / Conference website:
https://www.sdjt.si/jtdh-2022 / https://www.sdjt.si/jtdh-2022/en

Publikacija je brezplačno dostopna na: / *Publication is available free of charge at*:
https://nl.ijs.si/jtdh22/proceedings-sl.html / https://nl.ijs.si/jtdh22/proceedings-en.html



To delo je objavljeno pod licenco Creative Commons Priznanje avtorstva 4.0 Mednarodna.

This work is licensed under a Creative Commons Attribution 4.0 International License.

CIP - Kataložni zapis o publikaciji

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

COBISS.SI-ID 121176323

ISBN 978-961-7104-20-2 (PDF)

Predgovor k zborniku konference "Jezikovne tehnologije in digitalna humanistika"

Slovensko društvo za jezikovne tehnologije, skupaj z Inštitutom za novejšo zgodovino in Centrom za jezikovne vire in tehnologije Univerze v Ljubljani ter raziskovalnima infrastrukturama CLARIN.SI in DARIAH-SI, že četrtič po vrsti prirejajo konferenco "Jezikovne tehnologije in digitalna humanistika", po uspešni programski širitvi konference Jezikovne tehnologije, ki se je odvijala od 1998, na digitalno humanistiko leta 2016 ohranja povezovalni fokus med disciplinama, hkrati pa si prizadeva postati pomembno srečevališče raziskovalcev v regiji.

Letošnja konferenca je potekala na Fakulteti za družbene vede Univerze v Ljubljani. Ker smo želeli zagotoviti, da bi bila konferenca v čim večji meri dostopna vsem zainteresiranim, smo vabljeni predavanji in vse predstavitve posneli in po zaključku konference objavili na konferenčni spletni strani. Na spletni strani konference pa je bil že vnaprej objavljen tudi zbornik konference.

Konferenčne vsebine smo razvrstili v tri dni. Prvi dan je bil posvečen predkonferenčnima seminarjema na temo tematskega modeliranja parlamentarnih razprav in raziskovalne infrastrukture CLARIN.SI. Drugi in tretji dan pa so se zvrstile predstavitve vabljenih predavateljev in avtorjev sprejetih prispevkov. Ker je bila zasedba na konferenci mednarodna, smo program izvedli v ločenih slovenskih in angleških sekcijah. Zvrstili sta se tako slovenska kot angleška študentska sekcija, dve slovenski in tri angleške redne sekcije ter angleška in slovenska poster sekcija, tako za redne, kot za študentske prispevke. Ob zaključku konference smo nagradili najboljši študentski prispevek. V posebni sekciji so bili predstavljeni še dosedanji rezultati projekta Razvoj slovenščine v digitalnem okolju, po konferenci pa je sledil še redni letni občni zbor Slovenskega društva za jezikovne tehnologije.

Na letošnji konferenci sta se predstavila dva vabljena predavatelja ter avtorji 30 rednih prispevkov, 9 razširjenih povzetkov in 12 študentskih prispevkov. Vse prispevke so pregledali trije recenzenti. 20 prispevkov je napisanih v slovenskem, 31 pa v angleškem jeziku. Skupno število vseh avtorjev prispevkov je 120, od katerih je skoraj tretjina tujih (iz Avstralije, Bosne in Hercegovine, Brazilije, Bolgarije, Hrvaške, Finske, Francije, Italije, Luksemburga, Severne Makedonije in Srbije).

Urednika se najlepše zahvaljujeva vsem, ki so prispevali k uspehu konference: vabljenima predavateljema in avtorjem prispevkov za skrbno pripravljene prispevke, predstavitve in plakate, programskemu odboru za natančno recenzentsko delo, organizacijskemu odboru za izvedbo konference, moderatorjem diskusij, tehničnim urednikom za pripravo spletnega zbornika in raziskovalnima infrastrukturama DARIAH-SI in CLARIN.SI ter društvu SDJT za finančno podporo konference.

Ljubljana, september 2022

Darja Fišer in Tomaž Erjavec

Preface to the Proceedings of the Conference "Language Technologies and Digital Humanities"

The Slovenian Language Technologies Society, together with the Institute of Contemporary History, the Centre of Language Resources and Technologies at the University of Ljubljana, and the research infrastructures CLARIN.SI and DARIAH-SI has organised the 13th Conference on Language Technologies and Digital Humanities. After its successful expansion to Digital Humanities in 2016, the conference retains its focus on the integration of the two disciplines and at the same time aims to position itself as an important meeting hub for fellow researchers in the region.

This year's conference took place at the Institute of Contemporary History in Ljubljana. In order to make the conference as accessible as possible to all participants, we made recordings of the invited talks and the presentations. After the conference, we published the recordings on the conference webpage, while the proceedings were made available on the webpage in advance.

The conference took place over the course of three days. On the first day, two pre-conference seminars were organised, one on topic modelling of parliamentary debates and another the CLARIN.SI research infrastructure. Days two and three were dedicated to two invited talks and presentations of accepted papers. Since the conference was also attended by international scholars, the programme was divided into separate Slovenian and English sessions. There was a Slovenian and an English student session, two Slovenian and three English regular sessions, as well as an English and a Slovenian poster section both for regular and student contributions. In a special session, the results of the project *Development of Slovene in a Digital Environment – Language Resources and Technologies* were presented.

This year's conference saw presentations from two invited speakers and from the authors of 30 regular papers, 9 extended abstracts, and 12 student papers. All the papers were reviewed by three reviewers. 20 papers were written in Slovene and 31 in English. The total number of all authors of the accepted papers is 120, a third of which were from abroad (from Australia, Bosnia and Herzegovina, Brazil, Bulgaria, Croatia, Finland, France, Italy, Luxemburg, North Macedonia, Serbia).

The editors would like to thank everyone who has contributed to the success of this conference: the invited lecturers and the authors of the papers for inspiring contributions and recordings of their lectures, the programme committee for their detailed reviews, the organising committee for enabling the conference to be held virtually, the discussion moderators, the technical editors for preparing the online proceedings and the research infrastructures DARIAH-SI and CLARIN.SI as well as the SDJT society for financially supporting the conference.

Ljubljana, September 2022

Darja Fišer and Tomaž Erjavec

Programski odbor / *Programme committee*

Predsedstvo programskega odbora / *Steering committee*

Darja Fišer, predsednica / *Chair*
Filozofska fakulteta, Univerza v Ljubljani in Inštitut za novejšo zgodovino / *Faculty of Arts, University of Ljubljana and Institute of Contemporary History*

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

Tomaž Erjavec
Institut "Jožef Stefan" / *Jožef Stefan Institute*

Andrej Pančur
Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Matej Klemen, študentska sekcija / *student section*
Fakulteta za računalništvo in informatiko / *Faculty for computer science and informatics, University of Ljubljana*

Aleš Žagar, študentska sekcija / *student section*
Fakulteta za računalništvo in informatiko / *Faculty for computer science and informatics, University of Ljubljana*

Člani programskega odbora in recenzenti / *Programme committee members and reviewers*

Špela Arhar Holdt
Filozofska fakulteta, Univerza v Ljubljani / *Faculty of Arts, University of Ljubljana*

Petra Bago
Filozofska fakulteta, Univerza v Zagrebu / *Faculty of Arts, University of Zagreb*

Vuk Batanović
Fakulteta za elektrotehniko, Univerza v Beogradu / *Faculty of Electrical Engineering, University of Belgrade*

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

Narvika Bovcon
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer 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
Filozofska fakulteta, Univerza v Ljubljani / *Faculty of Arts, 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 / *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 / *Faculty of Arts, University of Ljubljana*

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

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

Miran Hladnik
Filozofska fakulteta, Univerza v Ljubljani / *Faculty of Arts, University of Ljubljana*

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 Science, University of Ljubljana*

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

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

Jakob Lenardič
Filozofska fakulteta, Univerza v Ljubljani / *Faculty of Arts, University of Ljubljana*

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 Information Science, University of Ljubljana*

Sanda Martinčić Ipšić
Univerza na Reki / *University of Rijeka*

Maja Miličević Petrović
Univerza v Bolonji / *University of Bologna*

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

Matija Ogrin
Inštitut za slovensko literaturo in literarne vede ZRC SAZU / *Institute of Slovenian Literature and Literary Sciences, ZRC SAZU*

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

Dan Podjed
Inštitut za slovensko narodopisje ZRC SAZU / *Institute of Slovenian Ethnology, ZRC SAZU*

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

Ajda Pretnar Žagar
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer Science, University of Ljubljana*

Marko Robnik-Šikonja
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani / *Faculty of Computer 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*

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

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

Branislava Šandrih Todorović
Filološka fakulteta, Univerza v Beogradu / *Faculty of Philology, University of Belgrade*

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

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

Simon Šuster
Univerza v Melbournu / *University of Melbourne*

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*

Andrej Žgank
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*

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

Organizacijski odbor / *Organising committee*

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

Ana Cvek
Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Kaja Dobrovoljc
Filozofska fakulteta, Univerza v Ljubljani, Institut "Jožef Stefan" / *Faculty of Arts, University of Ljubljana, Jožef Stefan Institute*

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

Katja Meden
Institut "Jožef Stefan" / *Jožef Stefan Institute*

Mihael Ojsteršek
Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Nataša Rozman
Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Organizatorji / *Organizers*

URNIK / TIMETABLE

Sreda / Wednesday, 14. 9. 2022

Inštitut za novejšo zgodovino / Institute of Contemporary History

09.00-09.30	**Registracija / Registration**
09.30-11.00	**Orange delavnica 1. del / Orange Tutorial Part 1 - *1. nadstropje, Stavba A / 1st floor, Building A***
11.00-11.30	**Odmor za kavo / Coffee break**
11.30-13.00	**Orange delavnica 2. del / Orange Tutorial Part 2 - *1. nadstropje, Stavba A / 1st floor, Building A***
13.00-14.30	**Kosilo / Lunch**
14.30-15.30	**CLARIN delavnica 1. del / CLARIN Tutorial Part 1 - *1. nadstropje, Stavba A / 1st floor, Building A***
15.30-16.00	**Odmor za kavo / Coffee break**
16.00-17.30	**CLARIN delavnica 2. del / CLARIN Tutorial Part 2 - *1. nadstropje, Stavba A / 1st floor, Building A***
17.30	**Neformalno večerno druženje/ Informal dinner**

Četrtek / Thursday, 15. 9. 2022

Fakulteta za družbene vede / Faculty of Social Sciences

08.30-09.15	**Registracija / Registration -** ***1. nadstropje / 1st floor***
09.15-09.30	**Otvoritev / Opening -** ***Room 20 / Soba 20***
09.30-10.00	**Študentska sekcija SLO / Student Session SLO -** ***Room 20 / Soba 20*** David Bordon: *Govoriš nevronsko? Kako ljudje razumemo jezik sodobnih strojnih prevajalnikov* Špela Antloga: *Korpusni pristopi za identifikacijo metafore in metonimije: primer metonimije v korpusu g-KOMET*
10.00-11.00	**Vabljeno predavanje 1 / Keynote 1 -** ***Room 20 / Soba 20*** Eetu Mäkelä (University of Helsinki): *Designing computational systems to support humanities and social sciences research* [Abstract]
11.00-11.30	**Odmor za kavo / Coffee break**

11.30-13.00	**Sekcija 1 SLO / Oral Session 1 SLO-** ***Room 20 / Soba 20***	**Sekcija 1 ANG / Oral Session 1 ENG-** ***Room 21 / Soba 21***
	Iztok Kosem, Jaka Čibej, Kaja Dobrovoljc and Nikola Ljubešić: *Spremljevalni korpus Trendi: metode, vsebina in kategorizacija besedil* Eva Pori, Jaka Čibej, Tina Munda, Luka Terčon and Špela Arhar Holdt: *Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref* Kaja Dobrovoljc, Luka Terčon and Nikola Ljubešić: *Universal Dependencies za slovenščino: nadgradnja smernic, učnih podatkov in razčlenjevalnega modela* Darinka Verdonik, Andreja Bizjak, Andrej Žgank and Simon Dobrišek: *Metapodatki o posnetkih in govorcih v govornih virih: primer baze Artur* Gregor Donaj and Mirjam Sepesy Maučec: *Primerjava načinov razcepljanja besed v strojnem prevajanju slovenščina-angleščina* Tomaž Erjavec, Kaja Dobrovoljc, Darja Fišer, Jan Jona Javoršek, Simon Krek, Taja Kuzman, Cyprian Laskowski, Nikola Ljubešić and Katja Meden: *Raziskovalna infrastruktura CLARIN.SI*	Jakob Lenardič and Kristina Pahor de Maiti: *Slovenian Epistemic and Deontic Modals in Socially Unacceptable Discourse Online* Jure Skubic and Darja Fišer: *Parliamentary Discourse Research in History: Literature Review* Maja Miličević Petrović, Vuk Batanović, Radoslava Trnavac and Borko Kovačević: *Cross-Level Semantic Similarity in newswire texts and software code comments:* *Insights from Serbian data in the AVANTES project* Ajda Pretnar Žagar, Nikola Đukić and Rajko Muršič: *Document enrichment as a tool for automated interview coding* Nikola Ljubešić and Peter Rupnik: *The ParlaSpeech-HR benchmark for speaker profiling in Croatian* Marta Petrak, Mia Uremović and Bogdanka Pavelin Lešić: *Fine-grained human evaluation of NMT applied to literary text: case study of a French-to-Croatian translation*
13.00-13.45	**Kosilo / Lunch**	

13.45-14.30 **Predstavitev plakatov ANG / Poster Session with coffee ENG - *Predprostor predavalnic, prvo nadstropje / Anteroom of the lecture halls, 1st floor***

Jasna Cindrič, Lara Kuhelj, Sara Sever, Živa Simonišek and Miha Šemen:
Data Collection and Definition Annotation for Semantic Relation Extraction

Katja Meden:
Speech-level Sentiment Analysis of Parliamentary Debates using Lexicon-based Approaches

Vladimir Polomac:
Serbian Early Printed Books: Towards Generic Model for Automatic Text Recognition using Transkribus

Branko Žitko, Lucija Bročić, Angelina Gašpar, Ani Grubišić, Daniel Vasić and Ines Šarić-Grgić:
Automatic Predicate Sense Disambiguation Using Syntactic and Semantic Features

Henna Paakki, Faeze Ghorbanpour and Nitin Sawhney:
An approach to computational crisis narrative analysis: a case-study of social media discourse interaction with news narratives about Covid-19 vaccinations in India

Petra Matović and Katarina Radić:
A Parallel Corpus of the New Testament: Digital Philology and Teaching the Classical Languages in Croatia

14.30-16.00 **Sekcija 2 SLO / Oral Session 2 SLO - *Room 20 / Soba 20***

Špela Arhar Holdt, Polona Gantar, Iztok Kosem, Eva Pori,
Nataša Logar Berginc, Vojko Gorjanc and Simon Krek:
Sovražno in grobo besedišče v odzivnem Slovarju sopomenk sodobne slovenščine

Martin Anton Grad and Nataša Hirci:
Raba kolokacijskega slovarja sodobne slovenščine pri prevajanju kolokacij

Tadeja Rozman and Špela Arhar Holdt:

Sekcija 2 ANG / Oral Session 2 ENG -*Room 21 / Soba 21*

Thi Hong Hanh Tran, Matej Martinc, Andraz Repar,
Antoine Doucet and Senja Pollak:
A Transformer-based Sequence-labeling Approach to the Slovenian Cross-domain Automatic Term Extraction

Michal Mochtak, Peter Rupnik and Nikola Ljubešić: *The ParlaSent-BCS dataset of sentiment-annotated*

	Gradnja Korpusa študentskih besedil KOŠ Maja Veselič and Dunja Zorman: *Uporaba Europeaninega podatkovnega modela (EDM) pri digitalizaciji kulturne dediščine: primer Skuškove zbirke iz Slovenskega etnografskega muzeja v projektu PAGODE-Europeana China* Matija Marolt, Mark Žakelj, Alenka Kavčič and Matevž Pesek: *Poravnava zvočnih posnetkov s transkripcijami narečnega govora in petja* Janez Križaj, Simon Dobrišek, Aleš Mihelič, Jerneja Žganec Gros: *Zadnji napredki pri samodejni slovenski grafemsko-fonemski pretvorbi*	*parliamentary debates from Bosnia-Herzegovina, Croatia, and Serbia* Petra Bago and Virna Karlić: *DirKorp: A Croatian corpus of directive speech acts* Sara Košutar, Dario Karl, Matea Kramarić and Gordana Hržica: *Automatic text analysis in language assessment: developing a MultiDis web application* Boshko Koloski, Senja Pollak and Matej Martinc: *What works for Slovenian? A comparative study of different keyword extraction systems* Andrejka Žejn, Mojca Šorli: *Annotation of Named Entities in the May68 Corpus: NEs in modernist literary texts*
19.00-21.00	**Konferenčna večerja / Conference dinner**	

Petek / Friday, 16. 9. 2022

Fakulteta za družbene vede / Faculty of Social Sciences

08.30-09.00	**Registracija / Registration -** ***1. nadstropje / 1st floor***
09.00-10.00	**Študentska sekcija ANG / Student Session ENG -** ***Soba 20 / Room 20*** Ruzica Farmakovski and Natalija Tomic: *Serbo-Croatian Wikipedia between Serbian and Croatian Wikipedia* Meta Jazbinšek, Teja Hadalin, Sara Sever, Erika Stanković and Eva Boneš: *Neural translation model specialized in translating English TED Talks into Slovene* Uroš Šmajdek, Maj Zirkelbach, Matjaž Zupanič and Meta Jazbinšek: *Preparing a corpus and a question answering system for Slovene* Tvrtko Balić: *The CCRU as an Attempt of Doing Philosophy in a Digital World*
10.00-11.00	**Vabljeno predavanje 2 / Keynote 2 -** ***Soba 20 / Room 20*** Benoît Sagot (INRIA): *Large-scale language models: challenges and perspective* [Abstract]
11.00-11.30	**Odmor za kavo / Coffee break**

11.30-12:45 **Sekcija 3 ANG / Oral Session 3 ENG -** ***Soba 20 / Room 20***

Taja Kuzman, Nikola Ljubešić and Senja Pollak:
Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments

Špela Vintar and Andraz Repar:
Human evaluation of machine translations by semi-professionals: Lessons learnt

Aleksandar Petrovski:
A Bilingual English-Ukrainian Lexicon of Named Entities Extracted from Wikipedia

Darja Fišer, Tjaša Konovšek and Andrej Pančur:
Populist and Non-Populist Discourse in Slovene Parliament (1992 – 2018)

Petra Bago:
Progress of the RETROGRAM Project: Developing a TEI-like Model for Pre-standard Croatian Grammars

12:45-13.30 **Kosilo / Lunch**

13.30-14.15 **Predstavitev plakatov z odmorom za kavo SLO / Poster Session with coffee SLO -** ***Predprostor predavalnic / Anteroom of the lecture halls***

Tina Mozetič, Miha Sever, Martin Justin and Jasmina Pegan:
Evalvacijska kategorizacija strojno izluščenih protipomenskih parov

Nina Sangawa Hmeljak, Anna Sangawa Hmeljak and Jan Hrastnik:
Ilukana - aplikacija za učenje japonskih zlogovnih pisav hiragana in katakana s pomočjo asociacij

Vili Grdič, Kaja Perme, Lea Turšič and Alja Križanec:
Šahovska terminološka baza

Lucija Gril, Simon Dobrišek and Andre Žgank:
Akustično modeliranje z različnimi osnovnimi enotami za avtomatsko razpoznavanje slovenskega govora

Saša Babič and Tomaž Erjavec:
Izdelava in analiza digitalizirane zbirke paremioloških enot

	Magdalena Gapsa: *Ocenjevanje uporabniško dodanih sopomenk v Slovarju sopomenk sodobne slovenščine – pilotna študija*
14.15-14.30	**Podelitev nagrad in zaključek / Award&Closing - *Soba 20 / Room 20***
14.30-16.00	**Občni zbor SDJT / SDJT Annual Meeting** **Razvoj slovenščine v digitalnem okolju – jezikovni viri in tehnologije:** **Predstavitev vmesnih rezultatov /** **Development of Slovene in a Digital Environment – Language Resources and Technologies: presentation of intermediate results - *Soba 20 / Room 20***

Kazalo / *Table of Contents*

Predgovor ……..…….…..…… i

Preface …….....……..….....…… ii

Programski odbor / *Programme committee* …….…… iii

Člani programskega odbora / *Programme committee members* …… iii

Organizacijski odbor / *Organising committee* …… vi

Organizatorji / *Organizers* ……. vi

Urnik / *Timetable* ……. vii

Kazalo / *Table of Contents* ….…… xv

VABLJENI PRISPEVKI / *INVITED TALKS* 1

Designing computational systems to support humanities and social sciences research

Eetu Mäkelä 1

Large-scale language modelsl challenges and perspective

Benoît Sagot 2

PRISPEVKI – *PAPERS* 3

The impact of a one-session-phonetic training on the improvement of non-native speakers' pronunciation of English

Amaury Flávio Silva 3

Sovražno in grobo besedišče v odzivnem Slovarju sopomenk sodobne slovenščine

Špela Arhar Holdt, Polona Gantar, Iztok Kosem, Eva Pori, Nataša Logar, Vojko Gorjanc, Simon Krek 10

Izdelava in analiza digitalizirane zbirke paremioloških enot

Saša Babič, Tomaž Erjavec 17

DirKorp: A Croatian Corpus of Directive Speech Acts

Petra Bago, Virna Karlić 23

Universal Dependencies za slovenščino: nadgradnja smernic, učnih podatkov in razčlenjevalnega modela

Kaja Dobrovoljc, Luka Terčon, Nikola Ljubešić 30

Primerjava načinov razcepljanja besed v strojnem prevajanju slovenščina–angleščina

Gregor Donaj, Mirjam Sepesy Maučec 40

Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec, Kaja Dobrovoljc, Darja Fišer, Jan Jona Javoršek, Simon Krek, Taja Kuzman, Cyprian Laskowski, Nikola Ljubešić, Katja Meden 47

ILiAD: An Interactive Corpus for Linguistic Annotated Data from Twitter Posts

Simon Gonzalez 55

Raba *Kolokacijskega slovarja sodobne slovenščine* pri prevajanju kolokacij

Martin Anton Grad, Nataša Hirci 63

Akustično modeliranje z različnimi osnovnimi enotami za avtomatsko razpoznavanje slovenskega govora

Lucija Gril, Simon Dobrišek, Andrej Žgank 71

What works for Slovenian? A comparative study of different keyword extraction systems

Boshko Koloski, Senja Pollak, Matej Martinc 78

Spremljevalni korpus Trendi: metode, vsebina in kategorizacija besedil

Iztok Kosem, Jaka Čibej, Kaja Dobrovoljc, Nikola Ljubešić 86

Automatic Text Analysis in Language Assessment: Developing a MultiDis Web Application

Sara Košutar, Dario Karl, Matea Kramarić, Gordana Hržica 93

Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments

Taja Kuzman, Nikola Ljubešić, Senja Pollak 100

Slovenian Epistemic and Deontic Modals in Socially Unacceptable Discourse Online

Jakob Lenardič, Kristina Pahor de Maiti 108

The ParlaSpeech-HR benchmark for speaker profiling in Croatian

Nikola Ljubešić, Peter Rupnik 117

Cross-Level Semantic Similarity in Newswire Texts and Software Code Comments: Insights from Serbian Data in the AVANTES Project

Maja Miličević Petrović, Vuk Batanović, Radoslava Trnavac, Borko Kovačević 124

The ParlaSent-BCS Dataset of Sentiment-annotated Parliamentary Debates from Bosnia and Herzegovina, Croatia, and Serbia

Michal Mochtak, Peter Rupnik, Nikola Ljubešić 132

Fine-grained human evaluation of NMT applied to literary text: case study of a French-to-Croatian translation

Marta Petrak, Mia Uremović, Bogdanka Pavelin Lešić 141

A Bilingual English-Ukrainian Lexicon of Named Entities Extracted from Wikipedia

Aleksandar Petrovski 147

Serbian Early Printed Books: Towards Generic Model for Automatic Text Recognition using Transkribus

Vladimir Polomac 154

Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref

Eva Pori, Jaka Čibej, Tina Munda, Luka Terčon, Špela Arhar Holdt 162

Document Enrichment as a Tool for Automated Interview Coding

Ajda Pretnar Žagar, Nikola Đukić, Rajko Muršič 169

Parliamentary Discourse Research in History: Literature Review

Jure Skubic, Darja Fišer 177

Annotation of Named Entities in the May68 Corpus: NEs in modernist literary texts

Mojca Šorli, Andrejka Žejn 187

A Transformer-based Sequence-labeling Approach to the Slovenian Cross-domain Automatic Term Extraction

Thi Hong Hanh Tran, Matej Martinc, Andraž Repar, Antoine Doucet, Senja Pollak 196

Metapodatki o posnetkih in govorcih v govornih virih: primer baze Artur

Darinka Verdonik, Andreja Bizjak, Andrej Žgank, Simon Dobrišek 205

Uporaba Europeaninega podatkovnega modela (EDM) pri digitalizaciji kulturne dediščine: primer Skuškove zbirke iz Slovenskega etnografskega muzeja v projektu PAGODE-Europeana China

Maja Veselič, Dunja Zorman 213

Human Evaluation of Machine Translations by Semi-Professionals: Lessons Learnt

Špela Vintar, Andraž Repar 220

Automatic Predicate Sense Disambiguation Using Syntactic and Semantic Features

Branko Žitko, Lucija Bročić, Angelina Gašpar, Ani Grubišić, Daniel Vasić, Ines Šarić-Grgić 227

POVZETKI –ABSTRACTS 235

Progress of the RETROGRAM Project: Developing a TEI-like Model for Croatian Grammars Books before Illyrism

Petra Bago 235

The CCRU as an Attempt of Doing Philosophy in a Digital World

Tvrtko Balić 239

Referencing the Public by Populist and Non-Populist Parties in the Slovene Parliament

Darja Fišer, Tjaša Konovšek, Andrej Pančur 243

Uporaba postopkov strojnega učenja pri samodejni slovenski grafemsko-fonemski pretvorbi

Janez Križaj, Simon Dobrišek, Aleš Mihelič, Jerneja Žganec Gros 248

Poravnava zvočnih posnetkov s transkripcijami narečnega govora in petja

Matija Marolt, Mark Žakelj, Alenka Kavčič, Matevž Pesek 252

A Parallel Corpus of the New Testament: Digital Philology and Teaching the Classical Languages in Croatia

Petra Matović, Katarina Radić 256

Pre-Processing Terms in Bulgarian from Various Social Sciences and Humanities (SSH) Domains: Status and Challenges

Petya Osenova, Kiril Simov, Yura Konstantinova 258

An Approach to Computational Crisis Narrative Analysis: A Case-study of Social Media Narratives Around the COVID-19 Crisis in India

Henna Paakki, Faeze Ghorbanpour, Nitin Sawhney 263

Gradnja Korpusa študentskih besedil KOŠ

Tadeja Rozman, Špela Arhar Holdt 267

ŠTUDENTSKI PRISPEVKI – STUDENT PAPERS 271

Korpusni pristopi za identifikacijo metafore in metonimije: primer metonimije v korpusu g-KOMET

Špela Antloga 271

Neural Translation Model Specialized in Translating English TED Talks into Slovene

Eva Boneš, Teja Hadalin, Meta Jazbinšek, Sara Sever, Erika Stanković 278

Govoriš nevronsko? Kako ljudje razumemo jezik sodobnih strojnih prevajalnikov

David Bordon 286

Data Collection and Definition Annotation for Semantic Relation Extraction

Jasna Cindrič, Lara Kuhelj, Sara Sever, Živa Simonišek, Miha Šemen 292

Serbo-Croatian Wikipedia Between Serbian and Croatian Wikipedia

Ružica Farmakovski, Natalija Tomić 300

Ocenjevanje uporabniško dodanih sopomenk v Slovarju sopomenk sodobne slovenščine – pilotna študija

Magdalena Gapsa 308

Angleško-slovenska šahovska terminološka baza

Vili Grdič, Alja Križanec, Kaja Perme, Lea Turšič 317

Speech-level Sentiment Analysis of Parliamentary Debates using Lexicon-based Approaches

Katja Meden 323

Evalvacijska kategorizacija strojno izluščenih protipomenskih parov

Tina Mozetič, Miha Sever, Martin Justin, Jasmina Pegan 331

Ilukana – aplikacija za učenje japonskih zlogovnih pisav hiragana in katakana s pomočjo asociacij

Nina Sangawa Hmeljak, Anna Sangawa Hmeljak, Jan Hrastnik 339

Filter nezaželene elektronske pošte za akademski svet

Anja Vrečer 345

Preparing a Corpus and a Question Answering System for Slovene

Matjaž Zupanič, Maj Zirkelbach, Uroš Šmajdek, Meta Jazbinšek 353

Designing computational systems to support humanities and social sciences research

Eetu Mäkelä
University of Helsinki, Finland
P.O. Box 24, 00014
eetu.makela@helsinki.fi

Abstract

From the viewpoint of the humanities and social sciences, collaborations with computer scientists often fail to deliver. In my research group, we have tried to understand why this is, and what to do about it. In this talk, I will discuss three key elements that we have discovered:

Often, datasets in the humanities and social sciences are not neatly representative of the object of interest. Systems need to provide ways in which to evaluate and counter the biases, confounders and noise in the data. Often, there is also a large gap between what is in the data, and what would be of interest. This gap needs to be bridged using algorithms, but care must be given that a) what the algorithm produces actually matches the interest and b) that its application does not introduce bias of its own (also interestingly, algorithm performance metrics of interest here often differ from those generally used in NLP/computer science). On a process level, collaboration between researchers from different disciplines is hard due to discrepancies in expectations relating to all facets of research, from research questions through methodology to the publication of results. Projects and systems need to acknowledge this, and be designed to facilitate iterative movement in the right direction.

Bio

Eetu Mäkelä is an associate professor in Human Sciences–Computing Interaction at the University of Helsinki, and a docent (adjunct professor) in computer science at Aalto University. At the Helsinki Centre for Digital Humanities, he leads a research group that seeks to figure out the technological, processual and theoretical underpinnings of successful computational research in the humanities and social sciences.

Additionally, he serves as a technological director at the DARIAH-FI infrastructure for computational humanities and is one of three research programme directors in the datafication research initiative of the Helsinki Institute for Social Sciences and Humanities. For his work, he has obtained a total of 19 awards, including multiple best paper awards in conferences and journals, as well as multiple open data and open science awards. He also has a proven track record in creating systems fit for continued use by their audience.

Large-scale language models: challenges and perspective

Benoît Sagot
Inria Paris (équipe ALMAnaCH)
2 rue Simone Iff CS 42112
75589 Paris Cedex 12, France
benoit.sagot@inria.fr

Abstract

The emergence of large-scale neural language models in Natural Language Processing (NLP) research and applications has improved the state of the art in most NLP tasks. However, training such models requires enormous computational resources and training data. The characteristics of the training data has an impact on the behaviour of the models trained on it, depending for instance on the data's homogeneity and size. In this talk, I will speak about how we developed the large-scale multilingual OSCAR corpus. I will describe the lessons we learned while training the French language model CamemBERT, the first large-scale monolingual model for a language other than English, especially in terms of the influence of size and heterogeneity of the training corpus. I will also sketch out a few research questions related to biases in large-scale language models, with a focus on the impact of tokenisation and language imbalance, in the context of the BigScience initiative. I will conclude with my thoughts on the future of language models and their impact on NLP and other data processing fields (speech, vision).

Bio

Benoît Sagot, Directeur de Recherches (Senior Researcher) at Inria, is the head of the Inria project-team ALMAnaCH in Paris, France. A specialist in natural language processing (NLP) and computational linguistics, his research focuses on language modelling, language resource development, machine translation, text simplification, part-of-speech tagging and parsing, computational morphology and, more recently, digital humanities (computational historical linguistics and historical language processing). He has been the PI or co-PI of a number of national and international projects, and is the holder of a chair in the PRAIRIE institute dedicated to research in artificial intelligence. He is also the co-founder of two start-ups where he uses his expertise in NLP and data mining for the automatic analysis of employee survey results.

The impact of a one-session-phonetic training on the improvement of non-native speakers' pronunciation of English

Amaury Flávio Silva

Technology College of Jacareí (FATEC Jacareí) - São Paulo, Brazil
Rua Faria Lima, 155 – Jardim Santa Maria, Jacareí – SP, Brazil, Zip Code 12328-070
amaury.silva@fatec.sp.gov.br

Abstract

Due to the difficulties L2[1] learners face regarding pronunciation, we conducted an experiment to find out if the participants of a one-session phonetic training would present any sign of improvement in their speech a week after the session. In order to evaluate their improvement, it was checked if the interword phonetic phenomena resyllabification, blending and hiding could be found in the subjects' speech. Furthermore, intraword-level pronunciation was also investigated. The findings have shown that betterment related to the presence of resyllabification occurred to all the subjects, but improvement to the other phenomena studied happened heterogeneously.

1. Introduction

Until the end of the 21st century, there was a limited number of studies regarding pronunciation (Derwing and Munro, 2005). This negligence is attributed to the fact that pronunciation was considered an aspect of language learning that could be naturally acquired through the learning process. However, since 2005 this viewpoint has been changing inasmuch as several studies, conferences, and articles about L2 pronunciation started to arise (Thomas and Derwing, 2014).

Despite the fact the importance of L2 pronunciation has become more evident, there are still L2 students, teachers and researchers who consider pronunciation teaching as being unnecessary as they reckon it can be learnt through exposure.

We regard pronunciation instruction as an essential part of the L2 teaching process. Its essential character becomes more evident when L2 learners, in spite of being studying the L2 for many years, still struggle to correctly pronounce the L2-language sounds, especially the ones that are not part of their L1 inventory systems. Nonetheless, we do not believe that achieving native-like pronunciation is necessary in that one's pronunciation being intelligible enough not to cause misunderstandings or hamper the flow of communication is what should be expected.

Owing to our belief that pronunciation instruction should be part and parcel of L2 language learning, we decided to carry out a study that aims to check the benefits of a one-session-pronunciation training in the improvement of the pronunciation of a group of subjects, Brazilian learners of English as a foreign language.

With regard to this one-session training, we hypothesize that there may be some kind of improvement in the subjects' pronunciation, but more sessions will be necessary to address all the pronunciation problems they may have. Moreover, the less proficient the students are, the higher will be the number of sessions necessary to help them deal with their pronunciation problems.

The dataset used during the training session was based on a study developed by Silva (2021) in which he studied examples of coarticulatory effects that we also incorporated in our pronunciation instruction session.

2. Goal of the paper

This paper, whose goal is to investigate the efficacy of a one-session phonetic training to enhance the participants' performance in pronunciation tasks also aims to provide a guideline that L2 teachers could use to assist their students improve. Furthermore, we hope that researchers could use the methods here applied to carry out new experiments in this area.

3. Theoretical Background

The increasing number of pronunciation-related studies since 2005 reveals the importance that pronunciation instruction has in the L2 learning process. Not only does it allow learners to become more confident when they speak, it also improves speech intelligibility as it helps to avoid misunderstandings.

Due to the importance of pronunciation, Thomas and Derwing (2014) wrote an article in which they evaluated 75 L2 pronunciation studies, most of which affirm that there was some kind of improvement in the speakers' pronunciation due to the training they took. The authors point out that diverging results take place owing to a few factors such as 'learner individual differences, goals and foci of instruction, type and duration of instructional input and assessment procedures' (p.1).

Most of the 75 studies focused on the achievement of native-like pronunciation by the learner and consisted in the use of computer-assisted tools. Moreover, the studies aimed at teaching the pronunciation of individual segments instead of teaching suprasegmental features, which would involve, for instance, resyllabification, prosodic boundaries, word stress, intonation, and speech rate.

In order to teach the pronunciation of segments, most of the time, the learners were engaged in activities that

[1] We use the term 'L2' to refer to the teaching of English as a foreign and as a second language.

required them to read texts aloud, instead of producing spontaneous speech.

When it comes to the quality of a pronunciation study, Thomas and Derwing (2014) mention a few features they should have. Firstly, they express their belief that pronunciation instruction should focus on 'helping students become more understandable' (p. 2). From this principle, they point out that an ideal pronunciation study should be able to give plenty of information on the subjects, have enough data that could be used to carry out statistical analyses, have a control group, and should not be limited to reading aloud tasks, i.e., it should also include spontaneous speech samples. Finally, it should include delayed assessment to verify the lasting effect of the pronunciation instruction.

With regard to qualitative analyses, they should encompass aspects such as motivation, type of interactions in the L2 and even social influences (Thomas and Derwing, 2014).

The training input of the studies surveyed, which was either classroom instruction or computer assisted pronunciation training, ranged from the manipulation of segments (Wang, 2002; Lee, 2009) to providing students with speech samples produced by native speakers so students could listen to them and compare them with their own productions (Gonzales-Bueno, 1997; Guilloteau, 1997, Weinberg and Knoerr, 2003; Lord, 2005; Pearson et al., 2011).

The learners' performances were evaluated by human listeners in 79 per cent of the studies and the other 21 per cent were evaluated using acoustic analyses.

The majority of the pronunciation training studies reviewed by Thomson and Derwing (2014) lacked explicit theoretical background so that the pronunciation training was solely based on the researchers' own experience. In our training, we considered the research about reduction phenomena led by Silva (2016, 2021), the findings on coarticulation conducted by Browman and Goldstein (1986, 1989), and the work developed by Vroomen and Gelder (1999) about resyllabification. We will be discussing this theoretical background later on in this section.

One important aspect that was not clear in the studies was the procedure taken during the training sessions (training input). The lack of clarity in the methodological procedures prevent other teachers and researchers from replicating the steps used in the studies in their own classes or research. Therefore, detailed methodological procedure is necessary 'for the benefit of other researchers and teachers' (Thomson and Derwing , 2014, p. 11).

The research on pronunciation training by Thomson and Derwing (2014) revealed that most of the participants showed some kind of improvement after the training. Nonetheless, the majority of studies only focused on the instruction of single sounds such as the contrast of /iː/ and /ɪ/. Should the studies be on several segmental and suprasegmental features, more time would be necessary so the learners could present significant improvement.

Another issue that questions the efficacy of the studies is whether or not the assessment used in them would reflect in the improvement of intelligibility when language is used in real-life contexts. For such issue to be solved, the studies should focus on 'more intelligible, as opposed to less-accented speech … (and) include a variety of assessment tasks' (Thomson and Derwing, 2014, p. 13-14). Furthermore, the authors state that evaluating the efficacy of the studies in a naturalistic fashion would take years, instead of weeks or months.

We believe that any research should depart from a well-established theoretical standpoint. Hence, since in our analyses we focused on the influence adjacent intra or interword segments have on one another, we turned to the studies developed by Browman and Goldstein (1986, 1989) on coarticulation.

According to Browman and Goldstein (1986, 1989), adjacent segments may be subjected to the phenomena called blending and hiding. Blending occurs when adjacent segments share the same articulator so that they cannot be produced without disturbance in their constriction location. An example of this phenomenon takes place when the segments [t] and [ð] from the context 'I want that' have to be produced one after the other. In this context, the constriction location of either segment may be disturbed as they are both characterized by a tongue tip gesture. Thus, the canonical production of the alveolar plosive and the interdental fricative may be realized as an approximant and as a dental fricative respectively.

Hiding occurs when adjacent segments do not share the same articulator so that the production of the first segment is overlapped by the production of the second one. Such phenomenon may occur when the segments [t] and [b] from the context 'I can't buy it' have to be produced one after the other. When this happens, the gesture of mouth closure to produce the bilabial consonant 'hides' the burst that would be caused by the release of the alveolar plosive.

Being aware of how these phenomena work allows speakers to reduce articulatory effort when they speak as the excursion of the articulators is decreased. The reduction in articulatory effort was studied by Silva (2016, 2021). In his investigations, he noticed that reduction is a strategy commonly used by native speakers and which can be characterized by the replacement of a segment that calls for high excursion of the articulators by one that does not (low-hierarchy reduction). Reduction can be also characterized by a segment deletion (high-hierarchy reduction).

Another phenomenon that causes reduction in articulatory effort is the one called resyllabification. It happens when 'consonants are attached to syllables other than those from which they originally came' (Vroomen and Gelder, 1999, p.413). An example of this phenomenon is the sentence 'you can evaluate this' in which the consonant /n/ of the word 'can' is coarticulated with the vowel /ɪ/ of the word 'evaluate.' This process contributes to maintain the speech flow as the speaker does not need to add a pause between adjacent words.

The analyses carried out in this study as well as the concepts explained during the training session were based on the phenomena blending and hiding (Browman and Goldstein, 1986-1989), reduction (Silva 2016, 2021), and resyllabification (Vroomen and Gelder, 1999).

4. Methods

In this section, we will describe details related to the subjects that participated in the study, the research dataset, the acoustic inspection and the training session.

4.1 Subjects

In order to conduct the analysis, we had the participation of four subjects, native speakers of Brazilian Portuguese (three males and one female), who study English as a foreign language. The subject 'English' is part of the Technological course the subjects were taking and all of them were enrolled in the same class, taking the third semester. It is important to point out that English is offered throughout the duration of the course, six semesters, and, despite the fact all the students were in the same class, their proficiency level was not the same.

The four participants will be referred to as subjects, 'S,' in this investigation.

4.2 Research dataset

The research dataset, table 1, is an extract from the program Actors' Studio (season 12, episode 13, released on July 2006) that was sent to the subjects so they would have to record and send it to the trainer before the training session. After the session, they would record it once more and send it to the trainer again so their improvement could be analyzed. We would like to point out that in our experiment, we asked the subjects to use their own smartphones or computers to record the dataset. This was done as they could not come to college to record it in its sound laboratory due to the restrictions related to the COVID-19 pandemic.

The same text was used in the pre and post-training phase as we aimed to analyze whether or not improvement could be observed in the second recording in terms of the group of words we selected that encompass the phenomena described in tables 2-4.

This dataset was also selected by Silva (2021) on his investigation about coarticulatory phenomena analysis.

It's funny, you know, someone comes into your life at a certain time and that's one of the great things that happens on Earth is you're mysteriously guided towards these people that you get to dance with, you know. And I thought "How great is that", he's kind of, like, I don't want to say an angel to her, but he's someone who needs as much as he's prepared to offer, and he has seen a lot of life, and he's not a typical lawyer-type.

Table 1: Research Dataset

Using the dataset above, we selected fragments in which the phenomena resyllabification, blending and hiding could take place. Furthermore, we also analyzed the pronunciation of a group of words that the students mispronounced in the pre-training recording.

The phenomenon resyllabification was investigated in 11 contexts, which we will present in the next table.

Contexts	Phonemes involved
'comes into'	/z/and /ɪ/
'at a certain'	/t̬/and /ə/
'one of'	/n/and /ə/
'on Earth'	/n/and /ɜr/
'and I'	/d/and /aɪ/
'great is'	/t̬/and /ɪ/
'kind of'	/d/and /ə/
'an angel'	/n/and /eɪ/
'as much as'	/tʃ/and /ə/
'seen a'	/n/and /ə/
'lot of life'	/t̬/and /ə/

Table 2: Resyllabification phenomenon

With regard to the phenomena blending and hiding, we analyzed eight contexts, presented in the next table.

Contexts	Phonemes involved
'certain time'	/n/and /t/
'great things'	/t/and /θ/
'guided towards'	/d/and /t/
'get to dance'	/t/and /t/
'prepared to'	/d/and /t/
'typical lawyer'	/l/and /l/
'these people'	/z/and /p/
'I don't want'	/t/and /w/

Table 3: Blending and hiding phenomena

Lastly, when it comes to word-level pronunciation, the words presented in the table below were investigated.

Words	Pronunciation errors found
'someone'	Phoneme substitution and insertion of a phoneme
'certain'	Phoneme substitution and word stress
'mysteriously'	Phoneme substitution and word stress
'towards'	Phoneme substitution
'thought'	Phoneme substitution
'offer'	Phoneme substitution and word stress
'lawyer'	Phoneme substitution and word stress

Table 4: Word-level pronunciation

4.3 Phonetic inspections

The phonetic inspection was carried out with the use of the free software PRAAT, version 6.0.39, developed by Paul Boersma and David Weenik (2018), from the Institute of Phonetic Science of the University of Amsterdam.

The inspections were based on the observation of the waveform, the broadband spectrogram, the fundamental frequency and the intensity of the phonemic segments.

4.4 Training Session

The training took place in a single 50-minute session of an online class. It was recorded so that the subjects could revisit it as many times as they wanted in order to review the concepts explained.

The training session the subjects participated was provided by the researcher of this work.

At the beginning of the training, which took place after the first recording of the dataset was sent by the subjects, the original recording of the dataset was played, and the corresponding script was projected on the screen for the subjects to follow it. The recording was played three times.

After that, the concept of resyllabification was explained and the first context where such phenomenon occurred according to table 2, 'comes into', was presented to the subjects (the orthography along with the recording). The context was played three times.

The subjects were asked to pay close attention to the recording of the context as they would have to repeat it afterwards. If they could not repeat it, the trainer would repeat the context himself at least three more times in order to assist the subjects grasp what and how they should say it.

Before moving on to the next context, the original recording was played one more time and the subjects were asked to repeat it. Not until all the subjects were able to repeat the context intelligibly, would the trainer teach the next context.

The procedure described above was followed to teach the other contexts including resyllabification, blending and hiding phenomena. Word-level pronunciation instruction followed the steps related to playing the recording three times before repetition. However, after analyzing the first recording, we reckoned the need to teach word stress and phoneme pronunciation.

It is important to point out that we did not use technical terms during the training as our focus was simply on improving their pronunciation.

When it comes to the difficulty the subjects presented to pronounce a word or group of words, the trainer noticed that it was necessary to teach the articulation of some phonemes, especially the ones not present in the subjects' L1inventory system. After the instruction of the articulation of such phonemes, improvements could be observed in their pronunciation.

The subjects, after the training session, had access to the original recording of the dataset and to a version recorded by the trainer, which was produced with a slower speech rate so that it could be helpful to less proficient subjects. These recordings were tools the subjects could use to improve their pronunciation before making the second recording that had to be sent within a week.

Once all the subjects had sent their recordings, we started the data analysis, whose results are presented in the next section.

5. Data analyses

The analyses in this chapter will feature figures that contain the waveform, spectrogram, segmentation, and spelling of a selection of the contexts investigated. However, at the end of section 5.1, a table with a summary of all the contexts investigated during the pre-training phase is provided and one at the end of section 5.2 with all the contexts analyzed in the post-training phase is available.

We reckon it is important to point out that the subjects reported that they recorded the dataset several times and that they sent us the version they judged to be the best.

5.1. Pre-training analyses

In this section, we will present the analyses that refer to the pre-training recordings. The first one refers to the context 'and I,' resyllabification.

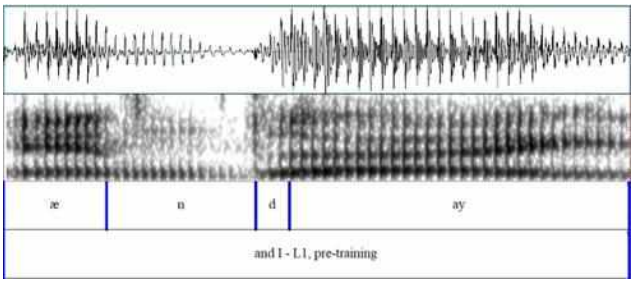


Figure 1: Production of 'and I' by L1, pre-training

Through the analysis of the broadband spectrogram and its corresponding waveform above, we can infer that there was no pause between the production of the adjacent segments [d] and [ay] so the phenomenon resyllabification was observed.

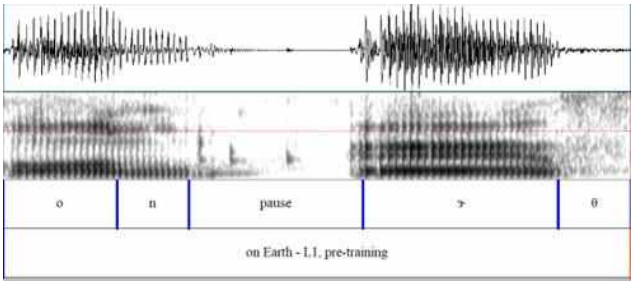


Figure 2: On Earth – S1, pre-training – Figure shows pause between the words 'on' and "Earth'

In the production of 'on Earth,' figure above, there was a pause between the segments [n] and [ɜr] so that the phenomenon resyllabification did not take place.

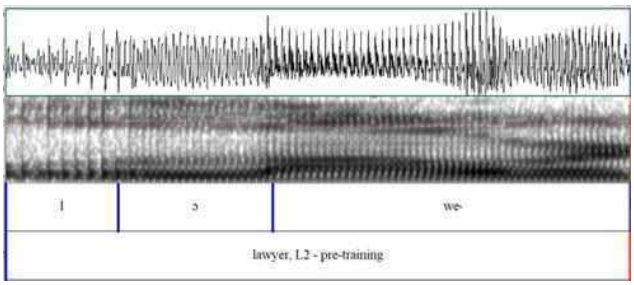


Figure 3: Production of 'lawyer' by S2

The figure above, which presents acoustic information, shows that the subject mispronounced the word lawyer in that [ləwər/] was produced instead of [lɔɪər/].

A summary of the data analyses that refer to the pre-training recordings is presented in the table below. The word or group of words written indicates that the phenomenon in the corresponding column was observed in their production.

Pre-training			
Subjects	**Resyllabification** (group of words in which the phenomenon was observed)	**Blending/Hiding** (group of words in which the phenomena were observed)	**Word-level pronunciation** (mispronounced words)
S1	'and I' 'great is' 'kind of' 'lot of life'	'guided towards' 'typical lawyer' 'these people' 'I don't want to'	All the words were mispronounced except 'someone' and 'offer'
S2	'at a certain' 'one of' 'on Earth' 'and I' 'kind of' 'lot of'	'great things' 'guided towards' 'get to dance' 'prepared to' 'typical lawyer' 'these people'	'mysteriously' 'thought' 'lawyer'
S3	'comes into' 'at a certain' 'one of' 'on Earth' 'and I' 'great is'	'certain time' 'get to dance' 'prepared to' 'these people'	'mysteriously' 'thought' 'lawyer'
S4	All the contexts except 'and I'	All the contexts except 'certain time'	'thought' 'offer' 'lawyer'

Table 5: Data analyses concerning the pre-training recordings

5.2. Post-training analysis

In this section, we will present the analyses that refer to the post-training recordings. The first one refers to the context 'on Earth,' resyllabification.

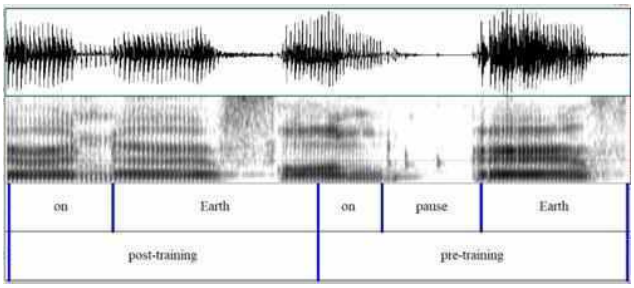


Figure 4: Concatenated productions of 'on Earth' by S1. Post-training left and pre-training right.

The concatenated productions of 'on Earth' by S1, presented in the figure above, demonstrate that the phenomenon resyllabification was observed in the post-training recording, but not in the pre-training recording. This fact is confirmed by the absence of pause between the segments [n] and [ɜr] in the post-training phase that did not occur in the pre-training phase as a pause is present in the spectrogram.

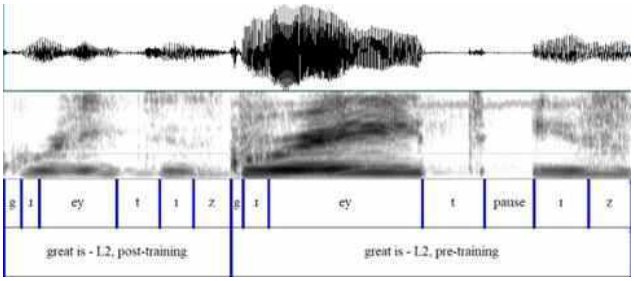


Figure 5: Concatenated productions of the post and pre-training versions of 'great is' by S2

As shown in the analysis of the context 'on Earth,' figure 4, in the context 'great is' by S2, figure above, the phenomenon resyllabification was observed in the post-training recording, but not in the pre-training one.

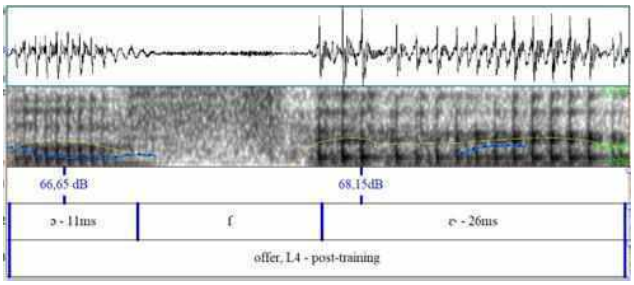


Figure 6: Production of the word 'offer' by S4

The analysis of the production of the context 'offer,' produced by S4, shows that the word stress was placed on the syllable '-fer' instead of the syllable 'of-', which is where the correct stress for the word 'offer' should occur. The word stress on the syllable '-fer' can be confirmed not only by the higher duration of the segment [ɛr], but also the higher intensity of this segment in comparison to the segment [ɔ]. What's more, S4 used the segment [ɛr] instead of /ər/ in the second syllable.

A summary of the data analyses that refer to the post-training recordings are presented in the table below. The word or group of words written indicates that the phenomenon in the corresponding column was observed in their production.

Post-training			
Subjects	**Resyllabification** (group of words in which the phenomenon was observed)	**Blending/Hiding** (group of words in which the phenomena were observed)	**Word-level pronunciation** (mispronounced words)
S1	All the contexts	'certain time' 'great things' 'get to dance' 'typical lawyer' 'these people' 'I don't want to'	All the words were mispronounced except 'someone' and 'offer'
S2	All the contexts	'certain time' 'great things' 'prepared to' 'these people'	'someone' 'mysteriously' 'thought' 'lawyer'
S3	'comes into' 'at a certain' 'one of' 'on Earth' 'and I' 'great is' 'kind of' 'a lot of life'	'certain time' 'great things' 'get to dance' 'prepared to'	'someone' 'mysteriously' 'thought'
S4	All the contexts except 'and I'	All the contexts	'thought' 'offer'

Table 6: Data analyses concerning the post-training recordings

6. Discussion

The analyses have shown that the one-session phonetic training was useful to help the subjects improve their pronunciation with regard to the resyllabification phenomenon. Nevertheless, no homogeneous improvement was observed in terms of the remaining phenomena analyzed.

The observed improvement in the resyllabification feature in the production of S1 and S2 was characterized by the use of this strategy in the production of all the contexts analyzed in the post-training recording, fact not observed in the pre-training one. S3 also demonstrated improvement in the use of this strategy in that it was used in two more contexts in the post-training recording. No improvement was observed in terms of resyllabification for S4, but they had already presented excellent performance of this strategy as there was only one context where it was not applied.

The presence of the phenomena blending, and hiding was found in the production of S1 in most of the contexts and in all the contexts produced by S4 in the post-training recording. Such phenomena were noticed in fewer contexts in the production of S2 and in the same number in the production of S3 in the post-training recording.

With regard to the last feature analyzed after the post-training session, word-level pronunciation, no improvements were observed in the production of S1, S2 made one more mistake and S3 improved the production of the word 'lawyer' but mispronounced a word he had produced correctly in the pre-training session, 'someone'. S4 improved the production of the word 'lawyer' but continued mispronouncing the words 'thought' and 'offer'.

Our findings have revealed different levels of improvement in the subjects' performance so that S1 is the one who presented the most improvement. S2 and S3's performances betterment was limited to the presence of the resyllabification phenomenon. S4 is the most proficient subject who presented only a few mistakes in the pre-training recording and was able to use the phenomena blending and hiding in all the contexts and to improve the pronunciation of a word after training.

The hypothesis we presented at the beginning of our work was confirmed as the subjects' pronunciation was somehow improved, but more sessions are necessary to address certain pronunciation problems such as word-level pronunciation and the phenomena hiding and blending.

In future studies, we could ask the subjects to report on the time they have dedicated to study and practice the pronunciation concepts studied during the training session. Furthermore, we could ask judges to evaluate the students' performance before and after the training session to find out if a perceptual betterment in their pronunciation was clear, i.e., if the level of intelligibility was enhanced.

We believe vehemently that, although the number of participants was not adequate through a quantitative standpoint as our aim was to conduct a qualitative investigation, the study has shown that improvement did occur, bringing to light the importance of phonetic instruction. Moreover, we expect that the procedure we used during the training session was clear enough so the study could be replicated by other researchers.

Lastly, we hope to continue our investigation by providing the subjects with more training sessions, evaluate them at least five months after the first training session and have more participants so we could carry out statistical analysis.

7. Reference

Paul Boersma, David Weenink. 2018. Praat: doing phonetics by computer, version 6.0.39. Available at: <http://www.praat.org>. Access on: 2 Dec. 2018

Catherine Browman, Louis Goldstein. 1986. Towards an articulatory phonology. Phonology, v. 3, pages. 219-252.

_____. Articulatory gestures as phonological units. Phonology, v. 6, 1989, pages 201-251.

Tracey M. Derwing, M, Murray J. Munro.2005. Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly* 16/1: pages 71-77.

Manuela Gonzales-Bueno. 1997. The effects of formal instruction on the acquisition of Spanish stop consonants. *Contemporary Perspectives on the Acquisition of Spanish* 2: pages 57-75.

Nancy Clarke Guilloteau. 1997. *Modification of phonetic categories in French as a second language: Experimental studies with conventional and computer-based intervention methods*. Unpublished Ph. D. thesis. University of Texas at Austin.

Lee Ji-Yeon. 2009. *The effects of pronunciation instruction using duration manipulation in the acquisition of English vowel sounds by pre-service Korean EFL teachers*. Unpublished Ph.D. thesis, University of Kansas.

Gillian Lord. 2005. (How) can we teach foreign language pronunciation? On the Effects of a Spanish Phonetics Course. *Hispania*, 88/3: pages 557-567.

Pamela Pearson, Lucy Pickering, Rachel DaSilva. 2011. The impact of computer assisted pronunciation training on the improvement of Vietnamese learner production of English syllable margins, In: Levis J. and LeVelle K. (eds). *Proceedings of the 2nd Pronunciation in Second Language Learning and Teaching Conference*, Iowa State University, pages 169-180.

Amaury F. Silva. 2021. Coarticulatory phenomena analysis in English based on the articulatory phonology. São Paulo. *CBTecLe* v.1, n.1.

_____.2016. *Percepção de reduções em inglês como L2*. Unpublished Ph.D. thesis, PUC-SP.

Ron I. Thomson, Tracey M. Derwing. 2014. The effectiveness of L2 pronunciation Instruction: a narrative review. Oxford, Oxford University Press.

Jean Vroomen, Beatrice De Gelder. 1999. Lexical access of resyllabified words: evidence from phoneme monitoring. Memory & cognition, 27(3), pages 413–421.

Xinchun Wang. 2002. *Training Mandarin and Cantonese speakers to identify English vowel contrasts: long term retention and elects on production*. Unpublished Ph.D. thesis, Simon Fraser University.

Alysse Weinberg, Hélène Knoerr. 2003. Learning French pronunciation: Audiocassettes or multimedia? *CALICO Journal*, 20/2: pages 315-336.

Sovražno in grobo besedišče v odzivnem Slovarju sopomenk sodobne slovenščine

Špela Arhar Holdt,*‡ Polona Gantar,* Iztok Kosem,* Eva Pori,*
Nataša Logar, Vojko Gorjanc,* Simon Krek***

* Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@ff.uni-lj.si, iztok.kosem@ff.uni-lj.si, eva.pori@ff.uni-lj.si,
vojko.gorjanc@ff.uni-lj.si, simon.krek@ff.uni-lj.si
** Fakulteta za družbene vede, Univerza v Ljubljani
Kardeljeva ploščad 5, 1000 Ljubljana
natasa.logar@fdv.uni-lj.si
‡ Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
spela.arharholdt@fri.uni-lj.si

Povzetek

V prispevku predstavljamo rešitve za identifikacijo in označevanje sovražnega ter grobega besedišča v okviru koncepta odzivnega Slovarja sopomenk sodobne slovenščine. Ker gre za prvi tovrstni projekt, so pripravljene rešitve v veliki meri inovativne, umeščene pa v okvir problematike avtomatske strojne izdelave slovarja, njegove odprtosti in vključenosti uporabniške skupnosti. Prispevek prikazuje identifikacijo sovražnega in grobega besedišča ter pripis oznak oziroma opozorilnih ikon z daljšimi pojasnili. Oznake temeljijo na sporočanjskem namenu oziroma učinku, pri čemer je njihovo bistvo informacija o možnih posledicah rabe. Pri označevanju tako kot pri izdelavi celotnega slovarja posvečamo veliko pozornost digitalnemu mediju in vizualizaciji rešitev v njem. Ker je odzivnost eden ključnih konceptov slovarja, se tudi pri rešitvah glede označevanja zavedamo pomembnosti sodelovanja z uporabniško skupnostjo, zato predlagamo še rešitve za sodelovanje s skupnostjo pri dodajanju oznak.

Extremely Offensive and Vulgar Vocabulary in the Responsive Thesaurus of Modern Slovene

In the paper we present the solutions for identification and annotation of extremely offensive and vulgar vocabulary, which can be found in the responsive Thesaurus of Modern Slovene. As this is the first project of its kind, the prepared solutions are to a great extent innovative, and have been devised considering the use of automatic methods in dictionary compilation, open access nature of dictionary data, and the inclusion of users into the compilation process. The paper describes the process of identification of extremely offensive and vulgar vocabulary, as well as the attribution of labels and warning icons containing longer explanations. The labels are based on their communicative purpose or effect, and are focussed on providing the information about potential consequences of word use. During the processes of labelling and dictionary compilation, considerable attention is paid to the digital medium and related visualisation solutions. As responsiveness is one of the key concepts of the dictionary, a part of preparing the labelling solutions was to design ways of including user community in labelling.

1. Uvod

Slovar sopomenk sodobne slovenščine (SSSS) je oblikovan po modelu odzivnega slovarja: v prvem koraku je bil pripravljen strojno, nadaljnje urejanje podatkov pa poteka po korakih in v sodelovanju jezikoslovcev ter širše zainteresirane skupnosti (Arhar Holdt et al., 2018: 404). V SSSS lahko slovarski uporabniki ob strojno pripravljeno sopomensko gradivo dodajo lastne predloge sopomenk, za vse sopomenke v slovarju pa je mogoče tudi glasovati in gradivo na tak način (pomagati) potrditi ali zavrniti.[1]

Vključevanje strojnih postopkov in predlogov uporabniške skupnosti v slovaropisne delotoke odgovarja na potrebe sodobnega časa, kot sta potreba skupnosti po odprto dostopnih jezikovnih podatkih in želja slovarskih uporabnikov po demokratičnem sodelovanju pri razvoju temeljne jezikovne infrastrukture. Na drugi strani pa ima neposredno objavljanje strojnega in uporabniško dodanega (nepregledanega) gradiva lahko tudi neželene posledice, ki jih je treba pri razvoju odzivnega modela predvideti in ustrezno obravnavati. Med prioritetami za razvoj SSSS je tako brez dvoma obravnava besedišča, ki vrednostno poimenuje posamezne družbene skupine in njihove pripadnike. Tako besedišče se trenutno v slovarju (lahko) pojavlja na različnih mestih in na različne načine.

Namen prispevka je predstaviti obseg problematike, ki se pri odzivnem slovarju pomembno razlikuje od tradicionalnih slovaropisnih projektov, in opisati rešitve, ki bodo vključene v prihajajočo nadgradnjo SSSS. Med temi želimo posebej izpostaviti nove načine identifikacije in označevanja sovražnega, grobega ter drugače negativno vrednotenega besedišča, ki SSSS presegajo in so uporabne za različne sodobne jezikovne vire.

2. Sovražno, grobo, tabuizirano, zaničljivo … v družbi, jeziku in slovarju

Na kratko je mogoče sovražni govor opredeliti kot "aktivno javno spodbujanje antipatije do določene, ponavadi šibke, družbene skupine" (Rebolj, 2008: 13), v daljši in bolj povedni obliki pa kot (Petković in Kogovšek Šalamon, 2007: 23):

> ustno ali pisno izražanje diskriminatornih stališč. Z njim širimo, spodbujamo, promoviramo ali opravičujemo rasno sovraštvo, ksenofobijo, homofobijo, antisemitizem, seksizem in druge oblike sovraštva, ki temeljijo na nestrpnosti. Mednje sodi tudi nestrpnost, ki se izraža z agresivnim nacionalizmom in etnocentrizmom, z diskriminacijo in sovražnostjo zoper manjšine, migrante in migrantke. Žrtve sovražnega govora praviloma niso posamezniki, pač pa ranljive družbene skupine. V osrčju sovražnega govora je prepričanje, da so nekateri ljudje manj vredni, zato je cilj sovražnega govora v razčlovečenju, ponižanju, ustrahovanju

[1] Slovar v vmesniku je na https://viri.cjvt.si/sopomenke/slv/, kot slovarska baza pa na repozitoriju CLARIN.SI (Krek et al., 2018). Strojno pripravo slovarja opisujejo Krek et al. (2017), koncept odzivnega slovarja pa Arhar Holdt et al. (2018).

in poslabšanju družbenega položaja tistih, proti katerim je naperjen.

Motl in Bajt (2016: 7) ugotavljata, da je sovražni govor deležen precejšnje pozornosti v različnih vedah, od prava, sociologije in komunikologije do psihiatrije in informatike, pridružimo pa jim lahko tudi jezikoslovje – predvsem jezikoslovje, povezano s slovarji. Ameriško slovaropisje (Hughes, 2011: 3. pogl.) je že pred desetletji v svoje vire načrtno vgradilo tudi občutljivost do ranljivih družbenih skupin, pri čemer ni zanemarilo nobenega od delov geselskega članka: razlag, oznak in zgledov rabe (Logar et al., 2020: 104). V manjši meri in pozneje, a vendarle so se opozorila o nujni tovrstni družbeni občutljivosti ter odgovornosti pojavila tudi v slovenskem prostoru (npr. Gorjanc, 2005; Kern, 2015; Logar et al., 2020: 91, 104), a jih kljub temu do sedaj ni polno upošteval še noben slovarski projekt.

Ni pa zgolj sovražni govor tisti, ki ga je treba v slovarjih obravnavati posebej pozorno. Kritično slovaropisje opozarja, da je treba pri slovarskih opisih izrecne (in nove) rešitve iskati pri vseh elementih, ki prinašajo vljudne in nevljudne vidike jezika, tabuiziranost, so usmerjeni v vrednotenje, konotacijo, kulturne aluzije ipd., še posebej pa je treba biti pozoren na nestabilna in spreminjajoča se poimenovanja vseh oblik drugosti (Moon, 2014: 85). Pri tem se sodobno slovaropisje ne more sklicevati na tradicionalne modele jezikovnega opisovanja in delovanja. Nikakor pri tem ni sprejemljivo tradicionalno razmišljanje, da "je slovar metajezikovni odsev dejanske hierarhizirane konceptualizacije sveta" (Vidovič Muha, 2013: 7), kar vodi v razpravljanje o resnicah v okviru slovaropisnega dela – prav nasprotno: slovaropisje mora jasno naslavljati vprašanja, ki so v svojem bistvu ideološka, saj gre za "uravnoteževanje opisa tega, kar prinašajo podatki glede pomena, s tem, na kakšen način 'naj bi bil' v postmoderni vključujoči družbi določen koncept obravnavan in predstavljen" (Moon, 2014: 89). Gre torej za to, da pri slovaropisnem delu končne rešitve preprosto ne morejo biti "samo jezikoslovne; neizogibno morajo biti tudi ideološke" (Moon, 2014: 94). Pomembno je, da se ideološkosti pri slovarskih opisih zavedamo, da odkrito in jasno povemo, da je slovaropisno delo težavno prav zato, ker je tudi ideološko (Gantar, 2015: 399), še posebej pri družbeno občutljivih elementih slovarja.

3. Problemi trenutnega SSSS

SSSS je pripravljen strojno in je trenutno na voljo v prvi, nepregledani različici, v kateri so kot iztočnice in sopomenke navedene leme (brez besednih vrst), pomensko členitev in opis začasno nadomeščajo strojno pripravljene pomenske gruče, slovar pa tudi ne vsebuje slovarskih oznak, razen področnih.

Navedene značilnosti imajo več posledic. Na eni strani se strojno pripravljene iztočnice in sopomenski kandidati pojavljajo brez oznak ali opozoril tudi pri izrazito problematičnih primerih, kot je npr. iztočnica *buzi* s sopomenkami *peder*, *buzerant*, *toplovodar*, *homič*, *poženščen moški*. Na drugi strani je problem potencialno zavajajoča (ne)zastopanost sopomenskega gradiva, npr. vse sopomenke, ki jih najdemo pri iztočnici *zmaj* – *ksantipa*, *vešča*, *strupenjača*, *babura*, *coprnica*, *pošast*, *kričava ženska* – so vezane na ženski spol in imajo izrazito negativno konotacijo, čeprav se beseda rabi tudi za moške in (v drugem pomenu) tudi s pozitivno konotacijo.

Tudi kolokacije in zgledi, ki so namenjeni primerjavi rabe dveh sopomenk, so iz referenčnega korpusa izvoženi strojno in so v slovarju brez oznak. Posledica je lahko sopostavitev pomensko neustreznih podatkov, npr. pri primerjavi besed *ženska – kura* najdemo prekrivne kolokacije [*stara, prava, gola*] *ženska* in [*stara, prava, gola*] *kura* ali *ženska* [*brez glave, v postelji, na odru*] in *kura* [*brez glave, v postelji, na odru*]. Korpusni zgledi načeloma pomagajo razdvoumiti problematične primere, vendar niso na voljo za vse primerjane besede, zgledi, ki so na voljo, pa niso izbrani po vsebinskih kriterijih. To je zlasti problematično pri sovražnem besedišču, npr. kolokacije [*sovražiti, tepsti, ubiti*] *pedra* ali zgledi tipa *In reskiral sem celo, da bi me imel za pedra.*

Pri uporabniško predlaganih sopomenkah ločujemo na eni strani zlonamerne vnose, kot je npr. uporabniški vpis *aljaz* pri iztočnici *gej*. Za takšne primere bi bilo treba določiti natančno uredniško politiko za sprotno obravnavo na ravni vmesnika. Na drugi strani uporabniki zaznamovano besedišče dodajajo kot dejanski sopomenski predlog, npr. pri iztočnici *južnjak*, kjer so uporabniki dodali dolg niz predlogov, mdr. *jugovič*, *južni brat*, *jugič*, *trenirkar, bosanec*, *z juga*. Uredniška naloga je presoditi, kateri predlogi so relevantni za vključitev v slovarsko bazo (in s katerimi oznakami), že uporabnikom pa omogočiti, da problematično besedišče označijo kot tako, da se torej oznaka v vmesniku prikaže istočasno kot dodana sopomenka.

Besedišče, ki je problematično na ravni same leme, je mogoče označiti že v obstoječi različici slovarja. Primeri, pri katerih je oznaka vezana na posamezen pomen besede ali specifičen kontekst rabe, pa zahtevajo predhodno pomensko členitev ter z njo povezan slovaropisni pregled kolokacij in zgledov rabe.

V projektu Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL bomo uresničili dva cilja: (a) identificirali besedišče, ki je problematično na ravni leme in ga označiti po celotnem slovarju SSSS ter (b) dodali v slovarski vmesnik možnost, da uporabniki sami označijo svoje predloge. V nadaljevanju natančneje pojasnjujemo, kako.

4. Identifikacija problematičnega besedišča

4.1. Slovaropisna izhodišča in sistem oznak

Prepoznavanje potencialnega, z vidika družbene občutljivosti problematičnega besedišča temelji na slovaropisnih izhodiščih, ki smo jih pred nekaj leti pripravili za slovarske vire na CJVT UL, prvič pa začeli uporabljati pri izdelavi Velikega slovensko-madžarskega slovarja (Kosem et al., 2018a). V izhodišča je vključeno prepoznavanje elementov sovražnega govora (oznaka *sovražno*), elementov nevljudnosti, žaljivosti (*grobo*) ter elementov negativnega vrednotenja ali konotacije (*izraža negativen odnos*). Omenjene oznake sodijo v širši okvir t. i. sporočanjskih oznak,[2] ki opredeljujejo izraze ali pomene z vidika njihove rabe v sporočanjskem procesu in v situacijah, v katerih sporočanje poteka. V predlaganem slovaropisnem opisu so sporočanjske oznake namenjene označevanju izrazov, z izbiro katerih govorci dosegamo ali želimo doseči določen učinek pri naslovniku. Ta učinek je lahko povzročen s pozitivnim ali negativnim

[2] Celotni sistem označevanja, ki ga razvijamo v okviru virov CJVT UL, poleg sporočanjskih oznak, ki jih notranje členimo na vrednotenjske, registrske in stilne, zajema še nabor pragmatičnih, kontekstualnih, področnih, slovničnih, časovnih in trendovskih oznak ter nabor oznak, vezanih na tuja poimenovanja in prevodne ustreznice.

vrednotenjem, z uporabo v določenem govornem položaju (npr. javnem, nejavnem) ali z namenom izraziti odnos do predmetnosti ali vsebine, ki temelji na določenih družbenih normah, pričakovanjih in odstopanjih od njih. Ta sistem se od tradicionalnega označevanja besed na podlagi odnosa do knjižne norme, kot ga pozna SSKJ (t. i. stilno-zvrstni in ekspresivni kvalifikatorji), ločuje v kvalificiranju besedišča na podlagi sporočanjskega namena oz. učinka, pri čemer izhodišče kvalificiranja ni v opozarjanju na odstop od knjižne norme, pač pa v informiranju glede možnih posledic rabe. S takim sistemom se želimo izogniti morebitnemu kvalificiranju govorca samega, hkrati pa opozoriti na kontekst potencialno problematične rabe v informativnem smislu. To pomeni, da ne želimo uporabnikov slovarja obveščati samo o možnih učinkih rabe grobega in sovražnega besedišča, pač pa pokazati tudi na okoliščine, v katerih je tako rabo mogoče prepoznati.

V slovarskem sistemu oznak označujemo z oznako *sovražno* izraze in pomene, ki so diskriminatorni, ksenofobični, rasistični in homofobični, ki so uperjeni proti predstavnikom skupin ali manjšin na podlagi njihove narodnosti, rase ali etničnega porekla, verskega prepričanja, spola, zdravstvenega stanja, spolne usmerjenosti, invalidnosti, gmotnega stanja, izobrazbe, družbenega položaja ter drugih lastnosti in prepričanj. Z oznako *sovražno* se torej opredeljujemo do vseh izrazov, ki spodbujajo sovraštvo, predsodke ali nestrpnost in s tem lahko predstavljajo – kot je bilo opredeljeno že v razdelku 2 – elemente sovražnega govora.

Na drugi strani z oznako *grobo* označujemo izraze ali pomene, ki so za naslovnika lahko žaljivi, z vidika družbenih in moralnih norm pa neprimerni. Tipično se nanašajo na človeško ali živalsko telo, spolnost, prehranjevanje in izločanje – zlasti torej na tabuizirano predmetnost.

Tretji sklop predstavlja besedišče, ki izraža neodobravanje, nenaklonjenost, posmehljivost ali kritiko do lastnosti posameznikov, predmetov ali dejanj. Z oznako *izraža negativen odnos* želimo tako opozoriti na izraze z izrazito negativno konotacijo ali vrednotenjem, ki so lahko za naslovnika žaljivi ali neprijetni.

4.2. Ročni pregled gradiva

Potencialno problematično besedišče v SSSS smo identificirali z ročnim pregledom iztočnic in sopomenk v slovarju. Na projektu smo se omejili na slovarske (jedrne in bližnje) sopomenke, saj pregled uporabniških predlogov zahteva dodatne uredniške premisleke in bo zato opravljen kasneje s prilagojeno metodologijo. Zaradi obilja gradiva smo delo organizirali v dva koraka: širši pregled, v katerem smo v grobem ločili potencialno problematično in neproblematično gradivo, nato pa natančnejši pregled problematičnih primerov.

Najprej smo iz slovarske baze izvozili nize sopomenk, urejenih na podlagi pomenskih gruč (Krek et al., 2017), npr. *speljati se; izginiti; pobrati se; skidati se; spokati se; spizditi*, pri čemer smo odstranili nize, ki so se glede nabora sopomenk podvajali, in tiste, ki so bili podmnožica kakega drugega niza. Na tak način smo pripravili 65.615 nizov različne dolžine: od posameznih sopomenskih parov do zelo dolgih nizov, ki pa so redki: več kot 30 sopomenk vsebuje le 156 nizov, povprečje je 5 sopomenk na niz.

Čeprav strojno pomensko gručenje ni povsem natančno in se razlikuje od slovaropisne pomenske členitve, tovrstna organizacija podatkov dobro naslovi dva pomembna problema: (a) tak pristop bistveno pohitri pregledovanje, kot bo razvidno v nadaljevanju; (b) presojanje je lahko bolj natančno, saj problematičnost posamezne leme nakazujejo ostale besede v nizu, prim. npr. *nategniti* v nizu *raztegniti; dilatirati; iztegniti; nategniti; pogrniti; razgrniti; razmakniti; razpreti; razprostreti; razviti; napeti; zavlačevati z; razpeti; prolongirati* in v nizu *pokavsati; nategniti; povaljati; porivati; pofukati; pojahati*.

Iz množice 65.615 nizov smo najprej umaknili 24.945 nizov (38,0 %), pri katerih sopomenke vsebujejo področne oznake, npr. *odbojnik, deflektor, ločilnik, membrana, opna, odbojna pregrada, zvočna stena* z oznako *elektrika* (ker so ti podatki terminološke narave, smo predvidevali zanemarljivo nizko vsebnost problematičnega besedišča in smo jih pustili za hiter pregled ob koncu naloge); ostalo je 496 nizov (0,8 %), ki vsebujejo lastnoimenske samostalnike, npr. *Antarktika, antarktično območje, južno polarno območje,* in 40.176 (61,2 %) občnoimenskih nizov, vsi relevantni za ročni pregled.

Podatke so pregledovali študentke in študenti jezikoslovnih smeri, in sicer po trije vzporedno. Pregledovanje je potekalo v okolju *Google Sheets*. Sopomenske nize smo organizirali v vrstice tabele, kjer jim je bilo mogoče pripisati eno od naslednjih odločitev: (1) niz vsebuje sovražno ali grobo besedišče; (2) niz vsebuje besedišče, ki je drugače negativno ali (v določenem pomenu, kontekstu) izraža negativen odnos; (3) z vidika sovražnosti, grobosti, negativnosti je niz neproblematičen. Če so pregledovalci želeli, so lahko opredeliti tudi, da je (4) v nizu kako drugače zaznamovano besedišče, da (5) ne razumejo vseh besed v nizu, lahko pa so vpisali tudi dodaten komentar na svoje odločitve ali podatke.

Kljub ogromni količini podatkov je bila tako oblikovana naloga izvedljiva v relativno kratkem času, saj so študentje lahko odločitev podali takoj, ko so v nizu našli eno samo problematično besedo, natančnejše razmisleke o vrsti zaznamovanosti oz. označevanja posameznih besed pa so prepustili za drugi korak dela s podatki.

4.3. Rezultati ročnega pregleda

Študentske odločitve smo pretvorili v končne odločitve po naslednjem ključu: (1) **sovražno/grobo:** če je vsaj eden od študentov presodil, da se v nizu pojavlja sovražno ali grobo besedišče; (2) **drugače negativno:** kombinacije odločitev "druga negativnost" in "neproblematično" ali (3) **neproblematično:** če so vsi študenti presodili, da je z vidika sovražnosti, grobosti, negativnosti niz neproblematičen. Rezultate prikazuje Tabela 1.

Kategorija končne odločitve	Število nizov v kategoriji	Delež glede na vse pregledano
Sovražno/grobo	1.810	4,5 %
Drugače negativno	12.730	31,3 %
Neproblematično	26.132	64,3 %
Skupaj	40.672	100,0 %

Tabela 1: Številčna zastopanost in delež nizov glede na končno odločitev glede potencialne problematičnosti.

V Tabeli 2 navajamo nekaj nizov s po tremi sopomenkami, ki so jim študentke in študenti pripisali skladne ali različne odločitve. Kot je razvidno, lahko posamezen niz vsebuje raznoliko zaznamovano besedišče,

kot tudi nezaznamovano besedišče. Tabela obenem ponazarja gradivo, ki bo deležno celovite in natančnejše obravnave v zaključnem delu projekta Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL (odločitev 1), podatke, ki so na tak ali drugačen način relevantni za nadaljnje delo (odločitev 2), in gradivo, ki ga z vidika negativne zaznamovanosti ne bomo nadalje obravnavali (odločitev 3).

Niz sopomenk	Študentske in končna odločitev
fukati; porivati; natepavati	111 -> 1
skozlati; izbruhati; zbruhati	111 -> 1
pedrski; buzerantski; toplovodarski	111 -> 1
črnuhinja; zamorka; zamorklja	111 -> 1
pofukanka; prasica; zajebanka	111 -> 1
debilen; bebast; duševno zaostal	121 -> 1
kripelj; pohabljenec; pohabljenka	211 -> 1
kurnik; pajzelj; temačna luknja	222 -> 2
bedastoča; glupost; nesmisel	222 -> 2
eliminirati; likvidirati; usmrtiti	222 -> 2
izmozgano; izčrpano; mršavo	223 -> 2
imenski; nazivni; nominalni	333 -> 3
kopirni papir; indigo; karbon	333 -> 3
zaustaviti se; izklopiti se; izključiti se	333 -> 3

Tabela 2: Primeri nizov s študentskimi odločitvami in končno odločitvijo o nadaljnji obravnavi.

V sodelovanju s študenti bomo v 1.810 nizih z odločitvijo (1) določili besede in zveze, ki so relevantne za slovarsko označevanje. Slednje bo potekalo ob upoštevanju pojavljanja oz. rabe identificiranega besedišča v raznovrstnih kontekstih, s čimer bomo željo po pohitritvi postopka prve selekcije ustrezno nadzorovali in obranili pred črno-belim presojanjem primernosti. Za ponazoritev navajamo nekaj primerov, ki so na seznamu za presojo:

- **sovražno:** *črnuh, črnuhinja, zamorklja, hlapčevski črnec, rdečuh, rdečuhinja, beli prasec, bela prasica, lezba, lezbača, peder, buzerant, pička, prasica, kripelj;*
- **grobo:** *podjebavati, v kurcu, zdrkati, nabrisati, pokavsati, nategniti v rit, pofafati ga, sranje, poscan, fentati, crkniti, razpizden, sfukan, kurbarija, joški.*

V SSSS želimo poleg sovražnega in grobega označiti tudi besedišče, ki izraža negativen odnos. To se najde predvsem v nizih z odločitvijo (2), mestoma pa tudi v (1). Kot problematične so študentje prepoznavali tako izraze (npr. *budala*, *avša*, *bedast*) kot potencialne problematične pomene besed (npr. *nataknjen*, *zabit*, *nasekati*). Prve je mogoče označiti že v trenutni različici slovarja, saj je njihova problematičnost vezana na lemo ne glede na morebitno večpomenskost. V drugem primeru bi oznaka morala biti pripisana pomenu, zato bo označevanje možno šele, ko bo slovar vseboval pomenske členitve. Primeri besedišča, ki ga je mogoče označiti na ravni izraza:

- **izraža negativen odnos:** *trapa, bebav, počasne pameti, lolek, kozlarija, zarukan, špeglarca, luftar, blefer, snobovski, drhal, težakinja, mlatenje prazne slame, avša, otročaj.*

V "drugače negativno" so raznorodni primeri, saj so poleg zaznamovanih izrazov in pomenov študentje označevali tudi besedišče, ki poimenuje **negativne vsebine in predmetnost**. Gre zlasti za poimenovanja agresivnega obnašanja: *uničiti*, *dotolči*, nekaterih osebnih lastnosti: *pokvarjen*, *hudoben*, *ničvreden*, *grozljiv*, *grd*, *apatičnost, pokvarjenost*; videza, stanja: *neurejenost*, *razdejanje*, *zanikrnost* itd. V slovarju večina teh besed ne potrebuje oznake. Čeprav besed ne bomo označevali, so seznami tovrstnega besedišča pomemben rezultat ročnega pregleda, saj so koristni za različne druge namene na področju slovaropisja in strojne obdelave jezika, npr. za filtriranje gradiva z negativnim pomenom iz jezikovnih iger ali učnih gradiv, strojno pripisovanje sentimenta ipd.

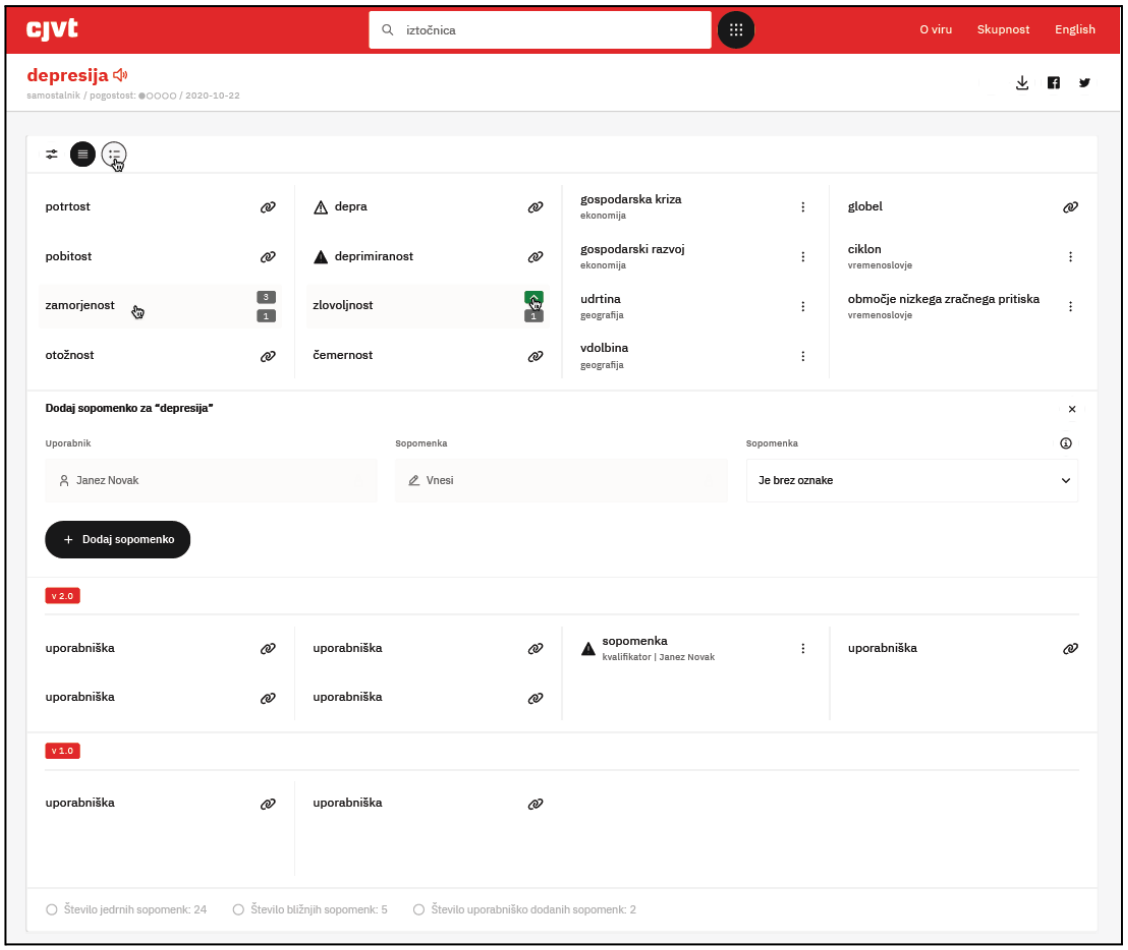
5. Vizualizacija v vmesniku SSSS 2.0

V slovarskem vmesniku SSSS 2.0 bomo na besedišče, o katerem razpravljamo tu, opozorili s kombinacijo opozorilne ikone in daljšega pojasnila, ki se bo izpisalo ob kliku nanjo. Trenutno rešitev, ki jo bomo po potrebi še nadgradili, kaže Tabela 3. Namenoma smo se odrekli pripisovanju (eno-)besednih oznak, saj bi te pri označevanju (mestoma tudi homonimnih) lem lahko vodile v napačno interpretacijo podatkov. Pri pomensko členjenih geslih bodo oznake seveda pripisane posameznim pomenom, pri pomensko nečlenjenih geslih pa bo kombinacija ikone in pojasnila omogočila, da je problematično besedišče na prvi pregled zelo opazno, pojasnilo pa je lahko daljše in vsebuje informacije o možnem učinku na naslovnika oz. možnih posledicah rabe označene besede.

Oznaka	Ikona	Pojasnilo
Sovražno	⚠	Z uporabo besede lahko izražamo sovražni, nestrpni odnos do posameznika ali družbene skupine.
Grobo	⚠	Zaradi družbenih in moralnih norm se marsikateremu uporabniku jezika beseda lahko zdi groba ali neprimerna. Uporaba lahko povzroči nelagodje, razburi ali užali.
Izraža negativen odnos		Beseda lahko ni nevtralna. Z uporabo besede se lahko posmehujemo, izražamo neodobravanje ali kritiko do nekaterih lastnosti posameznikov, predmetov ali dejanj.

Tabela 3: Predvidene ikone in izhodiščna različica pojasnil za označevanje besedišča v SSSS.

Slika 1 kaže oblikovalski predlog vmesnika SSSS 2.0, kakršen je na voljo v času priprave prispevka. Slovarske informacije na sliki so provizorične. Slika ponazarja, kakšna bo vizualizacija pri bližnjih in jedrnih (pri *depra* in *deprimiranost*) ter pri uporabniških sopomenkah (pri *sopomenka*). Razvidne so tudi nekatere druge novosti, npr. delitev uporabniških sopomenk glede na slovarsko verzijo, v kateri so bile predlagane, ter možnost dodajanja slovarskih oznak ob predlagane sopomenke.



Slika1. Oblikovalski predlog vmesnika SSSS 2.0 (vsebina je provizorična).

6. Uporabniško dodajanje oznak

Nedavno izvedena raziskava o odnosu uporabniške skupnosti do SSSS, v kateri je sodelovalo 671 anketirancev, je pokazala naklonjenost do večine novosti, ki jih prinaša slovar, npr. stalno posodabljanje, strojni postopki, digitalni format, kolokacijski podatki, povezave na korpus, uporabniško vključevanje (Arhar Holdt, 2020: 470). Med problematičnimi značilnostmi sta bili izpostavljeni nezanesljivost (strojno pridobljenih) podatkov in primanjkljaj slovarskih oznak tako pri jedrnih in bližnjih sopomenkah kot pri uporabniško dodanih. To, da ni oznak, je motilo 37 % sodelujočih (ibid.: 472).

V trenutnem slovarskem vmesniku nekateri uporabniki in uporabnice težavo rešujejo tako, da oznako ali kako drugo pojasnilo v oklepaju pripišejo ob svoj sopomenski predlog, npr. *babica – nona (lokalno), bojazljivec – pezde (vulg.), Italijanka – makaronarka (slabš.)*. Kot omenjeno v poglavju 3, pa večina predlaganih sopomenk oznake nima.

Skladno z uporabniškimi željami in potrebami želimo nadgraditi protokol dodajanja sopomenk, da bodo predlagani besedi ali zvezi uporabnice in uporabniki lahko dodali tudi slovarsko oznako oz. oznake. Privzeta izbira bo, da je predlog "brez oznake" , ostale možnosti bodo na voljo v spustnem meniju (Slika 1). V različici SSSS 2.0 bodo na klik na voljo oznake *sovražno*, *grobo* in *izraža negativen odnos*, poleg tega pa bomo ponudili okence, v katerega bo mogoče vtipkati morebitno drugo oznako.

Pomen in raba oznak *sovražno*, *grobo* in *izraža negativen odnos* bo razložena in ponazorjena s primeri, s čimer bo lahko dosežena določena stopnja enotnosti uporabniškega označevanja (informacije bodo na voljo na klik, gl. ikono (i) na Sliki 1). Predvideno pa je, da bodo uporabniki oznake mestoma interpretirali in uporabljali drugače, kot bi jih slovaropisci. Vse dodane oznake bodo (skupaj z dodanimi sopomenkami) preverjene in označene sopomenke bodo dragoceno gradivo ne le za dopolnitev odprto dostopne slovarske baze sopomenk, ampak tudi za analize širšega dojemanja označevalnega sistema ter dometa in meja oznak. Prav tako pomemben uvid bodo ponudile ročno vpisane oznake, ki jih bomo analizirali z vidika vsebine in pogostosti ter uporabili izsledke za nadaljnji razvoj slovarja.

7. Sklep in nadaljnje delo

Sodobno slovaropisno delo ima ob zavedanju ideološkosti, vključevanju novih pristopov, uporabi tehnologije, moči množic itd. danes veliko možnosti, da tudi vprašanja označevanja konotacije naslavlja na novo in zanj pripravlja inovativne rešitve (Gorjanc, 2017: 154).

V prispevku smo opisali, kako poteka obravnava sovražnega in grobega besedišča v SSSS in katere spremembe so v načrtu za različico 2.0, ki bo objavljena jeseni 2022. Rešitve naslavljajo dve pomembni značilnosti SSSS: njegovo strojno izdelanost in odprtost, da pri razvoju slovarja sodeluje tudi uporabniška skupnost. V novi različici slovarja bodo sovražnemu in grobemu besedišču pripisane slovarske oznake oz. opozorilne ikone s pojasnili o možnih posledicah rabe in dodana bo možnost, da uporabniki pripišejo oznako svojim predlogom sopomenk.

Ker vse težave trenutnega SSSS niso enostavno in hitro rešljive, želimo slovarske uporabnike bolje opozoriti na trenutne omejitve. Čeprav je metodologija priprave SSSS pojasnjena v razdelku *O viru*, pri samih iztočnicah ni izrecnih opozoril, da je slovar pripravljen strojno, in to na vseh ravneh: sopomenke, kolokacije, korpusni zgledi, kar lahko vodi v napačne interpretacije slovarske vsebine. V naslednji različici SSSS želimo zato uvesti indikator stopnje gesla[3] in dodati v predstavitev slovarja opozorila o dometu in posledicah metodologije ter razlago korakov, po katerih se slovar razvija.

Prepoznano sovražno in grobo besedišče bo koristno tudi pri izdelavi drugih virov, kjer se za pomene izbirajo reprezentativne kolokacije in zgledi. Pri izdelavi novih gesel za Kolokacijski slovar sodobne slovenščine (Kosem et al., 2018b) npr. že zdaj pri pripravi podatkov (pred slovaropisno analizo) označujemo kolokacije, ki vsebujejo sovražno in grobo besedišče, pa tudi besedišče, ki izraža negativen odnos. Tako slovaropiske in slovaropisce opozorimo na potencialno problematične kolokacije in posledično pohitrimo delo oz. se izognemo vključevanju problematičnih vsebin. Seznami problematičnega besedišča, ki jih uporabljamo trenutno, so pripravljeni *ad hoc* iz odprto dostopnih jezikovnih virov in precej krajši od seznamov, ki bodo (lahko) nastali na osnovi predstavljenega dela.

Kot smo poudarili v prispevku, je izražanje negativnega odnosa večkrat vezano na posamezen pomen besede, zato bo velik del naloge izvedljiv šele ob pripravi pomensko členjenih gesel. Pri pomenski členitvi in nadaljnjem označevanju gradiva SSSS bomo uporabili metodologijo, ki jo razvijamo pri izdelavi Velikega slovensko-madžarskega slovarja (Kosem et al., 2018a) in podatke oz. informacije, ki so na voljo v obstoječih odprto dostopnih virih za slovenščino. Preizkus prenosa metodologije bomo izvedli že pod okriljem projekta Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL, kjer je med cilji tudi nadgradnja SSSS z 2.000 pomensko členjenimi gesli, ki bodo imela slovaropisno pregledane in razvrščene sopomenke, kolokacije ter korpusne zglede.

V nadaljnje premisleke glede sovražnega in grobega besedišča znotraj koncepta odzivnega slovarja bi bilo smiselno celoviteje vključiti vidike okoliščin rabe. Zanimivo bi bilo denimo obravnavati zaznavanje in presojanje sovražnosti, grobosti v različnih tipih besedil, npr. medijskih. Ob tem se odpira tudi vprašanje formalnosti in neformalnosti položajev, na katere se ta presoja nanaša: ali posega na vse ravni izražanja ali gre zgolj za formalne, javne položaje in ali je neodvisna od generacijske ali kake druge pripadnosti presojevalca.

8. Zahvala

Projekt Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL v letih 2021–2022 financira Ministrstvo za kulturo Republike Slovenije. Raziskovalna programa št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik) in št. P6-0215 (Slovenski jezik – bazične, kontrastivne in aplikativne raziskave), sofinancira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

[3] Po zgledu Kolokacijskega slovarja sodobne slovenščine (KSSS), ki z ikono petstopenjske piramide uporabniku na jasen in ekspliciten način posreduje informacijo o razvoju ter različnih stopnjah izdelanosti slovarskih gesel (Kosem et al., 2018b).

9. Literatura

Špela Arhar Holdt. 2020. How Users Responded to a Responsive Dictionary: The Case of the Thesaurus of Modern Slovene. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46(2): 465–482. https://doi.org/10.31724/rihjj.46.2.1

Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Apolonija Gantar, Vojko Gorjanc, Bojan Klemenc, Iztok Kosem, Simon Krek, Cyprian Laskowski in Marko Robnik Šikonja. 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 18th Euralex International Congress: Lexicography in Global Contexts*, str. 401–410. Znanstvena založba Filozofske fakultete, Ljubljana. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Polona Gantar. 2015. *Leksikografski opis slovenščine v digitalnem okolju*. Znanstvena založba Filozofske fakultete, Ljubljana. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/download/62/138/2602-1?inline=1

Vojko Gorjanc. 2005. Neposredno in posredno žaljiv govor v jezikovnih priročnikih: diskurz slovarjev slovenskega jezika. *Družboslovne razprave*, 21(48): 197–209.

Vojko Gorjanc. 2017. *Nije rečnik za seljaka*. Biblioteka XX vek, Beograd.

Geoffrey Hughes. 2011. *Political Correctness: A History of Semantics and Culture.* Wiley-Blackwell, MA.

Boris Kern. 2015. Politična korektnost v slovaropisju. V: D. Zuljan Kumar in H. Dobrovoljc, ur., *Zbornik prispevkov s simpozija 2013*, str. 144–154, Založba Univerze, Nova Gorica.

Iztok Kosem, Júlia Čeh Bálint, Vojko Gorjanc, Anna Kolláth, Attila Kovács, Simon Krek, Sonja Novak-Lukanovič in Jutka Rudaš. 2018a. *Osnutek koncepta novega velikega slovensko-madžarskega slovarja*. Univerza v Ljubljani, Filozofska fakulteta, Ljubljana. https://www.cjvt.si/komass/wp-content/uploads/sites/17/2020/08/Osnutek-koncepta-VSMS-v1-1.pdf

Iztok Kosem, Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej in Cyprian Laskowski. 2018b. Kolokacijski slovar sodobne slovenščine. V: D. Fišer in A. Pančur, ur., *Jezikovne tehnologije in digitalna humanistika.* Znanstvena založba Filozofske fakultete, Ljubljana. http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Kosem-et-al_Kolokacijski-slovar-sodobne-slovenscine.pdf

Simon Krek, Cyprian Laskowski in Marko Robnik Šikonja. 2017. From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. V: I. Kosem et al., ur., *Proceedings of eLex 2017: Lexicography from Scratch*, str. 93–109, Leiden, Netherlands. https://elex.link/elex2017/wp-content/uploads/2017/09/paper05.pdf

Simon Krek, Cyprian Laskowski, Marko Robnik Šikonja, Iztok Kosem, Špela Arhar Holdt, Polona Gantar, Jaka Čibej, Vojko Gorjanc, Bojan Klemenc in Kaja Dobrovoljc. 2018. Thesaurus of Modern Slovene 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1166

Nataša Logar, Nina Perger, Vojko Gorjanc, Monika Kalin Golob, Neža Kogovšek Šalamon in Iztok Kosem. 2020.

Raba slovarjev v slovenski sodni praksi. *Teorija in praksa*, 57:89–108. https://www.fdv.uni-lj.si/docs/default-source/tip/tip_pos_2020_logar_idr.pdf?sfvrsn=0

Rosamund Moon. 2014. Meanings, Ideologies, and Learners' Dictionaries. V: A. Abel et al., ur., *Proceedings of the XVI EURALEX International Congress: The User in Focus*, str. 85–105, Bolzano/Bozen. Institute for Specialised Communication and Multilingualism. https://euralex.org/elx_proceedings/Euralex2014/euralex_2014_004_p_85.pdf

Andrej Motl in Veronika Bajt. 2016. *Sovražni govor v Republiki Sloveniji: Pregled stanja.* Mirovni inštitut, Ljubljana. https://dlib.si/stream/URN:NBN:SI:DOC-F2YZP2RB/c117f4c6-8fe9-437d-8c64-5b7987a856b6/PDF

Brankica Petković in Neža Kogovšek Šalamon. 2007. *O diskriminaciji: Priročnik za novinarje in novinarke*. Mirovni inštitut, Ljubljana. https://www.mirovni-institut.si/wp-content/uploads/2014/08/Prirocnik-o-diskriminaciji-final-all.pdf

Dušan Rebolj. 2008. Uporabnejša opredelitev politične korektnosti. V: S. Autor in R. Kuhar, ur., *Politična (ne)korektnost.* Mirovni inštitut, Ljubljana, str. 4–15. https://www.mirovni-institut.si/wp-content/uploads/2014/08/nestrpnost-6.pdf

SSKJ. 2014. *Slovar slovenskega knjižnega jezika: Uvod*. Druga, dopolnjena in deloma prenovljena izdaja. Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, Ljubljana. https://fran.si/130/sskj-slovar-slovenskega-knjiznega-jezika

Ada Vidovič Muha. 2013. *Moč in nemoč knjižnega jezika*. Znanstvena založba Filozofske fakultete, Ljubljana.

Izdelava in analiza digitalizirane zbirke paremioloških enot

Saša Babič*, Tomaž Erjavec†
* Inštitut za slovensko narodopisje ZRC SAZU
Novi trg 2, 1000 Ljubljana
sasa.babic@zrc-sazu.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

Članek obravnava digitaliziranje zbirke slovenskih pregovorov Inštituta za slovensko narodopisje ZRC SAZU. Zbirka je nastajala od leta 1947 dalje, digitalizacija pa se je začela v samem začetku 21. stoletja z iniciativo Marije Stanonik. V predstavljenem delu smo izhajali iz Excel razpredelnic paremioloških enot in virov, iz katerih smo najprej izločili neustrezne enote in neuporabljene vire. Nato smo tabeli pretvorili v zapis TEI in pregovore avtomatsko jezikoslovno označili. Tu so bile besede posodobljene, lematizirane, oblikoskladenjsko označene, povedi pa skladenjsko razčlenjene po formalizmu Universal Dependencies. Kanonični zapis TEI smo pretvorili v več izvedenih formatov in zbirko objavili pod odprto licenco na repozitoriju CLARIN.SI, kjer jo je mogoče prevzeti, in na konkordančnikih CLARIN.SI, ki so primerni za jezikoslovne analize zbirke. V članku orišemo tudi način iskanja po zbirki v konkordančnikih, ki omogočajo temeljitejšo etnolingvistično in semiotično raziskavo.

Creation and analysis of a digitised collection of Slovenian paremiological units

The article discusses the digitization of the collection of Slovenian proverbs from the Institute of Slovenian Ethnography ZRC SAZU. The collection was created from 1947, and its digitization began at the start of the 21st century on the initiative of Marija Stanonik. The departure point of the presented were two Excel spreadsheets with paremiological units and their bibliographical sources, from which we removed inappropriate units, and unused sources. The two spreadsheets were then converted to a TEI encoding, and the paremiological units automatically linguistically annotated: words were modernised, lemmatised, morphosyntactically annotated and the sentences syntactically parsed according to the Universal Dependencies formalism. We converted the canonical TEI encoding into several derived formats and published the collection under an open licence on the CLARIN.SI repository, where it can be downloaded, and on the CLARIN.SI concordancers, which allow for linguistic analyses of the collection. The paper also outlines searching the collection in the concordancers, which enables detailed ethnolinguistic and semiotic research.

1. Uvod

Jezik je ohranjevalec in nosilec kulture, s katerim človeštvo ustvarja in vključuje refleksije o samem sebi (Pitkin, 1972; Bartmiński, 2005; Tolstaja, 2015). Ena od najpogosteje rabljenih jezikovnih oblik so pregovori oz. paremiološke enote.

Paremiološke enote ali pregovori v širšem pomenu so eden od najkrajših žanrov slovstvene folklore; pregovore lahko opišemo kot relativno stalne povedi, ki jih uvrščamo med kratke folklorne obrazce. Pogosto so označeni z besednimi zvezami, kot »modrost ljudstva« (Mieder, 1993), »stara modrost« in »poezija vsakdanjega jezika« (Matičetov, 1956). V vsakem primeru lahko trdimo, da so pregovori »skrčeni moralno-etični obrazci določene skupnosti; so neke vrste tradicionalni stereotipi njenega samozavedanja in samoidentifikacije, bili so iz generacije v generacijo prenašani jezik vsakdanje kulture« (Kržišnik, 2008: 38). Prav zato velja, da so pregovori kratki stereotipi na sentenčni ravni s prenesenim ali generalizirajočim pomenom ter so načeloma splošno znani (Grzybek, 2012).

Pregovori so kulturna besedila z velikim semantičnim potencialom (Grzybek, 2015), saj gre za »zaključene misli« (Mlacek, 1983: 131), vendar pa se ne razlikujejo le po besedilu, temveč tudi glede na teksturo in kontekst (Dundes, 1965). Zaradi prozodičnih značilnosti si jih je lažje zapomniti, dandanes pa zato ponujajo možnosti za nadaljnjo uporabo, na primer pri oglaševanju, sodobnem prenosu mnenj, grafitih ali modifikacijah v različnih medijih. Semiotična kompleksnost pregovorov in prepletenost med sintaktično (kratkost), pragmatično (prenašanje skozi različne generacije) in semantično (stereotipno, splošno znanje) razsežnostjo ponujajo raziskovanje pregovorov kot kulturni znak, ki ohranja zgodovino kulture oz. družbe, hkrati pa sprejema nove funkcije, ki širijo in porajajo nove kontekste. Prav zato so paremiološke enote oz. pregovori označeni za narodni zaklad, neprecenljivo modrost in dediščino prednikov, in ne preseneča, da so (bili) predmet sprotnega terenskega zapisovanja ali celo namenskega zbiranja (Arewa in Dundes, 1966; Stanonik, 2015) ter analiz rabe (Meterc, 2021).

Inštitut za slovensko narodopisje ZRC SAZU je sistematično gradil arhiv različnih žanrov slovstvene folklore, v sklopu katerega je nastajala tudi zbirka pregovorov. Ti so bili zabeleženi na kartotečnih listkih ali v tematskih arhivskih mapah. V začetku 21. stoletja se je pojavila potreba po digitalizaciji gradiva, ki bi omogočala lažje delo z gradivom.

Pri projektu *Tradicionalne paremiološke enote v dialogu s sodobno rabo* (2020–2023) smo predvideli združitev etnolingvističnih pristopov in semiotike z namenom diahronega vpogleda v družbo s pomočjo pregovorov. Da bi bila analiza temeljitejša, je pomemben del projekta pretvorba gradiva v sprejemljivo obliko za računalniško besedilno analizo.

V članku opišemo pripravo in jezikoslovno označevanje digitalizirane zbirke pregovorov, ki je sedaj dostopna na repozitoriju in konkordančnikih CLARIN.SI, ter uporabo digitalizirane zbirke v namene etnolingvistične obravnave paremioloških enot. Na koncu podamo zaključke in načrte za nadaljnje delo.

2. Priprava gradiva

Inštitut za slovensko narodopisje (ISN) ZRC SAZU v arhivu hrani folklorno gradivo v analogni obliki, tj. ročno napisani, natipkani ali natisnjeni na kartotečnih listkih, v arhivskih predalih in omarah. Težnja po digitalizaciji folklornega gradiva se je najprej začela pri pregovorih, za katere je Marija Stanonik že leta 1997–1999 pridobila projekt *Slovenski pregovori in rekla* (Stanonik, 1996), v katerem je začela širiti arhivsko zbirko pregovorov na ISN. Z mislijo na digitalizacijo je nadaljevala v projektih *Informatizacija neoprijemljive dediščine za etnologijo in folkloristiko* (2005–2008) (Stanonik, 2004) in *Slovenski pregovori kot kulturna dediščina: klasifikacija in redakcija korpusa* (2010–2013) (Stanonik, 2009; Stanonik, 2015). Gradivo je bilo dodano k obstoječi zbirki v računalniškem prepisu, sprva v programu Word, pozneje v programu Excel, kar je predstavljalo temelj, na katerem smo lahko izvedli pretvorbo v druge digitalne formate.

2.1. Priprava gradiva v razpredelnicah

V urejanje smo dobili dve excelovi tabeli: prva je vsebovala 59.543 večinoma paremioloških enot, druga pa 2.742 virov teh enot. Tabeli sta bili povezani s kodo, ki je bila določena viru. Ob pregledu gradiva smo ugotovili, da precej enot ne spada v paremiološki nabor; te smo ročno izločili (uganke, dele folklornih pesmi, pozdrave, frazeme ipd.), pri pregledu virov smo ročno izločili vse tiste, ki niso bili navedeni ob paremioloških enotah. Poleg tega so nekatere paremiološke enote vsebovale širši kontekst, ki smo ga ročno izbrisali; tako smo dobili poenoteno obliko samostojnih paremioloških enot. Pri vremenskih paremioloških enotah se je pojavil problem pojasnjevanja svetniškega poimenovanja dnevov in praznikov: v originalnem zapisu (časopisi, koledarji, zvezki ipd.) so bili navedeni kot pojasnilo, npr. *Če je na Velike maše dan [15. avgust] lepo vreme, potem bo ozimna pšenica lepa*; *Če na ta dan [Florijanovo, 4. maj] dež gre, potlej ga celo leto manjka*. V excelovi tabeli, ki predstavlja del Inštitutskega arhiva, smo te pustili zabeležene v oglatem oklepaju.

Po ročnem urejanju smo Excel dokumente združili z OpenRefine[1] in tako poenotili korektorske opombe in kategorije označevanja pregovorov. Osnovne popravke smo vnesli tudi pri preverjanju shematiziranih vnosov (npr. navajanje virov, odstranjevanje presledkov na koncu besedil v posameznih celicah ipd.). Sledil je prenos podatkov v delovno bazo SQLite[2], kjer so potekali popravki preprostih slovničnih napak in zatipkov (velike začetnice, dvojni presledki, nepravilna raba ločil ipd.) ter zaznava uporabljenih črkovnih naborov, kjer gre izpostaviti nestandardizirane zapise dajnčice, metelčice, bohoričice in gajice. Pregovore so namreč začeli prepisovati v računalniško obliko že v začetku 21. stoletja, ko nabor črkovnih znakov še ni bil tako pester in so prepisovalci reševali zagate z različnimi zapisi z improviziranim izborom znakov. Po osnovnih popravkih paremioloških enot smo nadaljevali z iskanjem enakih oz. podvojenih enot in odstranjevali dvojnike, pri čemer smo vse vire dodali k eni paremiološki enoti. Ob koncu urejanja smo podatke izvozili v format TSV (tab-separated values), ki je bil izhodišče za izdelavo korpusa.

Gradivo je tako po ročnem in strojnem urejanju vsebovalo 36.349 relativno enotno urejenih paremioloških enot ter 2.515 virov.

Razpredelnica z viri vsebuje za vsak bibliografskimi vir njegov identifikator, identifikator z izvornega seznama virov, zaporedno številko vira (ki tudi združuje vire, ki spadajo v nadrejeno enoto), letnico izida (in letnico prvega izida, kjer se ta razlikuje), ime vira (avtor, naslov) ter kategorizacijo vira v 18 kategorij, npr. Leposlovje in literarjenje, Muzejske zbirke, Periodika – pratike in koledarji, Ustni viri itd.

Razpredelnica s paremiološkimi enotami vsebuje identifikator enote, zaporedno številko iz izvornega seznama enot, seznam identifikatorjev virov skupaj s številko strani, na kateri je enota v viru omenjena, diplomatično transkripcijo enote (torej zapis enote, kot se pojavi v viru) in kritično transkripcijo enote, ki enote, zapisane v bohoričici, transkribira v gajico. Tako ima npr. enota PREG-00-00001 zaporedno številko 1, seznam virov bib14.1: 202; bib23.1: 51; bib7.1: 524, diplomatično transkripcijo »Bres muje ſe zhreul ne obuje.« in kritično transkripcijo »Brez muje se črеul ne obuje.«

2.2. Zapis TEI

V naslednjem koraku smo podatke iz dokumentov TSV pretvorili v zapis, ki je bolj primeren za hrambo kot tudi za nadaljnje obdelave, in sicer XML s shemo po priporočilih iniciative za kodiranje besedil TEI (TEI Consortium, 2020). Celotna zbirka je bila formirana kot en TEI dokument (element <TEI>) s kolofonom (element <teiHeader>) in besedilnim delom (<text>).

Kolofon vsebuje bibliografske in druge metapodatke o zbirki, kot je npr. taksonomija kategorizacije virov. V opisu vira (<sourceDesc>) vsebuje tudi celoten seznam virov paremioloških enot; zapis je ilustriran v sliki 1.

Besedilni del vsebuje paremiološke enote, vsako s svojim identifikatorjem, diplomatični in kritični prepis ter seznam njihovih virov; zapis ilustriramo v sliki 2.

2.3. Posodabljanje besed in drugo jezikoslovno označevanje

Precejšnjo težavo za uporabo izdelane zbirke je predstavljal zapis v arhaični slovenščini, ki oteži iskanje po pregovorih, kot tudi njihovo nadaljnjo analizo. Oteženo je tudi avtomatsko jezikoslovno označevanje zbirke, saj orodja za jezikoslovno označevanje delujejo dobro le na sodobni standardni slovenščini.

Za posodabljanje zbirke smo uporabili odprtokodno[3] orodje za normalizacijo cSMTiser (Scherrer in Ljubešić, 2016), ki temelji na principu statističnega strojnega prevajanja in orodju Moses (Koehn, 2010). cSMTiser smo naučili posodabljanja na ročno posodobljene korpusu slovenščine goo300k (Erjavec, 2016), podobno, kot smo že pred tem naredili za posodabljanje zbirke slovenskih romanov v okviru korpusa ELTeC (Schöch et al., 2021). Z orodjem smo nato normalizirali kritični prepis, pri čemer orodje sicer približa zapis besed sodobni slovenščini, dela pa tudi napake (npr. besedo »črеul« prevede v »čevlj« namesto »čevelj«).

[1] https://openrefine.org/
[2] https://www.sqlite.org/
[3] https://github.com/clarinsi/csmtiser

```
<ab xml:id="PREG-00-00001" n="1">
    <seg xml:lang="sl-bohoric" xml:id="PREG-00-00001.dipl" type="dipl">Bres muje
        fe zhreul ne obuje.</seg>
    <seg xml:lang="sl" xml:id="PREG-00-00001.crit" type="crit">Brez muje se čreul
        ne obuje.</seg>
    <bibl corresp="#bib14.1">
        <biblScope unit="page">202</biblScope>
    </bibl>
    <bibl corresp="#bib23.1">
        <biblScope unit="page">51</biblScope>
    </bibl>
    <bibl corresp="#bib7.1">
        <biblScope unit="page">524</biblScope>
    </bibl>
</ab>
```

Slika 1: Primer virov paremioloških enot v zapisu TEI.

```
<listBibl xml:lang="sl">
    <bibl xml:id="bib1.1" n="15" corresp="#bibl.dictionary">Bohorič,
        Adam, <date type="reprint" when="1970">1970</date>
        (<date type="firstEdition" when="1584">1584</date>):
        Arcticae horulae succisivae. Faksimile. Mladinska knjiga.
        Ljubljana.</bibl>
    <bibl xml:id="bib2.1" n="3" corresp="#bibl.dictionary">Kastelec, Matija,
        (<date from="1680" to="1685">1680-1685</date>): Dictionarivm
        Latino-Carniolicvm. NUK, rokopisna zbirka, MS 803. Ljubljana.</bibl>
    <bibl xml:id="bib54.2" n="901/9" corresp="#bibl.yearbook">Erjavec, Fran,
        (<date when="1880">1880</date>): Iz potne torbe. V: Letopis Matice
        Slovenske za leto 1880. Matica Slovenska. Ljubljana.</bibl>
```

Slika 2: Primer kodiranja paremiološke enote v zapisu TEI.

```
<seg xml:id="P1.crit.norm.ana" type="crit.norm.ana" xml:lang="sl">
  <s xml:id="P1.crit.s1">
    <w ana="mte:Sg" msd="UPosTag=ADP|Case=Gen" lemma="brez" xml:id="P1.crit.s1.t1">Brez</w>
    <w ana="mte:Ncfsg" msd="UPosTag=NOUN|Case=Gen|Gender=Fem|Number=Sing" lemma="muja" xml:id="P1.crit.s1.t2">muje</w>
    <w ana="mte:Px------y" msd="UPosTag=PRON|PronType=Prs|Reflex=Yes|Variant=Short" lemma="se" xml:id="P1.crit.s1.t3">se</w>
    <w norm="čevlj" ana="mte:Ncmsn" msd="UPosTag=NOUN|Case=Nom|Gender=Masc|Number=Sing" lemma="čevlj" xml:id="P1.crit.s1.t4">čreul</w>
    <w ana="mte:Q" msd="UPosTag=PART|Polarity=Neg" lemma="ne" xml:id="P1.crit.s1.t5">ne</w>
    <w ana="mte:Vmer3s" msd="UPosTag=VERB|Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin"
      lemma="obuti" join="right" xml:id="P1.crit.s1.t6">obuje</w>
    <pc ana="mte:Z" msd="UPosTag=PUNCT" xml:id="P1.crit.s1.t7">.</pc>
    <linkGrp type="UD-SYN" targFunc="head argument" corresp="#P1.crit.s1">
      <link ana="ud-syn:case" target="#P1.crit.s1.t2 #P1.crit.s1.t1"/>
      <link ana="ud-syn:obl" target="#P1.crit.s1.t6 #P1.crit.s1.t2"/>
      <link ana="ud-syn:expl" target="#P1.crit.s1.t6 #P1.crit.s1.t3"/>
      <link ana="ud-syn:nsubj" target="#P1.crit.s1.t6 #P1.crit.s1.t4"/>
      <link ana="ud-syn:advmod" target="#P1.crit.s1.t6 #P1.crit.s1.t5"/>
      <link ana="ud-syn:root" target="#P1.crit.s1 #P1.crit.s1.t6"/>
      <link ana="ud-syn:punct" target="#P1.crit.s1.t6 #P1.crit.s1.t7"/>
    </linkGrp>
  </s>
</seg>
```

Slika 3: Primer kodiranja jezikoslovno označene paremiološke enote v zapisu TEI.

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 »čevlja«:

- oblikoskladenjsko oznako po priporočilih MULTEXT-East (Erjavec, 2012), npr. »Ncmsg« za »Noun Type=common Gender=masculine Number=singular Case=genitive« (pri čemer obstaja tudi ekvivalentna oznaka v slovenščini, tu »Somer« in njena razširitev v pare lastnost=vrednost);
- lemo oz. osnovno obliko besede, tu »čevelj«;
- oblikoskladenjske oznake po sistemu Universal Dependencies za slovenski jezik (Dobrovoljc et al., 2017), npr. »NOUN Case=Gen Gender=Masc Number=Sing«. Te oznake so sicer podobne oznakam MULTEXT-East, vendar z drugače izpisanim naborom lastnosti in vrednosti, občasno se pa od njih tudi sistemsko razlikujejo;
- odvisnostno skladenjsko razčlenitvijo povedi po sistemu Universal Dependencies.

Jezikoslovno označena različica posamezne paremiološke enote je bila dodana v zapis TEI po njenih besedilnih zapisih; format je ilustriran v sliki 3. V različici korpusa, ki vsebuje posodobljene in jezikovno označene enote, je dopolnjen tudi kolofon s taksonomijo skladenjskih oznak Universal Dependencies in z opisom uporabljenih orodij.

[4] https://github.com/clarinsi/classla

2.4. Objava zbirke

Zbirko smo objavili na dva načina. Za prevzem je dostopna na repozitoriju CLARIN.SI (Babič et al., 2022) pod odprto licenco CC BY. Poleg obeh različic zbirke (brez in z jezikoslovno označenimi enotami) v formatu TEI je tam na voljo tudi v izvedenem formatu TSV, torej kot razpredelnici z viri in enotami, in v t. i. vertikalnem formatu, ki služi kot vhodni format za konkordančnike CLARIN.SI.

Zbirka je dostopna tudi na konkordančnikih noSketch Engine in KonText CLARIN.SI. Prek teh dveh storitev je omogočen analitični vpogled v digitalizirano zbirko.

3. Analiza gradiva

Zbirka ima v diplomatičnem zapisu zavedenih 36.066 paremioloških enot (283 jih je v kritičnem prepisu). Največ paremioloških enot je izpisanih iz že obstoječih zbirk pregovorov (10.187 enot) ter iz leta 1974, tj. zbirke pregovorov *Pregovori in reki na Slovenskem*, ki jo je uredil Etbin Bojc (4.884 enot). Treba je upoštevati dejstvo, da je Bojc precej paremioloških enot zbral tudi iz že prej obstoječih zbirk (npr. Kocbek (1887), Kocbek-Šašelj (1934) ter starejše slovnice in slovarji), velja pa njegova zbirka za prvo sodobnejšo. Najstarejši pregovori v zbirki so iz leta 1587, in sicer iz *Predgovora k Postili* Jurija Juričiča.

Zbirka paremioloških enot ISN vsebuje precej enot iz slovnic in slovarjev, kar pomeni, da so bile te enote zapisane kot izolirane entitete, brez konteksta. Poleg tega, navedeno ne izpriča dejanskega poznavanja in rabe paremioloških enot, kot ga lahko predvidevamo pri zbiranju paremiološkega gradiva na terenu ali iz tiskanih besedil, v katerih avtor predvideva poznavanje posameznih paremioloških enot in s tem bralčevo razumevanje napisanega. Navedeno je v folkloristiki pomemben del raziskav in analiz, saj razkriva tudi konceptualni in etnolingvistični vidik folklornega gradiva. Če predvidimo dobro poznavanje posameznega pregovora (npr. *Brez muje se še čevelj ne obuje*), lahko predvidimo tudi konceptualno ozadje in etnolingvistično sliko, ki nam jo tovrstno gradivo lahko ponudi. Za takšen vpogled se poslužujemo ne le etnolingvističnega pristopa (povezovanja jezikoslovja in etnologije s poudarkom na stereotipni predstavi pojava), temveč tudi semiotične analize (pomen znaka).

3.1. Etnolingvistična in semiotična analiza s pomočjo konkordančnikov

Čeprav pregovori tradicionalno spadajo na področje paremiologije, so pogosto tudi raziskovalni predmet folkloristike, sociologije, pedagogike, jezikoslovja itd. Semiotika, kot veda o znakih, ponuja metodologijo za raziskovanje globljih dimenzij prepletenih kulturnih ozadij pregovorov (Grzybek, 2014). Semiotika s poudarkom na pragmatični (razmerje med označenim in označencem), sintaktični (formalni odnosi med znaki) in semantični dimenziji (odnosi znakov s predmeti, za katere je mogoče uporabiti znake) (Morris, 1938) omogoča opazovanje pregovorov z globljim vpogledom v kulturne pomene, pojme in svetovne nazore. Do svetovnega nazora v pregovorih pa je moč dostopati z etnolingvističnimi raziskovalnimi metodami, vključno z diahronim in sinhronim pristopom.

Etnolingvistika kot samostojno področje daje jeziku posebno mesto v družbi: v jeziku se oblikujejo kulturni pomeni; jezik v besedah, frazeologiji, celo v slovnici posreduje podobe sveta. Jezik je s tega vidika »gradivo kulture«, medtem ko je hkrati tudi kulturni meta-jezik: skupaj s folkloro velja za enega ključnih kulturnih kodov in kulturno ekspresivnih oblik. Jezik je zato eden najpomembnejših virov za raziskovanje folklore in rekonstrukcij njenih zgodnjih stanj; povezava med jezikom in kulturo je vzajemna (Tolstaja, 2006) in skupaj tvorita znakovni sistem. Vsi kulturni pomeni se zberejo v semantiko poimenovanja z besedami; te je lublinska etnolingvistična šola poimenovala jezikovni stereotipi (Bartmiński, 2005), ki kažejo naš poskus nadzora sveta. Analize relativno stalnih besednih zvez in besed v določenih kontekstih nam prikazujejo jezikovni zemljevid sveta z najpomembnejšimi družbenimi podobami in predstavami.

Hitro razvijajoče se področje digitalne humanistike omogoča raziskovalcem sprejemanje novih, korenito drugačnih metod raziskovanja in, kar je prav tako pomembno, daje na voljo elektronske zbirke z naprednimi možnostmi iskanja podatkov (Rassmusen Neal, 2015). Korpusno jezikoslovje in trenutno priljubljena »metodologija branja na daljavo« (tj. uporaba e-virov) poskuša izkoristiti velike jezikovne vzorce, da bi pridobili (kvantitativni) vpogled v besedišče, uporabo, trende in vizualizacije na področjih jezikovnega interesa. Hkrati pa takšne računalniške tekstovne oblike zbirk omogočajo natančnejše in hitrejše kvalitativne analize večjih zbirk: posameznih konkordančnih kombinacij in besednih okolij.

Semiotična analiza v namene etnolingvistične raziskave (Bartmiński, 2005) paremiološkega gradiva poteka predvsem na ravni semantike: pri besedah želimo zaznati tako metaforične pomene kot stereotipne oznake, ki jih (posamezna) beseda vsebuje in hkrati posreduje prek metafore v širši kontekst, torej s semiotičnega vidika, kakšni znaki se tvorijo znotraj paremiološke enote.

Statistični vpogled v celotno zbirko pokaže med drugim tudi najbolj pogosto rabljene besede, ki lahko podajo tudi splošnejša predvidevanja o družbeni naravnanosti. Najpogostejša polnopomenska beseda v zbirki paremioloških enot je:

- samostalnik *dan* se pojavi 1.657-krat; ta metaforično ali metonimično označuje tako časovno omejeno obdobje, ki pomeni dolgo (*Premislek je boljši kot dan hoda.*) ali kratko (*Bitke ne dobiš v enem dnevu.*), konec obdobja (*Po večeru se dan pozna.*), poimenovanje konkretnega dneva (*Ni vsak dan praznik./Pavla dne lepo, leto dobro bo.*), sledenje dobrega oz. označevanje konceptualne cikličnosti (*Za vsako nočjo pride dan*). Najpogostejša pojavnost ne presenеča, saj je ta samostalnik zelo pogost sestavni element vremenskih in kmetijskih paremioloških enot, poleg tega pa je je tudi v splošnem sodobnem jeziku izredno pogost: v Gigafidi v2.0 je tretji najpogosteje rabljeni samostalnik[5]. Po drugi strani je smiselno izpostaviti, da se nasprotje, tj. *noč* pojavi le 318-krat (pojavlja se kot nasprotje dnevu (*Ljubezen vidi noč, kjer sije beli dan.*), temen čas, ko se ne vidi (*Ponoči so vse krave črne.*), vpliven čas (*Noč ima svojo moč.*), mejni čas (*Ne hvali*

[5] http://hdl.handle.net/11346/QHKH

dneva pred nočjo.), slab čas (*Dan se zjutraj išče, noč pa sama pride.*), oznaka prazničnih časov *velika noč, božična noč*) itd.).

- Glagol *biti* se pojavi 19.301-krat (zanikan pa 3.501-krat), kar ne preseneča, glede na to, da gre za enega najosnovnejših glagolov; glagol je najpogostejši tudi v splošnem sodobnem jeziku.[6]
- Pridevnik *dobro* se pojavi 1.367-krat, največkrat v osnovni obliki, najmanjkrat pa kot presežnik (prim. *slabo* se pojavi 301-krat, osnovnik najpogosteje, presežnik najmanjkrat). Na podlagi izoliranih enot bi lahko sklepali, da pregovori na semantični ravni pogosto izražajo vrednotenje stanja ali delovanja, kar poleg izražanja družbenega nazora potrjuje tudi njihov pedagoški potencial.
- Predlog *v* je najpogostejši predlog v paremioloških enotah, tj. pojavi se v 4.538-krat. Iz tega podatka lahko sklepamo, da izvorno konceptualno najpogosteje uvrščamo pojave znotraj časovno-prostorskega koncepta pojava, pa čeprav se pomensko raba predloga razširi tudi na izražanje namena, sredstva, odnosa do celote, dejanja/stanja ipd. Enako je opaziti tudi v sodobnem splošnem jeziku.[7]

Ob najpogostejši prisotnosti besed v paremioloških enotah *dan*, *biti*, *dobro* in *v* se izkaže, da te povsem ustrezajo tudi pogostnosti rabe v splošnem sodobnem jeziku, ne glede na to, da gre za večinoma arhivsko gradivo.

Za natančnejši etnolingvistični in konceptualni vpogled je primernejša analiza s posamezno sestavino (npr. samostalnikom *čevelj*, *medved*) in njenimi vezavami, na podlagi katerih lahko s semiotično metodo podamo interpretacije družbenih konceptualnih vidikov. Za tako analizo je najširše uporabno enostavno iskanje, ki v primeru te zbirke naniza vse sklone iskanega samostalnika, vključno s starejšimi zapisi, npr. pri iskanju besede *čevelj* (68 enot) iskalnik izloči vse sklone, prav tako pa zapis *črevelj*, *čevl* ipd. Ob zahtevnejših iskanjih je možno slediti tudi številu posamezni obliki zapisa: *črevelj* (2), *črevlju* (1), *čevle* (3), seznam besed pa omogoča tudi sledenju starejšim zapisom, virom in njihovi pogostnosti v časovnem razponu, variantnim rabam in morebitnim prenovitvam.

Enako pri iskanju vseh zapisov in sklonov besede *lisica* (starejša oblika *lesica*, 7 enot) iskalnik najde 93 paremioloških enot. Ob zahtevnejših iskanjih je možno slediti tudi številu posamezni obliki zapisa: *lisica* (31), *lisice* (5), *lisici* (7), *lisico* (6), *lesica* (7), itd. Kontekstualna raziskava objav po različnih virih poda poveden podatek: slovnice in slovarji navajajo paremiološke enote z besedo *lisica*, ki so v celoti metaforične in se nanašajo na ljudi, medtem ko koledarji navedejo tudi paremiološke enote, ki veljajo za vremenske napovedi.

Iskalnik omogoča tudi iskanje želene besede v navezavi z drugo besedno vrsto, npr. lema *medved*, ki mu sledi glagol. Sicer je tako moč ugotoviti marsikatere povezave, vendar sam statistični del v nasprotju s pričakovanji prikaže tudi rezultate iz drugih (predhodnih ali sledečih) pregovorov, ne le rezultate, vezane na posamezni pregovor. Na primeru besede *medved* statistični del prikaže 79 ustreznic, vendar je teh znotraj enega pregovora 66. Ob ročnem pregledu kaj hitro ugotovimo, da se ta beseda najpogosteje veže z glagolom *prodajati*. Ob navezavah na samostalnik se pojavlja *koža*, kar tvori pregovor, ki metaforično svari pred preuranjeno hvalo. Pregovor nakazuje semantično polje, ki se v etnolingvistični interpretaciji veže na ekonomski odsev družbe, tj. prodaje medvedove kože, ki v zgodovinskem kontekstu pokaže svojo veliko ekonomsko vrednost.

Kljub vsemu iskalnik zaradi starejših in narečnih izrazov ne poišče vedno vseh kombinacij, npr. pri *Lep čevelj vidiš, a ne veš, kje me gloje* ali *Kdor stare čevlje flika, pride do zlatnika*, kjer konkordančnik ni zaznal kombinacije samostalnika in glagola.

Variante posameznega pregovora najlažje najdemo z iskanjem po besednih zvezah, npr. iskanje besedne zveze *lastovka ne* poda štiri rezultate: *Ena lastovka ne naredi poletja*, *Ena lastovka ne naredi pomladi*, *Ena lastovka ne prinese pomladi*, *Ena lastovka ne prinese nikoli spomladi*. Pri glagolski besedni zvezi *gre samo enkrat na led* pa rezultat poda tako *osla* kot *lisico* (*Osel/lisica gre samo enkrat na led*), prav tako *svoj rep hvali* lahko tako *lisica* kot *mačka* (*Vsaka lisica/mačka svoj rep hvali*).

Etnolingvistični vpogled v korpus pregovorov je z digitaliziranim gradivom in možnostjo zahtevnejšega iskanja temeljitejši. Že pogostost posameznih besed v pregovorih ali pa podatek o variantnosti posameznega pregovora je odlično izhodišče, ki ga z analognim arhivom le težko dosežemo.

4. Sklep

Digitalizacija folklornega gradiva olajša analizo le tega, hkrati pa postane bistveno bolj natančna – iskalniki omogočajo izpis vseh želenih enot, hkrati pa je primerjava gradiva bolj dosledna.

Vzpostavitev digitalne zbirke paremioloških enot ISN pomeni premik v slovenski folkloristiki. Gradivo je dostopnejše in analitično lažje obvladljivo. Hkrati takšna oblika ne terja (semantične, tematske, funkcijske, abecedne ipd.) kategorizacije pregovorov, temveč so razvrščeni kot najmanjša zaključena besedila, na katerih izvedemo analizo. Nedvomno je glede problema kategorizacije takšna rešitev najugodnejša, saj sama kategorizacija pogosto pokaže več pomanjkljivosti kot prednosti.

Pri zbirki pregovorov vsekakor najdemo mesto za izboljšave: poleg odprave nekaterih pravopisnih napak, se poraja vprašanje variantnosti ter povezave med variantami; na ta način bi bili odstranjeni tudi še nekateri podvojeni pregovori (predvsem tisti, ki so vpisani z različnimi ločili, npr. eden z vejico, drug s podpičjem). Ker je ponekod diplomatični prepis problematičen (gajica, bohoričica, metelčica, fonetični zapis), se poraja vprašanje smiselnosti knjižnega zapisa pregovora, ki bi moral biti ročno preverjen. Zbirka bo vsekakor tudi dopolnjena z novimi paremiološkimi enotami (iz starejših virov kot sodobne rabe). Poleg teh pa bi bilo smiselno uvesti tudi razdelitev virov po kategorijah, ki bi natančneje prikazal prisotnost paremioloških enot v posamezni kategoriji virov, kar bi omogočalo tudi primerjalno analizo (npr. enote v koledarjih in enote v slovnicah).

Za izdelavo digitalne paremiološke zbirke smo posegli po sistemih, ki so ustaljeni v jezikoslovju. V premisleku pa ostaja, kako digitalizirati slovstveno folkloro, ki je daljša (npr. zgodbe, molitve) in ima specifične funkcije (npr. uganke, zagovori).

[6] http://hdl.handle.net/11346/XNRI

[7] http://hdl.handle.net/11346/ZYVZ

Zahvala

Digitalizirana zbirka paremiolških enot ne bi nastala brez projektnih sodelavcev, še posebej Miha Pečeta: njegov občutek za folklorno gradivo in poznavanje računalniškega sveta sta omogočila hiter potek dela in sprotno reševanje zagat.

Delo opisano v prispevku je podprl temeljni raziskovalni projekt »Tradicionalne paremiološke enote v dialogu s sodobno rabo« (ARRS J6-2579).

5. Literatura

Ojo Arewa in Alan Dundes. 1966. Proverbs and the Ethnography of Speaking Folklore. *American Anthropologist*, 64: 70–85.

Saša Babič, Miha Peče, Tomaž Erjavec, Barbara Ivančič Kutin, Katarina Šrimpf Vendramin, Monika Kropej Telban, Nataša Jakop, in Marija Stanonik. 2022. *Collection of Slovenian paremiological units Pregovori 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1455.

Jiři Bartmiński. 2005. *Jazykovoj obraz mira: očerki po etnolingvistike*. Indarik, Moskva.

Kaja Dobrovoljc et al. 2017. The Universal Dependencies Treebank for Slovenian. V: *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, str. 33–38, Association for Computational Linguistics, doi:10.18653/v1/W17-1406.

Alan Dundes. 1965. *The study of folklore*. Prentice-Hall, Englewood Cliffs.

Tomaž Erjavec. 2021. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46(1): 35–57.

Tomaž Erjavec. 2015. *Reference corpus of historical Slovene goo300k 1.2*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1025.

Diana Faridovna Khakimzyanova in Enzhe Kharisovna Shamsutdinova. 2016. Corpus Linguistics in Proverbs and Sayings Study: Evidence from Different Languages. *The Social Sciences*, 11(15): 3770–3773.

Peter Grzybek. 2012. Proverb Variants and Variations: A New Old Problem? V: O. Lauhakangas, ur., in R. J. B. Soares, ur., *Proceedings of the Fifth Interdisciplinary Colloquium on Proverbs*, str. 136–152, AIP-IAP, Tavira.

Peter Grzybek. 2014. Semiotic and Semantic Aspects of the Proverb. V: H. Hrisztova-Gotthardt, (ur.) in M. A. Varga, ur., *Introduction to Paremiology: A Comprehensive Guide to Proverb Studies*, str. 68–111, De Gruyter, Warsaw/Berlin.

Dell Hymes, D. 1962. The ethnography of speaking. V: T. Gladwin, ur., in W. C. Sturtevant, ur., *Anthropology and Human Behavior*, str. 13–53, Anthropological Society of Washington, Washington.

Fran Kocbek. 1887. *Pregovori, prilike in reki*. Založil Anton Trstenjak, Ljubljana.

Fran Kocbek in Ivan Šašelj. 1934. *Slovenski pregovori, reki in prilike*. Družba Sv. Mohorja, Ljubljana.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Erika Kržišnik. 2008. Kulturološka interpretacija frazema. V: M. Kalin Golob, ur., N. Logar Berginc, ur., in A. Grizold, ur., *Jezikovna prepletanja*, str. 149–165, Fakulteta za družbene vede, Ljubljana.

Nikola Ljubešić in Kaja Dobrovoljc. 2019. 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, Association for Computational Linguistics, doi:10.18653/v1/W19-3704.

Milko Matičetov. 1956. Pregovori in uganke; ljudska proza. Slovenska matica, Ljubljana.

Matej Meterc. 2021. Aktualna raba in pomenska določljivost 200 pregovorov in sorodnih paremioloških izrazov. *Jezikoslovni zapiski* 27(1): 45–61.

Jozef Mlacek. 1983. Problémy komplexného rozboru prísloví a porekadiel. *Slovenská reč* 48(2): 129–140.

Wolfgang Mieder. 1993. *Proverbs are never out of season: Popular wisdom in modern age*. Oxford University Press.

Hanna F. Pitkin. 1972. *The concept of representation*. University of California Press.

Diana Rassmusen Neal. 2015. *Indexing and retrieval of non-text information*. De Gruyter Saur, Chicago, Vancouver.

Yves Scherrer in Nikola Ljubešić. 2016. Automatic Normalisation of the Swiss German ArchiMob Corpus Using Character-Level Machine Translation. V: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, str. 248–55.

Christoph Schöch, Roxana Patraş, Tomaž Erjavec, Diana Santos. 2021. Creating the European Literary Text Collection (ELTeC). *Modern languages open*, doi: 10.3828/mlo.v0i0.364.

Marija Stanonik. 1996. *Slovenski pregovori in rekla*. Projektna prijava.

Marija Stanonik. 2004. *Informatizacija neoprijemljive dediščine za etnologijo in folkloristiko*. Projektna prijava.

Marija Stanonik. 2009. *Slovenski pregovori kot kulturna dediščina: klasifikacija in redakcija korpusa*. Projektna prijava.

Marija Stanonik. 2015. Slovenski pregovori kot kulturna dediščina. Klasifikacija in redakcija korpusa. *Traditiones*, 44(3): 171–214.

Kathrin Steyer. 2017. Corpus Linguistic Exploration of Modern Proverb Use and Proverb Patterns. V: R. Mitkov, ur., *Europhras 2017. Computational and corpus-based phraseology: Recent advances and interdisciplinary approaches. Proceedings of the Conference Volume II*, str. 45–52, London, Geneva.

TEI Consortium. 2022. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. https://tei-c.org/guidelines/P5/

Svetlana M. Tolstaja. 2015. *Obraz mira v tekste i rituale*. Univerza Dimitrija Požarskega, Moskva.

DirKorp: A Croatian Corpus of Directive Speech Acts

Petra Bago*, Virna Karlić†

* Department of Information and Communication Sciences
† Department of South Slavic Languages and Literatures
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
{pbago, vkarlic}@ffzg.hr

Abstract

In this paper we present recent developments on a new version (v2.0) of DirKorp (*Korpus direktivnih govornih činova hrvatskoga jezika*), a Croatian corpus of directive speech acts developed for the purposes of pragmatic research. The corpus contains 800 elicited speech acts collected via an online questionnaire with role-playing tasks. Respondents were 100 Croatian speakers, all undergraduate or graduate students of the Faculty of Humanities and Social Sciences University of Zagreb. The corpus has been manually annotated on the speech act level, each speech act containing up to 12 features. It contains 12,676 tokens and 1,692 types. The corpus is encoded according to the *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, developed and maintained by the *Text Encoding Initiative Consortium* (TEI). We describe applied pragmatic annotation as well as the structure of the corpus.

1. Introduction

Corpus pragmatics is an interdisciplinary field of study that incorporates linguistic pragmatics and computer science, focusing on the development of natural language corpora in machine-readable form and their application for the purposes of studying pragmatics phenomena in written and spoken language. For a long time have linguists regarded a corpus approach to language incompatible with pragmatics (Romero-Trillo, 2008: 2). While the corpus approach to studying language implies processing authentic language material implementing quantitative research methods, pragmatic research is still predominantly of qualitative nature – based on the researcher's introspection, data obtained by elicitation methods or an analysis of authentic linguistic material of small size. The application of corpus analysis in the research of pragmatics phenomena represents a major turnaround in the development of pragmatics, primarily because it allows a systematic analysis of authentic language material of large size, and thus the detection of patterns of language use that "go below radar" through qualitative analyses (ibid.). In addition, it should be pointed out that the application of new technologies in linguistics, including pragmatics, did not only ensure, facilitate or accelerate numerous research processes, but opened the door to a new, different way of thinking about language (Leech, 1992).

The application of corpus methods on large pragmatic corpora allows one to systematically carry out empirically based pragmatic research (Bunt, 2017: 327). While the implementation of corpus research can result in minor adjustments to existing theories on the one hand, it can lead to a rethinking of pragmatics concepts and theoretical frameworks on the other hand, for example the development of the theory of dialogue acts (ibid.).

According to Rühlemann and Aijmer (2015), one of the major methodological problems that corpus pragmatics researchers encounter is the disproportionate relationship between pragmatic functions and language forms by which these functions are expressed. One form can perform multiple pragmatic functions in discourse, while one function can be expressed by different forms, which makes the process of querying a corpus according to the pragmatic function criterion considerably difficult. It is for this reason that corpus pragmatics researchers most often investigate conventional speech acts or functions performed by a limited number of language forms (Jucker, Scheier, and Hundt, 2009: 4). The aim of this paper is to present the first Croatian corpus of directive speech acts DirKorp, manually annotated for corpus pragmatic research.

The paper is structured as follows: Section 2 describes selected work related to pragmatic corpora, while the subsequent three section present the DirKorp corpus. Section 3 gives a description of the developed corpus, Section 4 describes 12 annotation features, and Section 5 presents the structure of the corpus encoded according to the *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (TEI Consortium, 2021). Finally, Section 6 contains conclusion and future work.

2. Related Work

The number of large corpora with systematically implemented pragmatic annotation is small so far. Due to a disproportionate relationship between pragmatic functions and language forms by which these functions are expressed, automatic corpus annotation does not produce satisfactory results. For this reason, only a small number of researchers have engaged in the creation of larger corpora of this sort. Generally, for the purposes of corpus pragmatic research, specialized corpora of smaller size are produced for individual research purposes. In addition, pragmatic research is sometimes carried out on corpora without pragmatic annotation.

An example of a corpus that does not contain pragmatic annotation, but which was used for pragmatic research is the Birmingham Blog Corpus[1] (Kehoe and Gee, 2007; Kehoe and Gee, 2012). In fact, this is a subcorpus of a larger set of corpora being developed at the department *Research and Development Unit for English Studies* at the Birmingham City University. It consists of blog posts and reader comments, sizing 500M words in English that were collected between 2000 and 2010. Automatic POS annotation was performed using the

[1] https://www.webcorp.org.uk/wcx/lse/corpora

Stanford Core NLP tools[2] and include lemma annotations and part-of-speech categories[3] based on the Universal Dependencies framework[4], while documents contain metadata of the publication date. Pragmatic research on speech acts was conducted on this corpus: For example, Lutzky and Kehoe (2017a; 2017b) used it to analyze apologies as speech acts that contain formulaic expressions, which facilitate its querying in a corpus when using available tools.

Similarly, we (Karlić and Bago, 2021) conducted research on the pragmatic functions and properties of imperatives using corpora without pragmatic annotation. We used hrWaC and srWaC (Ljubešić and Klubička, 2014), two large web corpora of Croatian and Serbian language with morphosyntactic annotation. For the purposes of the analysis, an additional pragmatic annotation of a representative sample of verbs in an imperative form was carried out manually. Other corpora of the Croatian spoken and written language with no pragmatic annotation have also been used as a resource for a corpus pragmatic research. For example, Hržica, Košutar, and Posavec (2021) used the Croatian Corpus of the Spoken Language of Adults (HrAL) (Kuvač Kraljević and Hržica, 2016) and the Croatian National Corpus of the written language (HNK) (Tadić, 1996) for the search and analysis of connectors and discourse markers.

According to Bunt (2017) the majority of corpora with pragmatic annotation contain labels on discourse relationships in written texts and on spoken dialogue acts. An example of such a larger corpus is Penn Discourse Treebank or PDTB[5] (Prasad, Webber, and Lee, 2018) that contains labels on discourse relations, i.e. discourse structure and its semantics. Discourse annotations were added to a subcorpus consisting of texts published in the newspaper *Wall Street Journal* sizing 1M tokens, included in a bigger corpus *Penn Treebank* (PTB). Bunt (2017) states that there are corpora of other languages developed for the purposes of studying the co-occurrence of discourse labels, such as Chinese, Czech, Dutch, German, Hindi and Turkish – emphasizing that these corpora are manually annotated and of modest sizes. Additionally, for each corpora a new schema was developed based on various theoretical starting points.

DialogBank[6] (Bunt et al., 2019) is one of a rare dialogue corpus annotated with an ISO 24617-2 standard. It contains already existing dialogue corpora annotated with various schemas. Four corpora are of English: HCRC Map Task (Anderson et al., 1991), Switchboard (Godfrey, Holliman, and McDaniel, 1992), TRAINS (Allen et al., 1995) and DBOX (Petukhova et al., 2014); and four of Dutch language: DIAMOND (Geertzen et al., 2004), OVIS[7], Dutch Map Task (Caspers, 2000) and Schiphol (Prüst, Minnen, and Beun, 1984). Dialogue act annotation involves segmenting a dialogue into defined grammatical units and augmenting each unit with one or more communicative function labels.

Another example of a corpus with a pragmatic annotation is the *Engineering Lecture Corpus*[8] (Alsop and Nesi, 2013; Alsop and Nesi, 2014) that contains 76 transcripts based on an hour-long video recordings of engineering lectures held in English on three universities. It is manually annotated for three pragmatic features: humor, storytelling and summary[9]. Each feature can be augmented with one of the attributes containing additional information that describes the feature in more detail. Further, the corpus contains labels regarding significant breaks, laughter, writing or drawing in the board, etc.

Finally, we present SPICE-Ireland corpus (*Systems of Pragmatic Annotation in the Spoken Component of ICE-Ireland*) (Kallena and Kirka, 2012), a part of a larger set of corpora ICE-Ireland (*International Corpus of English: Ireland Component*) containing pragmatic, discourse and prosodic features. The corpus contains various types of private and public, formal and informal dialogues and monologues of a length of about 2,000 words, sizing 625K words. It consists of spoken English. The pragmatic annotation of speech acts is based on Searle's classification (Searle, 1969; Searle, 1976): representatives, directives, commissives, expressives and declaratives.

To the best of our knowledge, there exist no publicly available corpora of spoken or written Croatian language with pragmatic annotation. So far, Croatian linguists mostly dealt with speech acts from a theoretical perspective, referring primarily to the Austin's and Searle's theory (cf. Pupovac, 1991; Ivanetić, 1995; Miščević, 2018; Palašić, 2020). However, in recent times, the number of research based on qualitative and quantitative analysis of small-sized authentic linguistic materials (from literary texts and advertisements to email messages and political discourse in Croatian and other languages) has been increasing (cf. e.g., Pišković, 2007; Matić, 2011; Franović and Šnajder, 2012; Šegić, 2019).

In the following sections we present a new version (v2.0) of DirKorp, the first Croatian corpus of directive speech acts.

3. Corpus Description

DirKorp (*Korpus direktivnih govornih činova hrvatskoga jezika*) (Karlić and Bago, 2021) is a Croatian corpus of directive speech acts developed for the purposes of pragmatic research. The corpus contains 800 elicited speech acts collected via an online questionnaire with role-playing tasks applying the method of simulated communication that is implemented under pre-set conditions. This method is suitable for researching speech acts due to the ability to collect a great number of examples of speech acts of the equal propositional content and illocutionary purpose used in the same controlled situations. The questionnaire included eight closed-type role-playing tasks. These types of tasks imply recording the speaker's reactions (in this case in writing) to the stimulus without feedback. In each task, the participants are presented with one textually described hypothetical situation asking them to refer a directive speech act to

[2] https://stanfordnlp.github.io/CoreNLP/

[3] See more about the POS tagset used for the Birmingham Blog Corpus: https://www.webcorp.org.uk/wcx/lse/guide.

[4] https://universaldependencies.org/u/pos/index.html

[5] https://doi.org/10.35111/qebf-gk47

[6] https://dialogbank.uvt.nl/

[7] http://www.let.rug.nl/vannoord/Ovis/

[8] www.coventry.ac.uk/elc

[9] https://www.coventry.ac.uk/research/research-directories/current-projects/2015/engineering-lecture-corpus-elc/annotations-and-mark-ups/

their interlocutor. Their assignment was to imagine they were in the presented situation and to give a written statement they would use in the described situations. The presented situations are classified into two categories with regard to the relationship between the participants of the communication act: (1) situations involving interlocutors who are not in a familiar relationship; (2) situations involving interlocutors in a familiar relationship. Assignments of the two categories are organized into four pairs, asking respondents to share a speech act of similar propositional content: "I want you to return something that belongs to me" (for text of role-playing tasks see Example 1 when interlocutors have (a) an unfamiliar relationship and (b) a familiar relationship); "I want you to answer my inquiry"; "I want you to change something that bothers me"; "I want you to stop behaving inappropriately"[10].

Example 1

(a) Upravo si pojeo/la ručak u restoranu. Posluživao te stariji konobar koji se odnosio prema tebi ljubazno i profesionalno. Prilikom plaćanja računa konobar ti vraća 100 kuna manje nego što je trebao. Želiš da ti konobar vrati novac. Zamisli da se konobar nalazi pred tobom i napiši što bi mu točno rekao/la u danoj situaciji (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).

(Eng. *You just ate lunch at a restaurant. You were served by an elderly waiter who treated you kindly and professionally. When paying the bill, the waiter refunds you 100 kunas less than he should have. You want the waiter to give you your money back. Imagine the waiter was in front of you and write what exactly you would say to him in the given situation (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

(b) Posudio/la si knjigu najboljem prijatelju (ili prijateljici). Rekao ti je da će ti je uskoro vratiti, no nije održao riječ. Sjedite zajedno u kafiću, situacija je opuštena, razgovarate o svakodnevnim stvarima. Želiš mu dati do znanja da ti treba čim prije vratiti knjigu. Zamisli da se tvoj prijatelj nalazi pred tobom i napiši što bi mu točno rekao/la u danoj situaciji (nemoj prepričavati, već iskaz formuliraj kao da se izravno obraćaš sugovorniku).

(Eng. *You lent a book to your best friend. (S)he told you (s)he'd give it back to you soon, but (s)he didn't keep her/his word. You are sitting together in a café, the situation is relaxed, you talk about everyday things. You want to let her/him know you need to get your book back as soon as possible. Imagine if your friend was in front of you and wrote what exactly you would say to her/him in the given situation (do not recount but formulate the statement as if you were addressing the interlocutor directly).*)

Respondents were 100 Croatian speakers, all undergraduate (63 %) or graduate students (37 %) of the Faculty of Humanities and Social Sciences University of Zagreb, ages between 18 to 33. Croatian is the mother tongue for the majority of the respondents (96 %). The questionnaire was carried out during December 2020 and January 2021. All respondents voluntarily participated in the study. The questionnaire was conducted anonymously, and the collected language material was used exclusively for scientific purposes.

The elicitation of language production by the role-playing method has its advantages and disadvantages. On the one hand, it enables the collection of a large number of speech acts with the same propositional content and illocutionary purpose. On the other hand, users of the corpus should keep in mind that the language material collected by this method does not reflect the features of actual language use. It rather shows what speakers think they would say and/or do in hypothetical situations.

DirKorp contains 12,676 tokens and 1,692 types[11]. Since it consists of 800 speech acts, it is a relatively small corpus. However, as the first Croatian corpus with detailed pragmatic annotation, DirKorp can serve as a useful resource for researching speech acts, politeness strategies and other related pragmatic phenomena in the Croatian language. In addition, we hope that it will contribute to the development of larger corpora of the Croatian language with pragmatic annotation, and that it will encourage a wider application of the corpus-pragmatic research method.

We have conducted corpus pragmatic analyses of the collected speech acts in order to investigate ways and means of expressing directives, and their pragmatic characteristics and functions. For example, we confirmed that indirect directives are more frequent than direct, especially among interlocutors who are not in a familiar relationship. Regarding (un)familiar relationship between interlocutors, we detected that explicit illocutionary force is more frequent in communication between interlocutors with a familiar relationship, while implicit illocutionary force is more frequent in communication between interlocutors with an unfamiliar relationship. Additionally, we have identified that imperative utterances are a more frequent type of direct directives than utterances with a directive performative verb in 1st person. For more such corpus pragmatic analyses see Karlić and Bago (2021).

4. Corpus Annotation

Collected language material has been manually annotated on the speech act level by two independent annotators with university graduate degrees in the field of philology. Annotators received oral and written instructions, including illustrative examples for all the features they had to annotate.

The categorization of speech acts and their formal and pragmatic properties was carried out according to the theory of speech acts by Austin (1962), Searle (1969; 1976) and their successors; the politeness theory of Brown and Levinson (1978), and the grammars of contemporary Croatian and Serbian languages (Silić and Pranjković, 2007; Piper et al., 2005). For more on individual

[10] Full texts of role-playing tasks are available in the corpus header.

[11] Respondents' answers contain utterances, but also text about what they would do in the given situation. At this moment, we have not analyzed average length of a response. Generally, we can only state that some speech acts contain only one utterance, while some contain more than one.

categories, see Karlić and Bago (2021). In the new version of DirKorp (v2.0), each speech act can contain up to 12 features. The first 8 features were part of the corpus version v1.0, while features 9-12 are newly added. For frequency distribution of all features see Karlić and Bago (2021).

(1) **Respondent ID** – This mandatory feature contains information on identification of the respondent uttering the speech act.

(2) **Familiarity / unfamiliarity** – This mandatory feature contains information on the category of the proposed situation in which the speech act was uttered. Four situations are labelled 'unfamiliar' (involving interlocutors who are not in a familiar relationship), while the other four situations are labelled 'familiar' (involving interlocutors who are in a familiar relationship).

(3) **Utterance type** – This mandatory feature contains information on the utterance type regarding its structural organization. It contains five labels: (a) an imperative utterance, (b) an assertive utterance (a statement), (c) an utterance in the form of a question, (d) an utterance in the form of an ellipsis, (e) a nonverbal signal, (f) a case of avoidance of executing a speech act (see Example 2).

Example 2
(a) E vrati mi onu knjigu koju sam ti posudio.
(Eng. *Hey, give me back that book I lent you.*)
(b) Oprostite, ali mislim da ste mi krivo vratili novce.
(Eng. *Excuse me, but I think you gave me my money back wrong.*)
(c) Možete li molim vas zatvoriti prozore?
(Eng. *Could you please close the windows?*)
(d) E, moja knjiga??
(Eng. *Hey, my book??*)
(e) [Samo bih zavrtjela očima da vide moje neodobravanje, ali ne bih ništa rekla.]
(Eng. *[I'd just roll my eyes so that they see my disapproval, but I wouldn't say anything.]*
(f) [Ne bih ništa rekao.]
(Eng. *[I wouldn't say anything.]*)

(4) **Directive performative verb in 1st person** – This optional feature contains information on the representation of a directive performative verb in 1st person as part of the speech act, only for assertive utterances and utterances in the form of a question. It contains two labels: (a) yes and (b) no (see Example 3).

Example 3
(a) Oprostite, molim da odete na kraj reda.
(Eng. Excuse me*, I am imploring you to go to the end of the line.*)
(b) Gospođo, morate na kraj reda stati.
(Eng. *Madam, you must move to the end of the line.*)

(5) **Illocutionary force** – The optional feature contains information on explicitness or implicitness of the illocutionary force of a speech act. It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question and in the form of an ellipsis). It contains two labels: (a) explicit and (b) implicit (see Example 4).

Example 4
(a) Daj mi donesi više onu knjigu, treba mi!
(Eng. *Bring me that book already, I need it!*)
(b) Kaj je s onom knjigom koju sam ti posudio?
(Eng. *What happened to that book I lent you?*)

(6) **Propositional content** – This optional feature contains information on explicitness or implicitness of the propositional content of a speech act. It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question and in the form of an ellipsis). It contains two labels: (a) explicit and (b) implicit (see Example 5).

Example 5
(a) Gledaj na cestu, pusti mobitel.
(Eng. *Look at the road, leave the cell phone.*)
(b) Ti hoćeš da poginemo?
(Eng. *You want us to die?*)

(7) **T/V form** – This optional feature contains information on how the respondent addressed the interlocutor, using an informal (T-form) or a formal *you* (V-form). It is only applied to utterances that contain verbal means (an imperative utterance, an assertive utterance, an utterance in the form of a question and in the form of an ellipsis). It contains three labels: (a) T-form, (b) V-form and (c) impossible to determine (see Example 6).

Example 6
(a) Oprosti, dao si mi manje novca
(Eng. *Sorry, you$_{T\text{-}form}$ gave me less change.*)
(b) Oprostite, mislim da ste mi ipak još dužni100 kuna.
(Eng. *Excuse me, I think you$_{V\text{-}form}$ still owe me 100 kunas.*)
(c) Hmm... još 100 kuna, zar ne?
(Eng. *Hmm… another 100 kunas, right?*)

(8) **Exhortative** – This optional feature contains information on the representation of an exhortative as part of the speech act. It contains two labels: (a) yes and (b) no (see Example 7).

Example 7
(a) Daj mi više vrati knjigu, treba mi za knjižnicu.
(Eng. *Bring me back my book already, I need it for the library.*)
(b) Jel se sjećaš one knjige koju sam ti posudila? Potrebna mi je. Možeš li mi ju donijeti sutra na faks?
(Eng. *Do you remember that book I lent you? I need it. Could you bring it tomorrow to uni?*)

(9) **Request** – This optional feature contains information on whether the speech act includes a lexical marker of request. It contains two labels: (a) yes and (b) no (see Example 8).

Example 8
(a) E da, jel bi mi mogao/la vratiti knjigu, molim te?
(Eng. *Oh yeah, could you bring the book back, please?*)
(b) Zaboravio si mi vratiti knjigu, jel se možeš idući put sjetiti?

(Eng. *You forgot to bring me back the book, can you remember next time?*)

(10) **Apology** – This optional feature contains information on whether the speech act includes a lexical marker of apology. It contains two labels: (a) yes and (b) no (see Example 9).

Example 9
(a) Oprostite, ovdje fali još 100 kuna
(Eng. *Excuse me, 100 kunas is missing here.*)
(b) Možete li molim vas pritvoriti prozore, hladno mi je?
(Eng. *Could you please close the windows, I'm cold?*)

(11) **Gratitude** – This optional feature contains information on whether the speech act includes a lexical marker of gratitude. It contains two labels: (a) yes and (b) no (see Example 10).

Example 10
(a) Molim te mi samo javi da znam zbog organizacije hoćeš li doći. Hvala ti!
(Eng. *Please just let me know whether you're coming so that I know because of the organization. Thank you!*)
(b) Heej, jel dolaziš večeras na druženje? Moram znati zbog organizacije. xoxo
(Eng. *Heeey, are you coming tonight to hang out? I need to know because of the organization. xoxo*)

(12) **Honorific title** – This optional feature contains information on whether the speech act includes an honorific title. It contains two labels: (a) yes and (b) no (see Example 11).

Example 11
(a) Gospođo, kraj reda je dolje
(Eng. *Madam, the end of the line is back there.*)
(b) Oprostite, tamo je kraj reda!
(Eng. *Excuse me, the end of the line is there!*)

5. Corpus Format

DirKorp is encoded according to the *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, developed and maintained by the Text Encoding Initiative Consortium (TEI) (TEI Consortium, 2021). The TEI document is comprised of a header and the body of the corpus. The content of the elements and attributes are in Croatian. Metadata of the corpus is given in the header including: bibliographic information; the editorial practice; a structured taxonomy describing categories used for each of the 12 pragmatic features in the annotation process (see Figure 1 for an example), including full text of the eight situations on the questionnaire; a list of questionnaire participants with information on their age, gender, undergraduate or graduate level of study, enrollment in a philological/non-philological/combined study program and mother tongue (see Figure 2 for an example); and a list of revisions of the DirKorp versions. The body of the corpus is composed of one division containing utterances with pragmatic features (see Figure 3 for an example).

DirKorp is available for download under the CC BY-SA 4.0 license from GitHub in TEI format (https://github.com/pbago/DirKorp).

```
<taxonomy xml:id="tiVi">
  <category xml:id="ti">
    <catDesc>Govorni čin sadržava
obraćanje na ti (atribut se odnosi
na tipove iskaza koji uključuju
verbalna sredstva [imperativni,
tvrdnja, upitni,
eliptični]).</catDesc>
  </category>
  <category xml:id="vi">
    <catDesc>Govorni čin sadržava
obraćanje na Vi (atribut se odnosi
na tipove iskaza koji uključuju
verbalna sredstva [imperativni,
tvrdnja, upitni,
eliptični]).</catDesc>
  </category>
  <category xml:id="persNeodredivo">
    <catDesc>Nije moguće odrediti
sadržava li govorni čin obraćanje na
ti ili Vi (atribut se odnosi na
tipove iskaza koji uključuju
verbalna sredstva [imperativni,
tvrdnja, upitni,
eliptični]).</catDesc>
  </category>
</taxonomy>
```

Figure 1: An example of a pragmatic feature description – how the respondent addressed the interlocutor (V-form, T-form or impossible to determine).

```
<person xml:id="I001" sex="F">
  <p>ispitanik/ispitanica, 20
godina, spol Ž, preddiplomski studij
Filozofskog fakulteta, nefilološko
usmjerenje, materinji jezik
hrvatski</p>
</person>
```

Figure 2: An example of participant information.

```
<u who="#I001" ana="#NEFAM1 #tvdrnja
#dpg1N #isI #psI #vi #adhorativN
#molbaN #isprikaY #zahvalaN
#honorifikN">Ispričavam se, pardon,
fali još sto kuna. Oprostite.</u>
```

Figure 3: An example of an utterance containing all 12 pragmatic features.

6. Conclusion and Future Work

We have presented DirKorp, the first Croatian corpus of directive speech acts, containing 800 elicited speech acts collected via an online questionnaire with role-playing tasks, specifically developed for pragmatic research studies. Respondents were 100 Croatian speakers, all students of the Faculty of Humanities and

Social Sciences University of Zagreb. The corpus has been manually annotated on the level of a speech act, each speech act containing up to 12 features. It contains 12,676 tokens and 1,692 types. The corpus is available for download under the CC BY-SA 4.0 license from GitHub in TEI format.

Further work is planned on the corpus, which includes an evaluation of the developed scheme for annotating directive speech acts, annotation at the levels smaller than a speech act, as well as augmentation with additional features such as information on grammatical mood used in a speech act, information on representation of modal verb in 2nd person as part of a speech act, and information on various politeness strategies applied in a speech act.

7. Acknowledgements

This paper is generously co-financed by the institutional project of the Faculty of Humanities and Social Sciences "South Slavic languages in use: pragmatic analyses" (principle researcher Virna Karlić). We wish to thank all our annotators.

8. References

James F. Allen, Lenhart K. Schubert, Geoge Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel G. Martin, Bradford W. Miller, Massimo Poesio, and David R. Traum. 1995. The TRAINS Project: A Case Study in Building a Conversational Planning Agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 7(1):7–48.

Sian Alsop and Hilary Nesi. 2013. Annotating a Corpus of Spoken English: The Engineering Lecture Corpus (ELC). In: *Proceedings of GSCP 2012: Speech and Corpora*, pages 58–62. Firenze University Press, Florence.

Sian Alsop and Hilary Nesi. 2014. The Pragmatic Annotation of a Corpus of Academic Lectures. In: *The International Conference on Language Resources andEvaluation 2014 Proceedings*, pages 1560–1563. European Language Resources Association, Reykjavik.

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task Corpus, *Language and Speech*, 34(4):351–366.

John L. Austin. 1962. *How to Do Things with Words*. Clarendon Press, Oxford.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Harry Bunt. 2017. Computational Pragmatics. In: *Oxford Handbook of Pragmatics*, pages 326-345. Oxford University Press, New York.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Fang, and Kars Wijnhoven. 2019. The DialogBank: Dialogues with Interoperable Annotations. In: *Language Resources and Evaluation*, 53(2):213–249.

Johanneke Caspers. 2000. Melodic Characteristics of Backchannelsin Dutch Map Task Dialogues. In: *Proceedings, 6th International Conference on SpokenLanguage Processing*, pages 611–614. China Military Friendship Publish , Beijing, https://www.isca-speech.org/archive/icslp_2000/.

Tin Franović and Jan Šnajder. 2012. Speech Act Based Classification of Email Messages in Croatian Language. In: *Proceedings of the Eighth Language Technologies Conference*, pages 69-72. Information Society, Ljubljana.

Jeroen Geertzen, Yann Girard, Roser Morante, Ielka Van der Sluis, Hans Van Dam, Barbara Suijkerbuijk, Rintse Van der Werf, Harry Bunt. 2004. The DIAMOND Project. In: *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (CATALOG 2004),* Barcelona.

John Godfrey, Edward Holliman, and Jande McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1*, pages 517–520. IEEE Computer Society, San Francisco.

Gordana Hržica, Sara Košutar, and Kristina Posavec. 2021. Konektori i druge diskursne oznake u pisanome i spontanome govorenom jeziku. *Fluminensia: časopis za filološka istraživanja*, 33(1):25–52.

Nada Ivanetić. 1995. *Govorni činovi*. Zagreb: FF-press, Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu.

Andreas H. Jucker, Daniel Schreier, and Marianne Hundt. (eds.). 2009. *Corpora: Pragmatics and Discourse*. Rodopi, Amsterdam.

Jeffrey L. Kallen and John M. Kirk. 2012. *SPICE-Ireland: A User's Guide*. https://pure.qub.ac.uk/en/publications/spice-ireland-a-users-guide.

Virna Karlić and Petra Bago. *(Računalna) pragmatika: temeljni pojmovi i korpusnopragmatičke analize*. FF Press , Zagreb, 2021. https://openbooks.ffzg.unizg.hr/index.php/Ffpress/catalog/book/125.

Andrew Kehoe and Matt Gee. 2007. New Corpora from the Web: Making Web Text More 'Text-Like'. In: *Studies in Variation, Contacts and Change in English 2*. https://varieng.helsinki.fi/series/volumes/02/kehoe_gee/.

Andrew Kehoe and Matt Gee. 2012. Reader Comments as an Aboutness Indicator in Online Texts: Introducing the Birmingham Blog Corpus. In: *Studies in Variation, Contacts and Change in English 12*. https://varieng.helsinki.fi/series/volumes/12/kehoe_gee/.

Jelena Kuvač Kraljević and Gordana Hržica. 2016. Croatian Adult Spoken Language Corpus (HrAL). *Fluminensia: časopis za filološka istraživanja*, 28(2):87–102.

Geoffrey N. Leech. 1992. Corpora and Theories of Linguistic Performance. In: *Directions in Corpus Linguistics*, pages 105–122. De Gruyter, Berlin.

Ursula Lutzky and Andrew Kehoe. 2016. Your Blog is (the) Shit: A Corpus Linguistic Approach to the Identification of Swearing in Computer Mediated Communication. *International Journal of Corpus Linguistics*, 21(2): 165–191.

Ursula Lutzky and Andrew Kehoe. 2017a. 'I Apologize for My Poor Blogging': Searching for Apologies in the

Birmingham Blog Corpus. *Corpus Pragmatics*, 1(1):37–56.

Ursula Lutzky and Andrew Kehoe. 2017b. 'Oops, I Didn't Mean to Be so Flippant'. A Corpus Pragmatic Analysis of Apologies in Blog Data. *Journal of Pragmatics*, 116:27–36.

Nikola Ljubešić and Filip Klubička. 2014. {bs, hr, sr}WaC-Web Corpora of Bosnian, Croatian and Serbian. In: *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Association for Computational Linguistics , Gothenburg, https://aclanthology.org/W14-0405.pdf.

Daniela Matić. 2011. *Govorni činovi u političkome diskursu*. PhD thesis. Faculty of Humanities and Social Sciences, Zagreb.

Nenad Miščević. 2018. *Rođenje pragmatike*. Orion Art, Beograd.

Nikolina Palašić. 2020. *Pragmalingvistika – lingvistički pravac ili petlja?* Hrvatska sveučilišna naklada, Zagreb.

Volha Petukhova, Martin Gropp, Dietrich Klakow, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlicek, Blaise Potard, John Dines, Olivier Deroo, Ronny Egeler, Uwe Meinz, Steffen Liersch, and Anna Schmidt. 2014. The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 252–258. European Language Resources Association, Reykjavik.

Predrag Piper et al. 2005 = Предраг Пипер, Ивана Антонић, Бранислава Ружић, Срето Танасић, Људмила Поповић, Бранко Тошовић. 2005. *Синтакса савременог српског језика*. Проста реченица, Београд: Институт за српски језик САНУ, Београдска књига, Матица српска.

Tatjana Pišković. 2007. Dramski diskurs između pragmalingvistike i feminističke lingvistike. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje,* 33(1):325–341.

Olumide Popoola. 2017. A Dictionary, a Survey and a Corpus Walked into a Courtroom...: An Evaluation of Resources for Adjudicating Meaning in Trademark Disputes. In: *The 9th International Corpus Linguistics Conference*. Birmingham: Birmingham University.https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2017/general/paper134.pdf.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse Annotation in the PDTB: The NextGeneration. In: *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97. Santa Fe: Association for Computational Linguistics. https://aclanthology.org/W18-4710.pdf.

Hub Prüst, Guido Minnen, and Robbert-Jan Beun. 1984. Transcriptie dialooogesperiment juni/juli 1984, *IPORapport 481*. Institute for Perception Research, Eindhoven University of Technology, Eindhoven.

Milorad Pupovac. 1990. *Jezik i djelovanje*. Biblioteka časopisa Pitanja, Zagreb.

Jesús Romero-Trillo (ed.). 2008. *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. De Gruyter, Berlin.

Christoph Rühlemann and Karin Aijmer. 2015. Introduction. Corpus pragmatics: laying the foundations. In: *Corpus pragmatics*, pages 1-28.

John R. Searle. 1969. *Speech Acts*. Cambridge University Press, Cambridge.

John R. Searle. 1976. A classification of illocutionary acts. *Language in Society*, 5:1–23.

Josip Silić and Ivo Pranjković. 2007. *Gramatika hrvatskoga jezika za gimnazije i visoka učilista*. Školska knjiga , Zagreb.

Tea Šegić. 2019. Tata kupi mi auto und Nivea Milk weil es nichts Besseres für die Hautpflege gibt. *Filologija*, 73:103–116.

Marko Tadić. 1996. Računalna obradba hrvatskoga i nacionalni korpus. *Suvremena lingvistika,* 41-42:603–611.

TEI Consortium (ed.). 2021. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Universal Dependencies za slovenščino: nadgradnja smernic, učnih podatkov in razčlenjevalnega modela

Kaja Dobrovoljc[*†‡], **Luka Terčon**[†], **Nikola Ljubešić**[‡†]

[*]Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
kaja.dobrovoljc@ff.uni-lj.si

[†]Fakulteta za računalništvo in informatiko
Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
luka.tercon@fri.uni-lj.si

[‡]Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana
nikola.ljubesic@ijs.si

Povzetek

Universal Dependencies (UD) je mednarodno usklajena označevalna shema za medjezikovno primerljivo oblikoslovno in skladenjsko označevanje besedil po načelih odvisnostne slovnice, ki je bila ob več kot 130 drugih svetovnih jezikih uspešno uporabljena tudi za označevanje besedil v slovenščini. V prispevku predstavimo rezultate nedavnih aktivnosti v povezavi s shemo UD znotraj projekta *Razvoj slovenščine v digitalnem okolju*, v okviru katerega smo obstoječo infrastrukturo nadgradili s prenovo in podrobno dokumentacijo označevalnih smernic UD za slovenščino, razširitvijo drevesnice SSJ-UD za pisno slovenščino z novimi povedmi iz korpusov ssj500k in ELEXIS-WSD ter izdelavo novega strojnega modela skladenjskega razčlenjevanja v označevalnem orodju CLASSLA-Stanza. V podporo nadaljnjim aplikacijam na različnih področjih strojnega procesiranja slovenščine novi model podrobneje ovrednotimo, in sicer poleg splošne evalvacije natančnosti razčlenjevanja poročamo tudi o natančnosti na ravni posamičnih skladenjskih relacij in o najpogostejših tipih napak.

1. Uvod

Jezikoslovno označeni korpusi, tj. digitalizirane zbirke besedil, ki poleg besed na površini vsebujejo tudi ročno pripisane podatke o njihovih slovničnih lastnostih na različnih ravneh jezikoslovnega opisa (Ide in Pustejovsky, 2017), predstavljajo enega izmed temeljnih jezikovnih virov za razvoj jezikovnotehnoloških orodij na eni strani in korpusnojezikoslovne raziskave na drugi. Slovnične lastnosti so besedilom tipično pripisane na podlagi vnaprej opredeljenih označevalnih shem oz. označevalnih sistemov, ki poleg nabora možnih oznak običajno vsebujejo tudi smernice za njihovo pripisovanje konkretnim slovničnim pojavom. Ker so v preteklosti označevalne sheme nastajale ločeno za posamezne jezike, slovnične teorije ali celo korpuse, je njihova posledična raznolikost onemogočala kakršnokoli neposredno primerjavo označenih podatkov ali na njih temelječih računalniških orodij.

Kot protiutež tovrstni razdrobljenosti je bila leta 2013 vzpostavljena označevalna shema Universal Dependencies,[1] ki si prizadeva za mednarodno oz. medjezično usklajeno slovnično označevanje besedil na oblikoslovni in skladenjski ravni, da bi pospešila razvoj večjezičnih jezikovnih tehnologij, medjezičnega strojnega učenja in kontrastivnih jezikoslovnih analiz. Znotraj sheme UD je bil tako vzpostavljen univerzalni nabor kategorij in smernic (17 besednih vrst, 24 oblikoskladenjskih lastnosti, 37 odvisnostnih skladenjskih relacij), ki odslej omogoča enotno označevanje podobnih slovničnih pojavov v različnih svetovnih jezikih, obenem pa dovoljuje tudi jezikovnospecifične izpeljave, če je to potrebno. Shema temelji na načelih odvisnostne slovnice, ki je v primerjavi s frazno pragmatiko bolj primerna za jezike s prostim besednim redom in za neposredno uporabo v različnih jezikovnotehnoloških aplikacijah (Jurafsky in Martin, 2021), njena teoretična izhodišča pa so podrobneje predstavljena v prispevku De Marneffe et al. (2021).

Doslej je bilo z označevalno shemo UD ročno označenih že več kot 200 korpusov (t.i. odvisnostnih drevesnic, angl. *dependency treebanks*) v 130 svetovnih jezikih. Med njimi sta tudi univerzalni odvisnostni drevesnici pisne slovenščine SSJ (Dobrovoljc et al., 2017) in govorjene slovenščine SST (Dobrovoljc in Nivre, 2016), ki sta bili s tem neposredno vključeni v razvoj številnih najsodobnejših orodij za večjezično obdelavo naravnih jezikov (Zeman et al., 2018), kakor tudi raznolike primerjalnojezikoslovne raziskave (Futrell et al., 2015; Naranjo in Becker, 2018; Chen in Gerdes, 2018).

Glede na pomen razvoja slovenskih virov v tovrstnih mednarodnih standardizacijskih pobudah smo v okviru nacionalnega projekta Razvoj slovenščine v digitalnem okolju (RSDO),[2] ki si prizadeva za zadovoljitev potreb po

[1] https://universaldependencies.org/

[2] https://slovenscina.eu/

računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik, obstoječe vire in povezano infrastrukturo za označevanje slovenskih besedil po sistemu Universal Dependencies bistveno nadgradili.

Potek in rezultate te aktivnosti predstavimo v nadaljevanju prispevka, v katerem po kratki predstavitvi izhodiščne različice korpusa SSJ-UD pred začetkom projekta RSDO (2. razdelek) opišemo dokumentacijo rahlo prenovljenih označevalnih smernic UD za slovenščino (3. razdelek). Nadaljujemo s predstavitvijo označevalne kampanje (4. razdelek), v okviru katere je bilo ročno razčlenjenih več kot 5.000 novih povedi, ki skupaj z nekoliko izboljšanim prvotnih korpusom tvorijo najnovejšo različico korpusa SSJ-UD (5. razdelek). V drugem delu prispevka opišemo izdelavo na novem korpusu temelječega napovednega modela za strojno skladenjsko razčlenjevanje (6. razdelek), ki ga v sklepnem delu tudi ovrednotimo z analizo splošne natančnosti (7. razdelek) in analizo najpogostejših napak (8. razdelek).

2. Nastanek korpusa SSJ-UD

Prva različica univerzalne odvistnostne drevesnice za pisno slovenščino SSJ-UD[3] je nastala na podlagi na polavtomatske pretvorbe korpusa ssj500k (Krek et al., 2020), bogato označenega referenčnega učnega korpusa za slovenščino, ki je bil predhodno že ročno lematiziran, oblikoskladenjsko označen in skladenjsko razčlenjen po označevalnem sistemu JOS (Erjavec et al., 2010). Medtem ko so leme in oblikoskladenjske oznake JOS pripisane vsem pojavnicam korpusa ssj500k (586.248 pojavnic oz. 27.829 povedi), je skladenjsko razčlenjena zgolj slaba polovica korpusa (235.864 pojavnic oz. 11.411 povedi).

Pretvorba korpusa ssj500k iz označevalne sheme JOS v shemo UD (Dobrovoljc et al., 2016; Dobrovoljc et al., 2017) je temeljila na širokem naboru pravil za preslikavo za vse tri ravni sheme UD: besedne vrste, oblikoslovne lastnosti in odvisnostne skladenjske relacije.[4] Ker so si (z nekaj izjemami) načela označevanja obeh sistemov na ravni oblikoslovja precej podobna, je bilo mogoče s pravili za preslikavo v besedne vrste in oblikoskladenjske lastnosti UD pretvoriti celoten korpus ssj500k oz. na istem sistemu temelječi leksikon Sloleks (Dobrovoljc et al., 2019), pri čemer je bilo ročno razdvoumljanje potrebno zgolj pri besednovrstni kategorizaciji glagola *biti*.[5]

Po drugi strani pa je bil skladenjsko razčlenjeni del korpusa ssj500k v shemo UD pretvorjen le delno, saj zaradi robustnosti sistema JOS v primerjavi z UD kljub podrobnemu sistemu pravil za preslikavo vseh povedi ni bilo mogoče v celoti samodejno pretvoriti z dovolj zanesljivo natančnostjo. Med nepretvorjenimi so tako ostale zlasti povedi s strukturami, ki so bile v sistemu JOS označene kot t. i. povezave tretjega nivoja (oznaka *modra*), kot so stavčna priredja in soredja, pristavki in pojasnjevalne strukture, členki oz. nepropozicijskimi prislovi, vrivki in podobno.

Prvotna različica korpusa SSJ-UD, prvič objavljena kot del zbirke drevesnic UD v1.2 leta 2015, je tako obsegala 8.000 povedi oz. 140.670 pojavnic. Kljub kontinuiranemu izboljševanju korpusa s prilagajanjem spremembam v splošnih označevalnih smernicah in odpravljanjem posamičnih napak je njegova velikost do nedavne razširitve, ki jo predstavimo v 4. razdelku tega prispevka, ostajala ves čas nespremenjena.

3. Popis smernic UD za slovenščino

Splošne smernice UD, kakršne so dokumentirane na krovni spletni strani projekta,[6] so kot nadaljevanje predhodnih standardizacijskih iniciativ in večletnega kolaborativnega razvoja zasnovane tako, da skušajo na čim krajši način nasloviti skladenjske specifike čim širšega nabora jezikov. Tako v splošnih smernicah najdemo predvsem prototipične opredelitve posameznih oznak, opis najbolj tipičnih mejnih primerov in ponazoritve na primerih izbranih jezikov, naloga avtorjev drevesnic za posamezne jezike pa je, da te splošne smernice nato prenesejo na svoje konkretne jezikovne podatke. Pri tem infrastuktura UD omogoča, da se za vsak jezik ta načela popišejo kot jezikovnospecifične smernice na uradni spletni strani, vendar to ni obvezno, zato je dokumentacija označevalnih smernic UD za posamične jezike prepuščena predvsem samoiniciativnosti avtorjev podatkov.

Za slovenščino so bile ob prvi objavi korpusa SSJ-UD tako dokumentirane zgolj smernice za pripisovanje besednih vrst in oblikoskladenjskih oznak, ki so odtlej ob prehodu z UD v1 na UD v2 (Nivre et al., 2020) že nekoliko zastarele, smernice za pripisovanje skladenjskih relacij UD besedilom v slovenščini pa zaradi obsežnosti niso bile podrobneje dokumentirane oz. so bile razvidne zgolj implicitno iz pretvorbenih pravil na eni strani in objavljenega korpusa na drugi.

Prvi korak znotraj projekta RSDO je bil tako namenjen izčrpnemu popisu smernic UD za slovenščino na vseh treh ravneh označevanja (besedne vrste, oblikoskladenjske lastnosti in skladenjske relacije) v obliki priročnika, ki na slovenskih primerih razlaga in ponazarja uporabo posameznih oznak UD za označevanje besedil v slovenščini. Pri tem smo poleg opisa prvotnih smernic uvedli tudi nekaj manjših sprememb na mestih, kjer je bila prvotna označenost korpusa SSJ-UD nedosledna ali neustrezna glede na univerzalne smernice. Med njimi lahko izpostavimo predvsem spremembe v obravnavi primerjalnih struktur (lastnost kot nadredni element primerjave), poudarjalnih členkov (razlikovanje med modifikatorji samostalnikov na eni in povedkov na drugi strani), besedilnih povezovalcev (razlikovanje glede na stavčno pozicijo) in prostega morfema *se*/*si* (razlikovanje med zaimki v predmetni in ekspletivni vlogi), ki

[3] V tem prispevku namesto uradnega imena drevesnice (SSJ) zaradi podobnosti s poimenovanji sorodnih korpusov in projektov v slovenskem prostoru uporabljamo daljši akronim SSJ-UD.

[4] Pravila in skripte za pretvorbo iz sistema JOS v sistem UD so na voljo na `https://github.com/clarinsi/jos2ud`.

[5] V nasprotju s sistemom JOS, znotraj katerega so pojavitve glagola *biti* ne glede na skladenjsko vlogo ali pomen vedno označene kot glagol s podvrsto *pomožni*, sistem UD že na ravni besednih vrst ločuje med glavnimi (oznaka VERB) in pomožnimi glagoli (oznaka AUX), kamor se umeščajo glagoli v vlogi pomožnikov in veznih glagolov.

[6] `https://universaldependencies.org/guidelines.html`

so bili zaradi omejitev strojne pretvorbe iz sistema JOS prvotno označeni drugače kot predvidevajo splošne smernice UD.

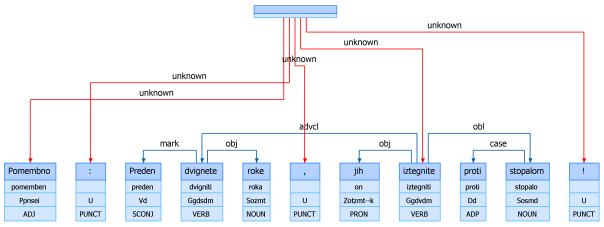
Priročnik s smernicami UD za slovenščino[7] poleg opisov posamičnih slovničnih kategorij in načel njihovega pripisovanja besedilom v slovenščini vsebuje še razdelek s podrobnejšo obravnavo težavnejših primerov, ki se je dopolnjeval tudi skozi označevalno kampanjo, opisano v 4. razdelku. V pripravi je tudi objava slovenskih smernic na uradni spletni strani UD (v angleščini) in popis odprtih vprašanj z izhodiščnimi priporočili za nadaljnje izboljšave (v sodelovanju z Univerzo v Novi Gorici) .

4. Nadgradnja korpusa SSJ-UD

V drugem koraku projekta je sledila označevalna kampanja, v okviru katere smo ročno označili več kot 5.000 novih povedi iz korpusov ssj500k oz. ELEXIS-WSD, nekoliko izboljšana pa je bila tudi označenost prvotne različice korpusa SSJ-UD. V vseh treh fazah je označevanje potekalo v označevalnem orodju Q-CAT (Brank, 2022), ki odslej podpira tudi standardni format CONLL-U, za primerjavo označenih datotek (kuriranje) pa smo uporabili lokalno inštalacijo orodja WebAnno (Eckart de Castilho et al., 2016), ki jo vzdržuje CLARIN.SI.[8] Podrobnejša analiza označevalnega procesa je opisana v prispevku Dobrovoljc in Ljubešić (2022), v nadaljevanju pa predstavimo zgolj najpomembnejše rezultate.

4.1. Razširitev s polpretvorjenimi povedmi iz ssj500k

Kot smo omenili že v 2. razdelku, nekaterih skladenjsko razčlenjenih povedi v korpusu ssj500k zaradi omejitev pretvorbenih pravil ni bilo mogoče v celoti pretvoriti v oznake UD, zato niso bile vključene v prvotno različico drevesnice SSJ-UD, predstavljale pa so logično izhodišče za nadaljnjo širitev podatkov UD za slovenščino. V prvi fazi razširitve so tako označevalci ročno pregledali teh 3.411 polpretvorjenih povedi oz. 96.194 pojavnic, med katerimi jih 22.377 (23,5 %) ni imelo pripisane skladenjske relacije UD. Te so bile za potrebe lažje vizualizacije označene z relacijo *unknown* (slika 1), označevalci (po dva na poved) pa so poleg ustvarjanja novih povezav preverjali tudi ustreznost že obstoječih (pretvorjenih) povezav.



Slika 1: Primer prikaza polpretvorjene povedi iz ssj500k z manjkajočimi relacijami UD (*unknown*) v označevalnem orodju Q-CAT.

Med pojavnicami, ki v izhodišču niso imele pripisane relacije UD, je bila skoraj polovica ločil (*punct*), kar je bilo glede na pretvorbena pravila pričakovano, saj so bila ločila večinoma na relevantno jedro povezana šele po določitvi vseh drugih pojavnic v povedi, zlasti korena povedi (*root*, običajno jedro povedka glavnega stavka ali drug hierarhično najpomembnejši element v povedi), ki predstavlja tudi drugo najpogostejšo vrsto nepretvorjenih pojavnic (12 %). Tej sledita še relaciji *parataxis* (9 %) in *conj* (6 %), ki se uporabljata za povezovanje stavčnih soredij oz. prireedij, torej struktur, kakršnih zgolj s pravili ni bilo mogoče pretvoriti z dovolj zanesljivo natančnostjo.

4.2. Razširitev s povedmi iz korpusa ELEXIS-WSD

V drugi fazi širitve je bil skladenjsko razčlenjen še korpus ELEXIS-WSD-SL, tj. slovenski del paralelnega korpusa ELEXIS-WSD (Martelli et al., 2021; Martelli et al., 2022), razvitega za potrebe strojnega pomenskega razdvoumljanja, ki vsebuje v več evropskih jezikov prevedena besedila iz Wikipedie (Schwenk et al., 2021). Slovenski korpus ELEXIS-WSD vsebuje 2.024 povedi (31.237 pojavnic), ki so bile predhodno že ročno tokenizirane, lematizirane in oblikoskladenjsko označene po sistemu JOS, na podlagi česar smo korpus s pretvorbeno skripto samodejno pretvorili še v besedne vrste in oblikoskladenjske oznake UD, pojavitve glagola *biti* pa razdvoumili ročno.

Tako označen korpus je bil izhodiščno skladenjsko razčlenjen z orodjem CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019), pravilnost strojno pripisanih razčlemb pa so nato pregledali trije označevalci in končni kurator. Na ta način je bilo ročno popravljenih 1.534 (4.91 %) skladenjskih relacij, med katerimi so prevladovale strukture z oznakami *nmod*, *advmod*, *obl*, *conj* in *punct*, kar se, kot bomo videli v nadaljevanju, sklada z najpogostejšimi tipi napak razčlenjevalnika nasploh (8. razdelek).

4.3. Izboljšanje označenosti v prvotnem korpusu

Poleg dodajanja novih razčlenjenih povedi smo glede na rahlo spremembo smernic (3. razdelek), analizo ročnih popravkov pretvorjenih relacij (razdelek 4.1.) in drugih identificiranih nedoslednosti izboljšali tudi označenost izhodiščne različice korpusa SSJ-UD.

Med približno 30 identificiranimi tipi napak oz. nedoslednosti so bile denimo pristavčne strukture, visok delež (neupravičenih) neprojektivnih povezav,[9] nedosledno ločevanje med sorednimi in prireedno vezanimi stavki, med premimi in nepremimi predmeti, itd. Za vsako izmed kategorij smo s hevrističnimi poizvedbami ustvarili podkorpuse povedi s potencialno problematičnimi oznakami, ki so jih nato označevalci ročno pregledali in popravili v skladu s smernicami. Na ta način je bilo v izhodiščnem korpusu popravljenih 1.670 skladenjskih oznak, kar sicer predstavlja razmeroma majhen del celotnega korpusa (1,2 %).

[7] Priročnik je trenutno na voljo v delovni različici, uradno pa bo objavljen ob zaključku projekta RSDO.

[8] https://www.clarin.si/webanno/.

[9] Povezava med besedo A in besedo B je projektivna, če je beseda A posredno nadrejena tudi vsem drugim besedam med A in B – obstaja torej pot od A do vseh besed med A in B. Če si to predstavljamo grafično, se povezave v neprojektivnem drevesu med seboj križajo. To je v jezikih s prostim besednim redom, kot je slovenščina, sicer možen pojav, a vendarle redek.

5. Nova različica korpusa SSJ-UD

V zadnjem koraku smo izhodiščni korpus SSJ-UD z nekoliko izboljšano označenostjo (razdelek 4.3.) združili z novimi povedmi iz korpusov ssj500k (razdelek 4.1.) in ELEXIS-WSD (razdelek 4.2.) ter tako dobili novo različico referenčne univerzalne odvistnostne drevesnice za pisno slovenščino SSJ-UD,[10] ki je bila prvič objavljena kot del uradnega izida UD v2.10 (Zeman et al., 2022). Ob zaključku projekta RSDO bo drevesnica SSJ-UD integrirana tudi v novi referenčni korpus učne slovenščine SUK.

5.1. Sestava korpusa

Kot prikazuje tabela 5.1., nova različica v primerjavi s prvotno vsebuje 5.435 novih razčlenjenih povedi (+67,9 %) oz. skoraj enkrat večje število pojavnic (126.427, +89,9 %), s čimer se korpus SSJ-UD po številu pojavnic danes umešča na 30. mesto med skupno 228 drevesnicami UD. Z razširitvijo je korpus SSJ-UD postal tudi bolj raznolik, saj se vsi trije podkorpusi (izvorne povedi iz ssj500k, nove povedi iz ssj500k, povedi iz ELEXIS-WSD) med seboj razlikujejo tako z vidika vrste vsebovanih besedil kot njihove skladenjske kompleksnosti.

Medtem ko besedila ssj500k kot vzorec korpusa Fida-PLUS (Arhar Holdt, 2007) vsebujejo predvsem izvorno slovenska leposlovna, neleposlovna in publicistična besedila, korpus ELEXIS-WSD vsebuje prevedena enciklopedična besedila iz Wikipedie. Po drugi strani sta si izvorni SSJ-UD in korpus ELEXIS-WSD podobna z vidika kompleksnosti (krajše in skladenjsko enostavnejše povedi), medtem ko so nove povedi iz ssj500k bistveno daljše.

Nenazadnje pa je z metodološkega vidika pomembno izpostaviti še, da se vsi trije podkorpusi razlikujejo tudi z vidika izvora pripisanih oznak UD, saj so oznake prvotnega SSJ-UD v veliki večini rezultat avtomatske pretvorbe iz sistema JOS, oznake novih povedi iz ssj500k kombinacija pretvorbe in ročnega pregleda, oznake povedi iz korpusa ELEXIS-WSD pa so bile v celoti pregledane ročno.

Podkorpus	Povedi	Pojavnice	Povp.
Prvotni SSJ-UD	8.000	140.670	17,58
Novo iz ssj500k	3.411	95.194	27,91
Novo iz ELEXIS-WSD	2.024	31.233	15,43
Skupaj novi SSJ-UD	13.435	267.097	19,88

Tabela 1: Zgradba nove različice korpusa SSJ-UD (od UD v2.10 naprej).

5.2. Delitev podatkovnih množic

Del objave drevesnice v uradni zbirki UD je tudi njena delitev na učno, validacijsko in testno množico, ki se standardno uporabljajo pri razvoju in evalvaciji na teh podatkih temelječih napovednih modelov. Pri tem smo sledili načelom delitve podatkov v prvotni različici, v kateri so bile podmnožice razdeljene glede na zaporedje pojavljanja v korpusu. Glede na to, da so nove povedi iz ssj500k enakomerno razpršene po celotnem korpusu, smo te zgolj priključili k že obstoječi delitvi povedi v prvotni različici in ohranili enako razmerje (80 % učna, 10 % validacijska, 10 % testna), nato pa smo vsaki izmed treh množic v enakem razmerju dodali še povedi iz korpusa ELEXIS-WSD. Sestava podmnožic tako odslikava raznolikost nove različice korpusa SSJ-UD, kakršno opisujemo v razdelku 5.1., in z reprezentativnostjo testnih podatkov glede na učne zagotavlja ustreznejšo, besedilnozvrstno manj pristransko evalvacijo.

6. Razčlenjevalni model

V drugi fazi projekta smo na novi, bistveno večji različici ročno označenega korpusa SSJ-UD naučili tudi nov napovedni model skladenjskega razčlenjevanja po sistemu UD v označevalnem orodju CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019),[11] ki se kot temeljno programsko orodje za označevanje besedil v slovenščini prav tako razvija v okviru projekta RSDO. Gre za izpeljavo odprtokodnega orodja Stanza (Qi et al., 2020), ki v primerjavi z izvornim orodjem uvaja nekatere izboljšave na ravni tokenizacije, oblikoskladenjskega označevanja in lematizacije, skladenjski razčlenjevalnik pa se od izvornega (Dozat in Manning, 2016), ki temelji na nadgrajeni metodi dvosmernega dolgega kratkoročnega spomina (BiLSTM), razlikuje predvsem po uporabi besednih vložitev CLARIN.SI-embed.sl (Ljubešić in Erjavec, 2018), ki so bile naučene na slovenskih besedilih v obsegu 3,5 milijard besed.

Tako pri učenju kot evalvaciji razčlenjevalnega modela smo uporabili ročno označene podatke na nižjih ravneh označevanja (tokenizacija, stavčna segmentacija, oblikoskladenjsko označevanje, lematizacija), saj nas je v tej fazi razvoja razčlenjevalnika zanimala predvsem natančnost napovednega modela v izolaciji, brez vpliva napovednih karakteristik orodja na nižjih ravneh.

Izgradnjo napovednega modela, njegovo primerjavo z modelom, naučenim na prvotni različici SSJ-UD, in evalvacijo glede na posamične podkorpuse podrobneje opisujeta Dobrovoljc in Ljubešić (2022), ki ugotavljata, da je model, naučen na novi različici korpusa SSJ-UD, zaradi povečanega obsega učnih podatkov in njihove diverzifikacije bistveno izboljšan v primerjavi z modelom, naučenim na prvotni različici.

Da bi osvetlili prednosti in pomanjkljivosti uporabe novega razčlenjevalnega modela v različnih jezikovnotehnoloških in jezikoslovnih aplikacijah ter obenem identificirali prioritete za njegove nadaljnje izboljšave, v nadaljevanju prispevka te ugotovitve nadgradimo s podrobnejšo evalvacijo splošne natančnosti modela (7. razdelek) na eni strani in analizo najpogostejših tipov napak (8. razdelek) na drugi.

[10] Čeprav infrastruktura UD dopušča objavo poljubnega števila drevesnic, smo se namesto objave novih drevesnic UD za slovenščino namenoma odločili za priključitev novih povedi k že obstoječi drevesnici SSJ-UD, da bi zagotovili kar najbolj učinkovito izrabo teh podatkov v širši jezikovnotehnološki skupnosti, kjer se zaradi poenostavitve dela modeli pogosto razvijajo zgolj na izbrani, običajno največji, drevesnici nekega jezika.

[11] https://pypi.org/project/classla/

7. Splošna natančnost

Za kvantitativno evalvacijo splošne natančnosti modela smo uporabili standardni protokol, po katerem smo model, naučen na učni oz. validacijski množici uporabili za razčlenjevanje testne množice, napovedane oznake pa nato primerjali z ročno pripisanimi. Za poročanje o natančnosti uporabljamo uveljavljeno metriko LAS (angl. *labeled attachment score*), ki prikazuje delež pojavnic s pravilno napovedano nadrejeno pojavnico in vrsto njunega skladenjskega razmerja, pri čemer ta delež povzemamo z oceno F1, ki prikazuje harmonično sredino med preciznostjo in priklicem.[12]

Rezultati, predstavljeni v tabeli 7., prikazujejo, da razčlenjevalni model dosega splošno natančnost 93,21 LAS F1, kar nekoliko poenostavljeno pomeni, da se model v povprečju na vsakih sto označenih pojavnic zmoti pri manj kot sedmih, tj. jim pripiše napačno nadrejeno pojavnico in/ali vrsto povezave med njima.[13]

Kot prikazujejo rezultati za posamične tipe relacij,[14] pa ta splošna ocena natančnosti ni reprezentativna za vse vrste skladenjskih struktur, saj je pri napovedovanju nekaterih relacij model bistveno natančnejši kot pri drugih.

Med relacijami z najvišjo natančnostjo napovedovanja so po pričakovanju funkcijske besede, kot so predlogi (*case*; 99,17), pomožni glagol *biti* (*aux*; 98,93), določilniški zaimki in prislovi (*det*; 98,79), podredni vezniki (*mark*; 98,69), ekspletivni zaimki (*expl*; 96,71) in priredni vezniki (*cc*; 96,27), skratka, pojavnice, ki se pojavljajo v zelo predvidljivih oblikah in skladenjskih položajih.

Poleg navedenih relacij model razmeroma dobro natančnost dosega tudi pri napovedovanju nekaterih jedrnih skladenjskih struktur, kot so samostalniški predmeti (*obj*; 95.53) in osebki (*nsubj*; 95.28), nadpovprečno uspešen pa je tudi pri identifikaciji korena povedi (*root*; 96,26), ki je običajno jedro povedka glavnega stavka, in veznega glagola *biti* (*cop*; 95,43), ki nastopa v strukturah s povedkovimi določili.

Med relacijami, pri napovedovanju katerih model dosega najslabše rezulate, pričakovano najdemo ogovore (*vocative*; 0,0), saj se v testni množici pojavi zgolj en primer, in nedoločene strukture (*dep*; 54,55), saj se ta oznaka kot skrajna možnost uporablja predvsem za povezovanje obrobnih, iregularnih pojavov, ki jim je nemogoče pripisati katerokoli drugo povezavo (npr. ostanki oštevilčenih strani pri digitalizaciji besedil).

Čeprav se je natančnost označevanja samostalniških pristavčnih določil (*appos*, 63,40), 'osirotelih' stavčnih členov v povedih z glagolsko elipso (*orphan*; 68,24), diskurznih členkov (*discourse*; 69,23), stavčnih soredij (*parataxis*; 70,35) in naštevalnih seznamov (*list*; 75,86) z novo različico korpusa SSJ-UD bistveno izboljšala glede na prvotni model (Dobrovoljc in Ljubešić, 2022), te relacije ostajajo med tistimi z najnižjo natančnostjo, kar je glede na njihovo ohlapnejšo slovnično povezanost s povedkom oz. nadrejenimi stavčnimi členi tudi pričakovano.

Med drugimi relacijami s podpovprečno natančnostjo označevanja lahko izpostavimo še podredne stavke različnih tipov, kot so prislovni (*advcl*; 75,86), prilastkovi (*acl*; 81,73), osebkovi (*csubj*; 85,53) in predmetni odvisniki (*ccomp*; 90,67). Poleg nepremih predmetov (*iobj*; 81,66), ki jih je težavno identificirati predvsem zaradi pomanjkljivosti trenutnih označevalnih smernic,[15] modelu precejšen izziv predstavljajo tudi priredja, zlasti medstavčna (*conj*; 85,91), samostalniški prilastki (*nmod*; 87,44) in prislovna določila povedkov, samostalnikov in pridevnikov (*advmod*; 89,95).

8. Najpogostejše napake

V drugem koraku evalvacije smo analizo zanesljivosti modela pri razčlenjevanju posameznih tipov relacij dopolnili še s podrobnejšo analizo najpogostejših tipov napak. Tabela 8. tako povzema distribucijo napak glede na to, pri katerem izmed obeh napovedanih podatkov (identifikator nadrejene pojavnice in vrsta skladenjske relacije med njima) se je model dejansko zmotil. Za vsak tip napake navajamo tudi pet najpogostejših podtipov glede na relacije, pri katerih se pojavlja, pri čemer štetje prikazujemo združeno za napake v obe smeri (npr. *obl-nmod* vključuje tako napovedovanje *obl* namesto *nmod* kot napovedovanje *nmod* namesto *obl*).

Identificirane pogoste tipe napak znotraj vsake kategorije na podlagi ročne analize napačno označenih primerov opišemo v nadaljevanju, pri čemer podrobneje predstavimo predvsem najpogostejše.

8.1. Napačna napoved nadrejenega elementa

Kot prikazuje tabela 8., dobro polovico (52,8 %) predstavljajo napake, pri katerih je model pravilno napovedal skladenjsko vlogo pojavnice (pravilno relacijo oz. oznako), zmotil pa se je pri napovedi njenega nadrejenega elementa (jedra oz. izvora relacije).

Najpogostejša napaka pri določanju nadrejenega elementa je povezana z relacijo **punct**, ki označuje ločila. Večinoma gre za primere, kjer so napačno določena tudi

[12] Izračuni temeljijo na uradni evalvacijski skripti tekmovanja CoNLL Shared Task 2018 (Zeman et al., 2018), ki smo jo dodatno prilagodili tako, da poleg splošnega izračuna natančnosti vrača tudi rezultate za posamične skladenjske relacije in druge relevantne oznake.

[13] Ta natančnost je v skladu z natančnostjo orodja Stanza za druge jezike oz. drevesnice (`https://stanfordnlp.github.io/stanza/performance.html`) oz. natančnostjo drugih sodobnih razčlenjevalnikov nasploh (`https://universaldependencies.org/conll18/results.html`), vendar neposredna primerjava zaradi specifik evalvacijske metodologije ni smiselna.

[14] V Tabeli 7. ni relacije *compound*, ki je glede na smernice v slovenščini ne uporabljamo. Pri relacijah *dislocated*, *goeswith* in *reparandum* podatkov o natančnosti ni (oznaka *n/a*), saj se v testni množici ne pojavijo. O natančnosti izpeljanih relacij oz. podoznak (npr. *flat:name*, *flat:foreign*) poročamo združeno z jedrno oznako (npr. *flat*).

[15] Zaradi kompleksnega prepletanja oblikoslovnih, skladenjskih in pomenskih razločevalnih lastnosti med premimi in nepremimi predmeti trenutne smernice UD priporočajo, da je v povedih z zgolj enim izraženim predmetom ta ne glede na sklon ali udeleženso vlogo označen kot premi predmet (*obj*). To pomeni, da se lahko tudi predmeti v dajalniku, kakršni tipično nastopajo kot nepremi predmeti, ob odsotnosti drugih predmetov označujejo kot premi.

Relacija	Izvorni opis	Slovenski prevod	LAS F1
acl	clausal modifier of noun	stavčni prilastki	81,73
advcl	adverbial clause modifier	prislovni odvisniki	75,86
advmod	adverbial modifier	prislovna določila (gl. opombo 16)	89,95
amod	adjectival modifier	pridevniški prilastki	98,9
appos	appositional modifier	pristavčna določila	63,4
aux	auxiliary verb	pomožni glagoli	98,93
case	case marking preposition	predlogi	99,17
cc	coordinating conjunction	priredni vezniki	96,27
ccomp	clausal complement	stavčna dopolnila (predmetni odvisniki)	90,67
conj	conjunct	priredno zloženi elementi	85,91
cop	copula verb	vezni glagoli	95,43
csubj	clausal subject	osebkovi odvisniki	85,53
dep	unspecified dependency	nedoločena povezava	54,55
det	determiner	določilniki	98,79
discourse	discourse element	diskurzni členki	69,23
dislocated	dislocated element	dislocirani elementi	n/a
expl	expletive	ekspletivne besede	96,71
fixed	fixed multi-word expression	funkcijske zveze	93,33
flat	flat multi word-expression	eksocentrične zveze	92,12
goeswith	disjointed token	razdruženi deli besed	n/a
iobj	indirect object	nepremi predmeti	81,66
list	list	seznami	75,86
mark	marker (subordinating conjunction)	podredni vezniki	98,69
nmod	nominal modifier	samostalniški prilastki	87,44
nsubj	nominal subject	samostalniški osebki	95,28
nummod	numeric modifier	številčna določila	94,23
obj	(direct) object	premi predmeti	95,53
obl	oblique nominal (adjunct)	odvisne samostalniške zveze	91,14
orphan	dependent of missing parent	elementi v eliptičnih strukturah	68,24
parataxis	parataxis	stavčna soredja	70,35
punct	punctuation symbol	ločila	93,08
reparandum	overriden disfluency	samopopravljanja	n/a
root	root element	koren povedi	96,26
vocative	vocative	ogovori	0
xcomp	open clausal complement	odprta stavčna dopolnila	92,87
Vse relacije			**93,21**

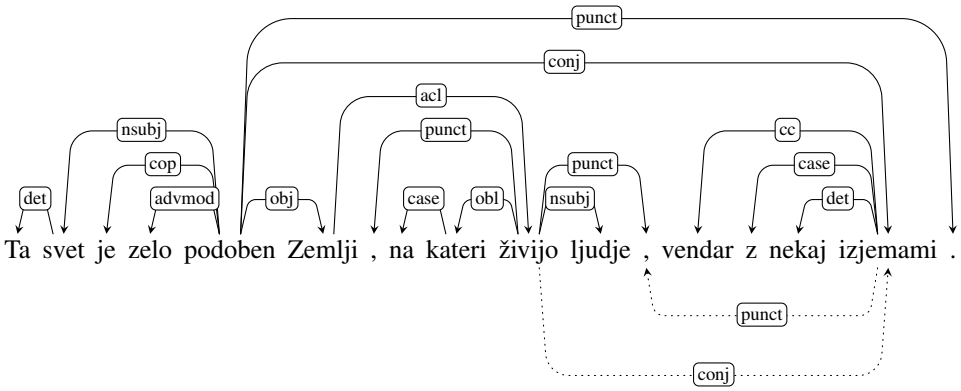
Tabela 2: Natančnost novega modela orodja CLASSLA-Stanza za skladenjsko razčlenjevanje po sistemu UD glede na metriko LAS.

jedra drugih struktur v povedi, na katera se ločila praviloma povezujejo. Napačno povezana ločila so torej predvsem posledica napak razčlenjevanja njihovih nadrejenih struktur, kot prikazuje primer na sliki 2, pri katerem razčlenjevalnik zadnji stavek zmotno interpretira kot priredje pred njim stoječega odvisnika, čemur ustreza tudi (napačno) označena vejica.

Druga pogosta skupina je povezana s t.i. poudarjalnimi členki oz. prislovi, kot so besedice *tudi*, *še*, *le*, *že* idr., ki jim pripisujemo relacijo **advmod**,[16] njihova stava pa je v slovenščini razmeroma prosta – modificirajo lahko tako povedek kot posamezne stavčne člene, kar je pogosto mogoče razbrati šele iz konteksta ali prozodičnih poudarkov pri branju. Kot prikazuje primer na sliki 3, razčlenjevalnik te besede namesto na poudarjeni samostalnik pogosto veže na povedek stavka. To ni presenetljivo, glede na to, da gre za eno izmed kategorij, pri kateri so se označevalci najpogosteje razhajali, prav tako pa je bila nedosledno označena v prvotnem korpusu, v katerem so bile ob pretvorbi te pojavnice ne glede na vlogo vedno povezane na povedek.

Pri preostalih treh analiziranih relacijah s pogosto napačno pripisanim izvorom povezave, tj. **nmod**, **conj** in **acl**, prihaja do podobne napake: razčlenjevalnik zanesljivo prepozna vrsto nadrejene strukture (npr. samostalniške zveze, pridevniške zveze ali povedki), vendar namesto prave strukture kot jedro izbere najbližjo ustrezno zvezo na levi, kar ni vedno prav, saj se včasih pravi izvor relacije v povedi pojavi že prej (slika 4).

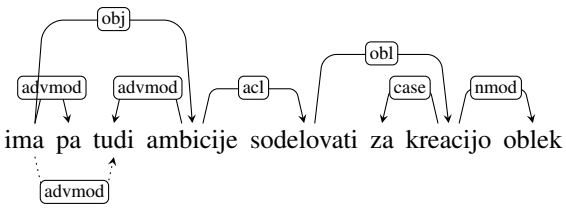
[16] Relacija *advmod* se uporablja za označevanje prislovov v vlogi modifikatorjev, kar vključuje tako prislove v vlogi okoliščinskih dopolnil povedkov (kakršna Slovenska slovnica imenuje prislovna določila, npr. *pridem takoj*) kot prislove v vlogi modifikatorjev pridevniških, prislovnih ali samostalniških besednih zvez (prislovni prilastki, npr. *izjemno prilagodljiv*).



Slika 2: Primer razhajanja med ročno (zgoraj) in strojno (spodaj) pripisanim jedrom relacije *punct*.

Tip napake	Število napak
Napačno jedro	**914**
punct-punct	248
advmod-advmod	166
nmod-nmod	111
conj-conj	99
acl-acl	53
Napačno jedro in oznaka	**517**
obl-nmod	141
parataxis-root	37
acl-advcl	22
root-nsubj	22
nsubj-nmod	19
Napačna oznaka	**299**
conj-parataxis	23
obl-nsubj	19
appos-conj	17
obl-obj	13
iobj-obj	13
Vse napake	**1730**

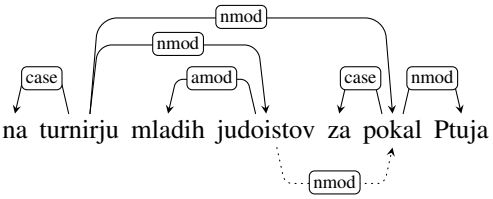
Tabela 3: Distribucija napak razčlenjevalnega modela glede na tip napake.



Slika 3: Primer napačne razčlembe poudarjalnih členkov (*advmod* zgoraj) kot prislovnih določil povedka (*advmod* spodaj).

8.2. Napačna napoved nadrejenega elementa in relacije

Po pogostosti sledijo napake, pri katerih se je model zmotil tako pri napovedi nadrejene pojavnice kot njune skladenjske relacije (29,9 %). Med njimi najbolj izstopa



Slika 4: Primer razhajanja med ročno (zgoraj) in strojno (spodaj) identificirano odnosnico predložne zveze v vlogi desnega prilastka (*nmod*).

zamenjevanje struktur z oznakama **obl**[17] in **nmod**, ki predstavlja tretji najpogostejši (pod)tip napak nasploh. Analiza primerov kaže, da gre večinoma za primere, v katerih predložna zveza v vlogi prislovnega določila povedka (**obl**) stoji tik za neko samostalniško zvezo, model pa prislovno določilo napačno tolmači kot njen desni prilastek, za katere se uporablja relacija **nmod**, kot prikazuje primer na sliki 5.

Manj pogoste v tej kategoriji so še napake pri določanju glavnega stavka v nizu dveh ali več soredno zloženih stavkov, zlasti kadar gre za vrinjene stavke ali premi govor (**parataxis-root**), napake ločevanja med prislovno-določilnimi odvisniki in stavčnimi prilastki, pogosto v kombinaciji z veznikom *kot* (**acl-advcl**), zamenjava osebka in povedkovega določila v strukturah z veznim glagolom *biti* (**root-nsubj**) in napake določanja osebka v povedih, kjer osebek ni eksplicitno izražen (**nsubj-nmod**).

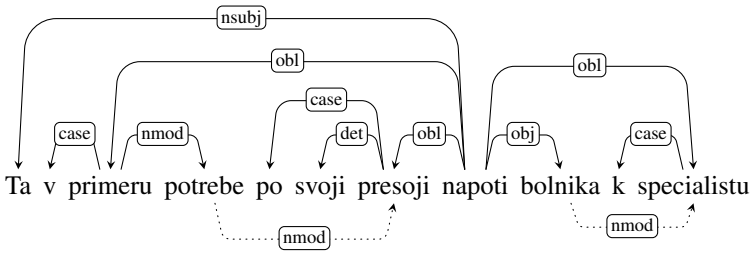
8.3. Napačna napoved relacije

Med vsemi tremi kategorijami napak pa je najmanj takih, pri katerih je razčlenjevalnik pojavnico povezal s pravim nadrejenim elementom, a tej relaciji pripisal napačno oznako (17,3 %). V primerjavi s prvima dvema kategorijama so tukaj tipi glede na relacije razpršeni bolj enakomerno.

Do zamenjav oznak **conj** in **parataxis**[18] prihaja pred-

[17] Relacija *obl* se uporablja za odvisne samostalniške in predložne zveze, ki nastopajo v vlogi nejedrnih argumentov povedka. Poleg teh se s to relacijo označujejo tudi neglagolske strukture s primerjalnimi vezniki.

[18] Relacija *parataxis* se uporablja za označevanje stavčnih soredij različnih vrst. To so razmerja med besedo (običajno jedrom glavnega stavka) in drugimi elementi, ki z njo niso v priredju, podredju ali kateremkoli drugem jedrnem slovničnem razmerju.



Slika 5: Primer napačne razčlembe predložnih prislovnih določil (*obl* zgoraj) kot desnih prilastkov (*nmod* spodaj).

vsem pri daljših povedih, pri katerih se med dva priredno zložena stavka oz. med priredni veznik in drugi stavek v priredju vrivajo druge strukture (npr. odvisniki). Samostalniška prislovna določila (ki prejmejo relacijo **obl**) so napačno označena kot osebki (**nsubj**) predvsem v zvezah z glagoli, kot so *imenovati*, *praviti*, idr., v katerih se pojavljajo v imenovalniku (npr. *pravimo jim mikroznaki*).

Med drugimi tipi napačno pripisanih relacij je pogosta še dvoumnost med samostalniškimi zvezami v vlogi pristavčnih določil (**appos**) na eni in priredno povezanih elementov (**conj**) na drugi strani, zlasti kadar zadnji element v brezvezniškem priredju stoji na koncu povedi. Pojavljajo se tudi napake ločevanja med prislovnimi določili in predmeti, predvsem pri samostalniških zvezah, ki izražajo časovni oz. prostorski okvir dogodka (**obl-obj**) in pa napačno določanje premega (**obj**) in nepremega predmeta (**iobj**).

9. Zaključek

V prispevku smo predstavili nadgradnjo drevesnice SSJ-UD, referenčnega ročno skladenjsko razčlenjenega korpusa po medjezično usklajeni shemi Universal Dependencies, v okviru katere smo po rahli prenovi in izčrpni dokumentaciji označevalnih smernic za slovenščino korpus razširili z novimi povedmi ter nato na novi učni množici naučili tudi nov napovedni model za skladenjsko razčlenjevanje slovenskih besedil. Podrobna kvantitativna in kvalitativna analiza njegove natančnosti je pokazala, da model v splošnem dosega razmeroma dobre rezultate, pri čemer je pri členjenju nekaterih struktur mogoče pričakovati bistveno večjo zanesljivost rezultatov kot pri drugih.

Glede na mednarodno relevantnost sheme UD rezultati predstavljajo pomemben doprinos k nadaljnjemu razvoju jezikovnih tehnologij za slovenščino tako v slovenskem kot mednarodnem prostoru, saj je glede na odprti dostop in standardizirano distribucijo drevesnic UD mogoče pričakovati, da bodo novi podatki za slovenščino kmalu integrirani tudi v številna druga razčlenjevalna orodja oz. na njih temelječe aplikacije (npr. Nguyen et al. (2021)). Poleg modelov za skladenjsko razčlenjevanje, kakršnega smo predstavili v tem prispevku, je skoraj enkrat večja količina učnih podatkov za slovenščino neprecenljiva tudi za nadaljnji razvoj modelov za lematizacijo in oblikoslovno označevanje po sistemu UD, ki v mednarodnem prostoru večinoma temeljijo zgolj na uradno izdanih drevesnicah UD, kot je SSJ-UD, ne pa virih, ki so bili razviti oz. distribuirani v lokalnem kontekstu, kot je denimo celotni korpus ssj500k oz. nastajajoči učni korpus SUK.

Čeprav je bila shema UD prvotno vzpostavljena predvsem za potrebe jezikovnotehnoloških raziskav, pa številne odmevne primerjalnojezikoslovne študije dokazujejo tudi njeno relevantnost na področju jezikoslovja, vključno s slovenistiko, kjer metodološki potencial skladenjsko razčlenjenih korpusov doslej še ni bil polno izkoriščen (Ledinek, 2018). Verjamemo, da izčrpno dokumentirane smernice, obsežen ročno označen korpus in sistematična evalvacija natančnosti na njem naučenega modela predstavljajo pomemben doprinos k nadaljnjim jezikoslovnim raziskavam ročno in strojno razčlenjenih slovenskih korpusov, pri čemer je glede na kompleksno strukturo tovrstnih korpusov nujno vzpostaviti tudi ustrezno infrastrukturo za njihovo analizo.

Seveda pa je tako z vidika jezikovnotehnološke kot jezikoslovne uporabe predstavljene rezultate smiselno kontinuirano nadgrajevati tudi v prihodnje, kar vključuje tako izboljšavo izhodiščnih smernic na eni strani kot njihovo dosledno implementacijo na drugi. Glede na v prispevku predstavljene metodološke razlike v nastanku posamičnih delov korpusa SSJ-UD in zaznane nedoslednosti med kvalitativno analizo napak je poleg nadaljnjega povečevanja korpusa vsekakor enako smiselna tudi konsolidacija obstoječega.

10. Zahvala

Predstavljeno delo sta podprla projekt Razvoj slovenščine v digitalnem okolju, ki ga financirata Ministrstvo za kulturo Republike Slovenije in Evropski sklad za regionalni razvoj, ter raziskovalni program Jezikovni viri in tehnologije za slovenski jezik (št. P6-0411), ki ga financira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Zahvala gre tudi označevalcem novih podatkov (Tina Munda, Ina Poteko, Rebeka Roblek, Luka Terčon, Karolina Zgaga) ter Tomažu Erjavcu, Luku Krsniku, Cyprianu Laskowskemu in Mihaelu Šinkcu za tehnično podporo.

11. Literatura

Špela Arhar Holdt. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52(2).

Janez Brank. 2022. Q-CAT corpus annotation tool 1.3. Slovenian language resource repository CLARIN.SI.

Xinying Chen in Kim Gerdes. 2018. How do Universal Dependencies distinguish language groups. *Quantitative Analysis of Dependency Structures*, 72:277–294.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre in Daniel Zeman. 2021. Universal Dependencies. *Computational linguistics*, 47(2):255–308.

Kaja Dobrovoljc, Tomaž Erjavec in Simon Krek. 2016. Pretvorba korpusa ssj500k v univerzalno odvisnostno drevesnico za slovenščino. V: *Proceedings of the Conference on Language Technologies and Digital Humanities*.

Kaja Dobrovoljc, Tomaž Erjavec in Simon Krek. 2017. The Universal Dependencies Treebank for Slovenian. V: *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, BSNLP@EACL 2017*, str. 33–38.

Kaja Dobrovoljc, Tomaž Erjavec in Nikola Ljubešić. 2019. Improving UD processing via satellite resources for morphology. V: *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, str. 24–34, Paris, France, August. Association for Computational Linguistics.

Kaja Dobrovoljc in Nikola Ljubešić. 2022. Extending the SSJ Universal Dependencies treebank for Slovenian: Was it worth it? V: *Proceedings of the 16th Linguistic Annotation Workshop (LAW 2022)*, June.

Kaja Dobrovoljc in Joakim Nivre. 2016. The Universal Dependencies treebank of spoken Slovenian. V: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, str. 1566–1573, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Timothy Dozat in Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank in Chris Biemann. 2016. 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, December. The COLING 2016 Organizing Committee.

Tomaž Erjavec, Darja Fišer, Simon Krek in Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. V: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Richard Futrell, Kyle Mahowald in Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Nancy Ide in James Pustejovsky. 2017. *Handbook of linguistic annotation*, zvezek 1. Springer.

Dan Jurafsky in James H. Martin. 2021. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 3nd Edition Draft*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Polona Gantar, Špela Arhar Holdt, Jaka Čibej in Janez Brank. 2020. The ssj500k training corpus for Slovene language processing. V: *Proceedings of the Conference on Language Technologies and Digital Humanities*, str. 24–33, Ljubljana, Slovenia, September. Institute of Contemporary History.

Nina Ledinek. 2018. Skladenjska analiza slovenščine in slovenski jezikoslovno označeni korpusi. *Jezik in slovstvo*, 63(2/3).

Nikola Ljubešić in Kaja Dobrovoljc. 2019. 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, August. Association for Computational Linguistics.

Nikola Ljubešić in Tomaž Erjavec. 2018. Word embeddings CLARIN.SI-embed.sl 1.0. Slovenian language resource repository CLARIN.SI.

Federico Martelli, Roberto Navigli, Simon Krek, Carole Tiberius, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langements, Kristina Koppel, Tiiu Üksik, Kaja Dobrovolijc, Rafael-J. Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamas Varadi, András Györffy, Simon László, Valeria Quochi, Monica Monachini, Francesca Frontini, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej in Tina Munda. 2021. Designing the ELEXIS parallel sense-annotated dataset in 10 European languages. V: *eLex 2021 Proceedings*, eLex Conference. Proceedings. Lexical Computing CZ.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej in Tina Munda. 2022. Parallel sense-annotated corpus ELEXIS-WSD 1.0. Slovenian language resource repository CLARIN.SI.

Matías Guzmán Naranjo in Laura Becker. 2018. Quantitative word order typology with UD. V: *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*, št. 155, str. 91–104. Linköping University Electronic Press.

Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh in Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. V: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers in Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. V: *Proceedings of the 12th Language Resources and Evaluation Conference*, str.

4034–4043, Marseille, France, May. European Language Resources Association.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton in Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong in Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. V: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, str. 1351–1361, Online, April. Association for Computational Linguistics.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre in Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. V: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, str. 1–21, Brussels, Belgium, October. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg et al. 2022. Universal dependencies 2.10. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (U´ FAL), Faculty of Mathematics and Physics, Charles University.

Primerjava načinov razcepljanja besed v strojnem prevajanju slovenščina–angleščina

Gregor Donaj, Mirjam Sepesy Maučec

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška cesta 46, 2000 Maribor
gregor.donaj@um.si; mirjam.sepesy@um.si

Povzetek

V nevronskih strojnih prevajalnikih smo pri današnji tehnologiji grafičnih procesnih enot omejeni z velikostjo slovarja, kar zmanjšuje kakovost prevodov. Uporaba podbesednih enot rešuje problem velikosti slovarja in pokritosti jezika. Z nadaljnjim razvojem tehnologije pa omejenost slovarja in uporaba podbesednih enot izgubljata pomen. V članku predstavljamo različne metode razcepljanja besed na podbesedne enote z različno velikimi slovarji in primerjamo njihovo uporabo v strojnem prevajalniku za jezikovni par slovenščina–angleščina. V primerjavo vključujemo tudi prevajalnik brez razcepljanja besed. Predstavljamo rezultate uspešnosti prevajanja, hitrosti učenja in prevajanja ter velikosti modelov.

A Comparison of Word Splitting Methods for Slovene-English Machine Translation

Given today's technology for graphical processing units, neural machine translation systems can use only a limited vocabulary, negatively affecting translation quality. The use of subword units can alleviate the problems of vocabulary size and language coverage. However, with further technological development, the limited vocabulary and the use of subword units are losing significance. This paper presents different word splitting methods with different final vocabulary sizes. We apply these methods to the machine translation task for the Slovene-English language pair and compare them in terms of translation quality, training and translation speed, and model size. We also include a comparison with word-based translation models.

1. Uvod

Strojno prevajanje je v zadnjem desetletju doseglo pravi razcvet, zahvaljujoč predvsem vedno večjim zbirkam dvojezičnih korpusov in razpoložljivosti vedno večje računske moči, ki omogoča učenje kompleksnih nevronskih mrež.

Danes najbolj intenzivno raziskovani pristopi strojnega prevajanja temeljijo na nevronskih mrežah (Stahlberg, 2020). Uveljavile so se tri osnovne arhitekture: nevronske mreže s povratno zanko (RNN – Recurrent Neural Network), konvolucijske nevronske mreže (CNN – Convolutional Neural Network) in arhitekture s samo-pozornostjo (self-attention).

Uporaba nevronskih mrež pa prinaša tudi tehnične izzive. Zaradi računske zahtevnosti je v praksi nujna uporaba grafičnih procesnih enot (GPU – Graphical Processing Unit). Le-te pa imajo omejeno velikost delovnega spomina, zaradi česar ne moremo uporabljati poljubno velikih nevronskih mrež. Velikost nevronske mreže v strojnem prevajanju je odvisna od izbrane arhitekture, nastavitev hiperparametrov nevronske mreže in velikosti slovarja. Omejena velikost slovarja pa pomeni slabo pokritost besedišča jezika in posledično dodatne napake pri prevajanju. Tudi jeziki, med katerimi prevajamo, predstavljajo različne izzive in imajo določene specifične lastnosti.

V tem članku bomo preizkusili različne podatkovno vodene metode za razcepljanje besed, s katerimi zmanjšamo velikost slovarja. Izbrali smo metode, ki so dobro znane in uveljavljene, vendar pa temeljijo na precej različnih optimizacijskih kriterijih. Te metode bomo uporabili na primeru strojnega prevajanja med angleščino in slovenščino. Predstavili bomo rezultate v smislu uspešnosti prevajanja, hitrosti učenja in prevajanja ter velikosti izdelanih modelov in njihove porabe pomnilnika GPU. Vse metode bomo tudi primerjali z besednim modelom brez razcepljanja.

2. Slovarske enote v strojnih prevajalnikih

Najbolj intuitivna izbira slovarske enote prevajalnika je beseda, ki je tudi najpogosteje izbrana osnovna enota v drugih postopkih jezikovnih tehnologij. Prinaša pa številne izzive. Za dovolj dobro pokritost besedišča jezika so potrebni veliki slovarji, kar je še posebej izrazit problem pri visoko pregibnih jezikih. Posledica premajhnih slovarjev pa je visok delež neznanih besed oz. besed izven slovarja, ki močno zmanjša kakovost prevodov.

Za obvladovanje omenjenih težav so bile predlagane različne alternativne slovarske enote. Kot najmanjša slovarska enota je bila uporabljena črka oz. znak, ki se je izkazal kot zelo robustna enota, manj občutljiva na šum in razlike v domeni učnega in testnega korpusa (Heigold et al., 2018; Gupta et al., 2019). Potrebne pa so določene prilagoditve v arhitekturi nevronske mreže, saj je dolžina segmenta nekajkrat daljša od segmenta, ki kot slovarske enote uporablja besede. Posledica je slabše modeliranje odvisnosti na velikih razdaljah v besedilih.

Preizkušene so bile tudi slovarske enote, ki so po velikosti med črko in besedo. Pod-besedne enote, dobljene s podatkovno vodenim razcepljanjem, ki kot enoto ohranja pogosta zaporedja črk, so se v splošnem izkazale kot najbolj učinkovite, saj v večji meri ohranjajo sintaktične in semantične lastnosti (Sennrich et al., 2016; Banerjee in Bhattacharyya, 2018). Ker je beseda lahko razcepljena na več različnih načinov, je bila predlagana tudi metoda regulacije razcepljanja (Kudo, 2018). Kot slovarsko enoto bi lahko uporabili tudi jezikoslovno enoto morfem, vendar bi za to potrebovali slovnično znanje.

2.1. Postopek Byte-Pair Encoding

Postopek BPE (Byte-Pair Encoding) je v izvoru postopek za stiskanje podatkov, ki deluje z iterativno zamenjavo najpogostejših parov simbolov z novim simbolom. Sennrich in drugi (Sennrich et al., 2016) so priredili ta algoritem za razcepljanje besed.

V postopku se najprej inicializira slovar, ki vsebuje vse črke in druge znake (številke, ločila), ki se pojavijo v korpusu, ter simbol za konec besede. Vsebina korpusa se obravnava kot zaporedje simbolov, ki so v prvem koraku le črke in drugi znaki. Nato sledi iterativni postopek, v katerem se najde najpogostejši par zaporednih simbolov in se le-ta nadomesti z novim simbolom. Te iterativne korake imenujemo združevanja. Pri postopku pa nimamo združevanj, ki bi vključevala simbol za konec besede, kar v končnem korpusu prepreči združevanje besed, namesto njihovega razcepljanja.

Parameter postopka je število združevanj, ki neposredno vpliva na velikost končnega slovarja. Natančna velikost končnega slovarja je nato enaka vsoti števila združevanj in števila znakov v začetnem slovarju.

Vsaka beseda v korpusu se pri uporabi modela razcepi na enote iz slovarja BPE. Ker pa so v tem slovarju tudi posamezne črke, je skoraj zagotovljeno, da bo delež (pod-) besednih enot izven slovarja po razcepljanju enak 0. Izjeme so zelo redke in se lahko pojavijo, kadar v testnem besedilu vidimo črko ali znak, ki ga ni v učnem korpusu.

Avtorji v (Sennrich et al., 2016) so predstavili implementacijo tega algoritma in predlagali možnost skupnega učenja razcepljanja (Joint BPE), kjer uporabimo besedilo obeh strani vzporednega korpusa kot učno gradivo za model razcepljanja. Tako tvorimo en model in dva slovarja, ločena za vsak jezik v paru. Nastavitev števila združevanj pa nato ustreza skupnemu številu združenih simbolov za oba jezika. Ni pa nujno, da se vsi združeni simboli pojavijo v obeh jezikih. Posledično sta slovarja v tem primeru tipično manjša od števila združevanj.

2.2. Morfessor

Program Morfessor (Creutz in Lagus, 2002) je bil razvit v želji po razcepljanju besed v kompleksnih jezikih na pod-besedne enote, ki približno ustrezajo morfemom – najmanjšim enotam besede, ki nosijo pomen. Želja je bila imeti podatkovno voden postopek, ki deluje za več jezikov brez dodatnega slovničnega znanja. Namen je bil zgraditi slovar jezikovnih enot, ki je manjši in bolj splošen kot pa slovar besed.

Predpostavka delovanja algoritma je, da so besede sestavljene iz zaporedja več segmentov, kot je to tipično v aglutinativnih jezikih. Razvita sta bila dva algoritma. Prvi temelji na principu najkrajše dolžine opisa, drugi pa na principu največje verjetnosti. Uporabili smo prvega.

Cilj algoritma je najti slovar pod-besednih enot, ki daje optimalno vrednost funkcije cene (cost function), ki vsebuje dva dela: ceno izvornega besedila T in ceno slovarja V. Ceno opišemo z

$$C = \mathrm{Cost}(T) + \mathrm{Cost}(V) = \\ = \sum_{m_i \in \mathrm{besedilo}} -\log p(m_i) + \sum_{m_j \in \mathrm{slovar}} k \cdot l(m_j), \quad (1)$$

kjer so m pod-besedne enote, $l(m_j)$ dolžina pod-besedne enote m_j (število črk) in k število bitov, ki so potrebni za predstavitev ene črke in ki ga v praksi lahko postavi na 5. Verjetnost posamezne pod-besedne enote v besedilu $p(m_i)$ se izračuna z oceno največje verjetnosti kot razmerje med absolutno frekvenco te enote in številom vseh enot v besedilu.

V svojem delu smo uporabljali novejšo implematacijo programa – Morfessor 2.0 (Virpioja et al., 2013). Iskalni algoritem v tej implementaciji poišče nabor pod-besednih enot, ki optimizirajo funkcijo cene, pri tem pa lahko ali ročno izbiramo uteži za obe komponenti funkcije cene ali pa izberemo želeno velikost slovarja.

2.3. Unigramski model

Zadnja metoda, ki smo jo pogledali, je razcepljanje besed na podlagi uporabe unigramskega modela (Kudo, 2018). V unigramskemu modelu je verjetnost zaporedja pod-besednih enot **x** modelirana kot produkt verjetnosti posameznih enot tega zaporedja:

$$P(\mathbf{x}) = \prod_{i=1}^{M} p(x_i), \quad (2)$$

kjer je M dolžina besedila, $p(x_i)$ pa verjetnost i-te enote v besedilu. Pri tem spadajo vse pod-besedne enote v določen slovar in vsota verjetnost vseh enot mora biti enaka 1.

Najverjetnejše razcepljanje **x*** besed vhodnega besedila X je tisto, za katero velja

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{S}(X)} P(\mathbf{x}), \quad (3)$$

kjer je $\mathcal{S}(X)$ množica vseh možnih razcepljanj besed iz besedila X.

Verjetnosti posameznih unigramov pod-besednih enot lahko določimo z algoritmom EM (Expectation Maximization), optimalno razcepljanje besed pa najdemo z Viterbijevim algoritmom (Kudo, 2018).

Primer implementacije opisanega postopka je v orodju SentencePiece (Kudo in Richardson, 2018), v katerem so sicer implementirani še drugi postopki, vključno z BPE. V tem orodju lahko prav tako izhajamo iz želene velikosti končnega slovarja.

2.4. Izbrane metode in orodja

Za naše eksperimente smo se odločili, da izberemo 4 metode razcepljanja besed:

- Joint BPE – postopek BPE s skupnim učenjem na vzporednem korpusu in implementacijo Rica Sennricha, imenovano Subword NMT.
- Morfessor – postopek na principu najkrajše dolžine opisa, kjer se uteži v funkciji cene prilagodijo glede na želeno velikost slovarja in implementacijo Morfessor 2.0.
- SentencePiece – BPE – implementacija postopka BPE z ločenim učenjem v orodju SentencePiece.
- SentencePiece – Unigram – postopek na podlagi unigramskih jezikovnih modelov in implementacija v orodju SentencePiece.

Nastavitev	Joint BPE (sl)	Morfessor (sl)	SP-BPE (sl)	SP-Unigram (sl)	Joint BPE (en)	Morfessor (en)	SP-BPE (en)	SP-Unigram (en)
10k	11.384	18.670	17.064	17.814	11.556	18.405	16.909	17.358
15k	16.273	28.251	25.525	26.716	15.739	27.822	25.455	26.177
20k	21.101	37.934	33.561	35.534	19.631	37.664	33.595	34.779
25k	25.883	46.879	41.297	44.175	23.299	46.298	41.395	43.051
30k	30.625	55.438	48.822	52.717	26.760	55.994	48.946	51.204
40k	39.890	73.766	63.478	69.530	33.593	73.960	63.132	66.839
50k	49.063	93.726	77.520	86.111	40.115	90.248	76.515	82.082
60k	58.155	109.989	91.015	102.242	46.404	105.558	89.312	96.924
80k	76.152	143.018	117.134	133.892	58.788	133.679	113.496	125.572
100k	93.938	174.026	142.294	164.877	71.043	159.419	136.198	153.190
120k	111.646	205.658	166.442	195.155	82.972	182.895	157.987	180.256
150k	138.006	238.620	201.334	239.515	101.013	210.425	188.859	218.140

Tabela 1: Velikost izdelanih slovenskih (sl) in angleških (en) slovarjev.

```
Joint BPE:     države  članice  bodo pregle-dale  sezna-me in od izdaja-te-ljice ...
Morfessor:     držav-e članic-e bodo pregled-a-le seznam-e in od izdajatelj-ice  ...
SP - BPE:      države  članice  bodo pregleda-le  sezna-me in od izdaja-telji-ce ...
SP - Unigram:  države  članice  bodo pregleda-le  seznam-e in od izdajatelj-ice  ...
```

Slika 1: Primer segmenta besedila iz testne množice z razcepljenimi besedami.

3. Eksperimentalni sistem

3.1. Korpusi

Eksperimenti so bili izvedeni na prosto dostopnem vzporednem korpusu ParaCrawl (Bañón et al., 2020). Korpus je bil zgrajen s spletnim pajkanjem (Web Crawling) in samodejno poravnavo. Za jezikovni par angleščina-slovenščina vsebuje približno 3,7 milijona poravnanih segmentov, kar predstavlja 65,5 milijona besed na angleški in 60,9 milijona besed na slovenski strani.

Korpus smo razdelili na 3 dele: učni korpus, razvojni korpus in testni korpus. Razvojni korpus je namenjen sprotni validaciji tekom učenja strojnega prevajalnika, testni korpus pa končnemu testiranju in vrednotenju rezultatov. Za vsakega izmed teh dveh korpusov smo izbrali 2.000 naključnih segmentov besedila iz izvornega korpusa. Preostanek je bil uporabljen kot učni korpus.

Nad vsemi deli korpusov smo izvedli standardno predprocesiranje za strojno prevajanje: čiščenje, normalizacijo ločil, tokenizacijo in truecasing[1]. Pri tem je bil učni korpus uporabljen tudi za učenje modela za truecasing. Končne velikosti vseh predprocesiranih korpusov so predstavljene v tabeli 3.1..

3.2. Razcepljanje besed

Pri razcepljanju besed smo uporabili orodja, ki so opisana v prejšnjem poglavju. Učni del korpusa smo uporabili za učenje modela razcepljanja, nato smo naučene modele uporabili za razcepljanje vseh delov korpusa. Tako smo dobili različice razcepljenih korpusov.

[1]Določanje pravilnega zapisa velikih in malih črk: zapis začetnih besed v vsakem stavku pretvorimo v najverjetnejši zapis z malo ali veliko črko in s tem zmanjšamo pomanjkanje podatkov

Korpus	Število segmentov
Učni	3.714.473
Razvojni	1.987
Testni	1.990
Skupaj	3.718.450

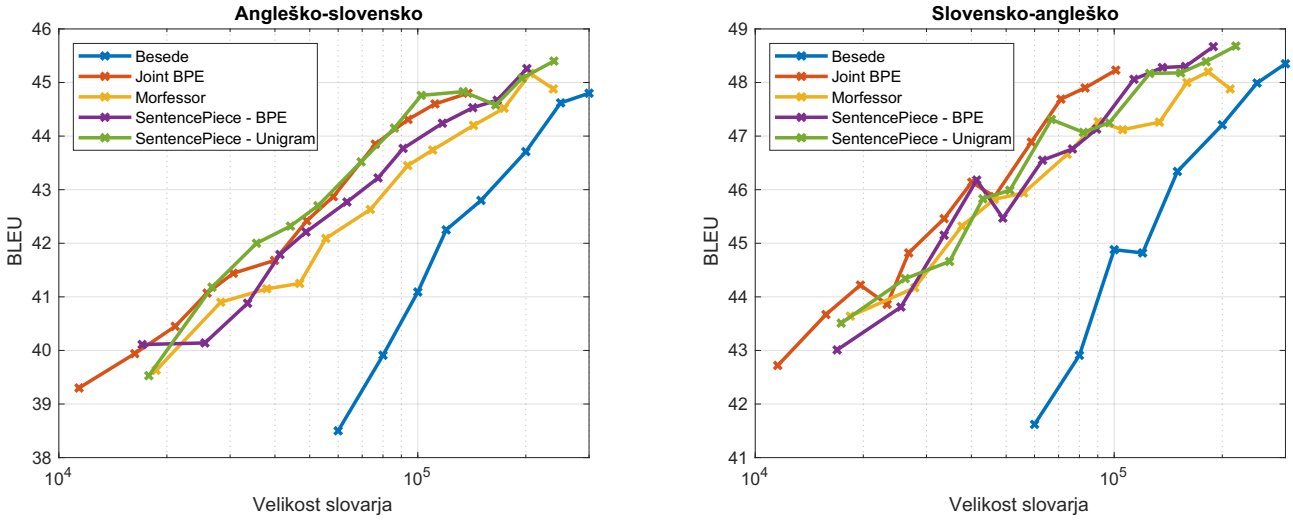
Tabela 2: Število segmentov besedila v učnem, razvojnem in testnem korpusu.

Ker smo izhajali iz želje po različnih velikostih končnih slovarjev, smo spreminjali ustrezne parametre pri uporabi orodij za učenje razcepljanja. Pri tem pa orodja uporabljajo te parametre na različne načine, kar pomeni, da velikosti končnih slovarjev ne ustrezajo natančno nastavljenim vrednostim parametrov. Želene vrednosti, ki smo jih nastavili, so: 10.000, 15.000, 20.000, 25.000, 30.000, 40.000, 50.000, 60.000, 80.000, 100.000, 120.000 in 150.000. V tabeli 2.4. so prikazane natančne velikosti slovarjev, ki jih dobimo na slovenski in na angleški učni množici pri teh nastavitvah.

Na sliki 1 je prikazan primer segmenta, kjer smo besede razcepili z uporabo vseh 4 postopkov. Uporabili smo ciljno velikost slovarja 20.000, saj so pri tej velikosti razcepljanja besed bolj pogosta in lažje prikažemo več razlik v enem segmentu. Na sliki so mesta delitve besed nakazana z vezaji.

V modelih brez razcepljanja besed smo uporabili velikosti slovarjev: 60.000, 80.000, 100.000, 125.000, 150.000, 200.000, 250.000 in 300.000.

V naslednjem koraku smo zgradili slovarje za vse različice razcepljenih učnih korpusov kot tudi za nerazce-



Slika 2: Rezultati uspešnosti prevajanja za vse modele.

pljen besedni učni korpus. Medtem ko v razcepljenih korpusih slovarji pokrijejo celotni korpus, se pri besednem korpusu pojavijo besede izven slovarja. V tabeli 3.2. smo prikazali deleže besed izven slovarja (OOV) na testnem delu korpusa za oba jezika. Po pričakovanjih vidimo, da so deleži večji na slovenski strani in da padajo z večanjem slovarja.

Slovar	OOV (en) [%]	OOV (sl) [%]
60k	2,57	6,66
80k	2,07	5,38
100k	1,77	4,44
125k	1,50	3,74
150k	1,30	3,22
200k	1,08	2,53
250k	0,95	2,11
300k	0,85	1,82

Tabela 3: Delež besed izven slovarja pri besednih slovarjih na angleškem (en) in slovenskem (sl) testnem korpusu.

3.3. Prevajalnik

Model prevajalnika je v vseh primerih nevronski strojni prevajalnik na podlagi arhitekture RNN z dimenzijo skritega stanja 1024 in dimenzijo vgrajenih vektorjev 512 (privzete nastavitve orodja Marian NMT). Naše dosedanje izkušnje na tej učni množici pa kažejo, da z uporabo arhitekture transformer in samo-pozornosti ne dosežemo bistvenih izboljšav. Dolžine segmentov besedila smo pri učenju omejili na 80 pojavnic (besed in ločil oz. pod-besednih enot in ločil), kar pomeni, da upoštevamo 99,7 % vseh segmentov v učni množici brez razcepljanja. Pri modelih, kjer uporabljamo razcepljanje pa tako upoštevamo med 96,3 % in 99,5 % vseh segmentov. Omejitve dolžine segmentov nismo več povečevali, saj glede na omenjeno pokritost predvidevamo, da ne bo več prišlo do znatnih sprememb rezultatov.

Učenje smo izvajali 10 epoh s preverjanjem rezultata na razvojni množici na vsakih 100 posodobitev parametrov modela. Najboljši model glede na razvojno množico smo nato uporabili pri vrednotenju rezultatov na testni množici.

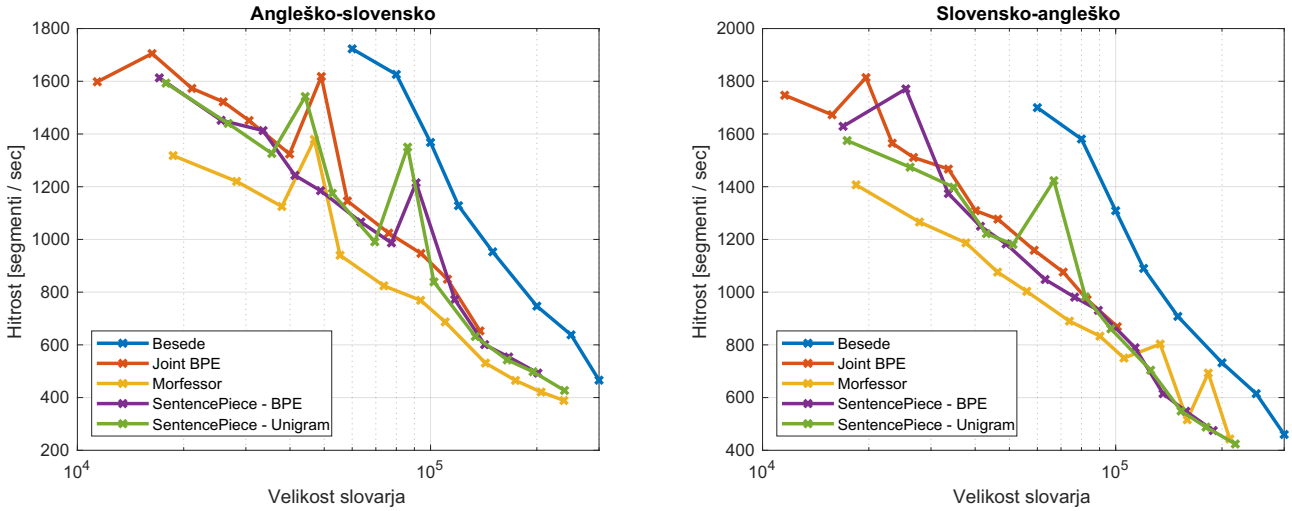
Pri prevajanju smo uporabljali mini serije (mini-batch) velikosti 64, medtem ko je pri učenju uporabljena fleksibilna velikost, ki je prilagojena velikosti delovnega pomnilnika enote GPU, na kateri izvajamo učenje.

3.4. Uporabljena orodja

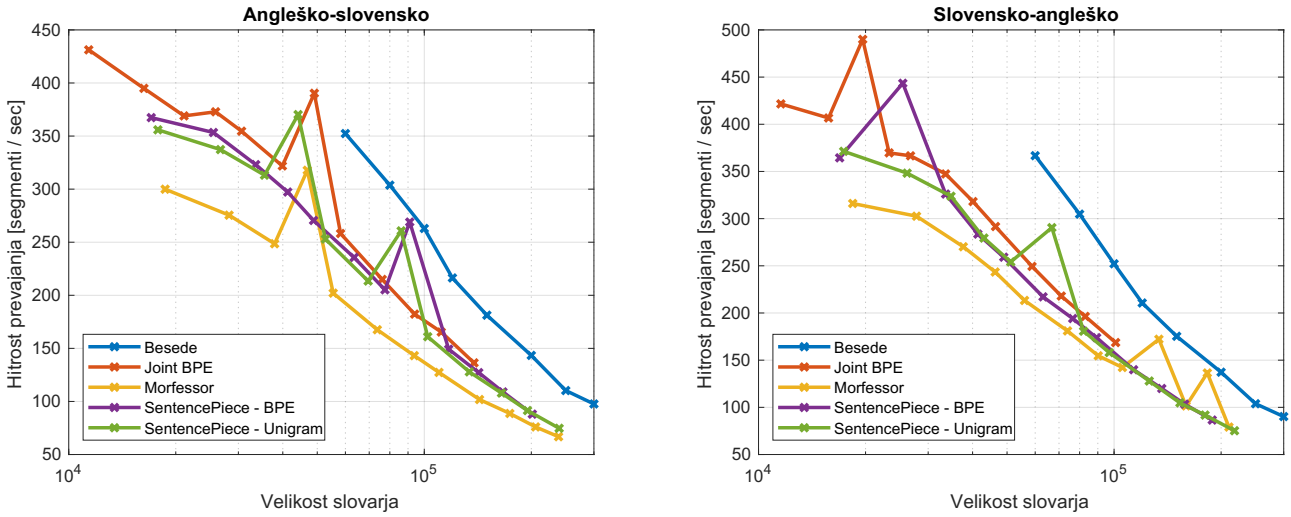
Za predprocesiranje (čiščenje, normalizacijo, tokenizacijo in truecaseing) ter postprocesiranje (detruecaseing in detokenizacijo) smo uporabljali skripte iz programskega paketa MOSES (Koehn et al., 2007). Za učenje prevajalnikov in prevajanje smo uporabljali orodje Marian NMT (Junczys-Dowmunt et al., 2018), ki smo ga poganjali na grafičnih procesnih enotah Nvidia Tesla V100. Za vrednotenje rezultatov z metriko BLEU smo uporabljali orodje SacreBLEU (Post, 2018), ki kot del vrednotenja izvaja tudi ponovno tokenizacijo in vrednoti tokenizirana besedila. Orodja za razcepljanje besed na pod-besedne enote so opisana v poglavju 2.

4. Rezultati in diskusija

Ker je bil osnovni namen uporabe pod-besednih enot zmanjšanje velikosti slovarja in s tem izvedljivost uporabe nevronskih strojnih prevajalnikov, najprej prikazujemo primer rezultatov na tipičnih velikostih slovarjev. Za besedni slovar smo izbrali velikost 60.000 besed, kar je pogosto uporabljena velikost slovarjev v procesiranju naravnega jezika. V tabeli 4. primerjamo rezultate prevajanja med besednim modelom in modelom Joint BPE z enako velikostjo slovarja. V tej točki lahko vidimo izboljšanje uspešnosti prevajanja z uporabo pod-besednih enot, kot jo tudi tipično zasledimo v obstoječi literaturi, npr. v (Sennrich et al., 2016). Na tej točki smo še dodali rezultate vrednotenja, ki jih dobimo z metriko ChrF ($\beta = 3$) (Popović, 2015). Čeprav se ta metrika uveljavlja za vrednotenje prevajanja pri morfološko kompleksnih jezik, smo preostale rezultate



Slika 3: Hitrost učenja prevajalnika za vse modele.



Slika 4: Hitrost uporabe prevajalnika za vse modele.

predstavili le z metriko BLEU, ki je še vedno uveljavljena in zadostuje za medsebojno primerjavo naših modelov.

Metrika	Besedni	Joint BPE
BLEU (en-sl)	38,50	42,87
BLEU (sl-en)	41,62	45,87
ChrF (en-sl)	58,43	63,13
ChrF (sl-en)	60,68	65,76

Tabela 4: Primer rezultatov uporabe besednega modela in modela z uporabo Joint BPE pri slovarju velikost 60.000.

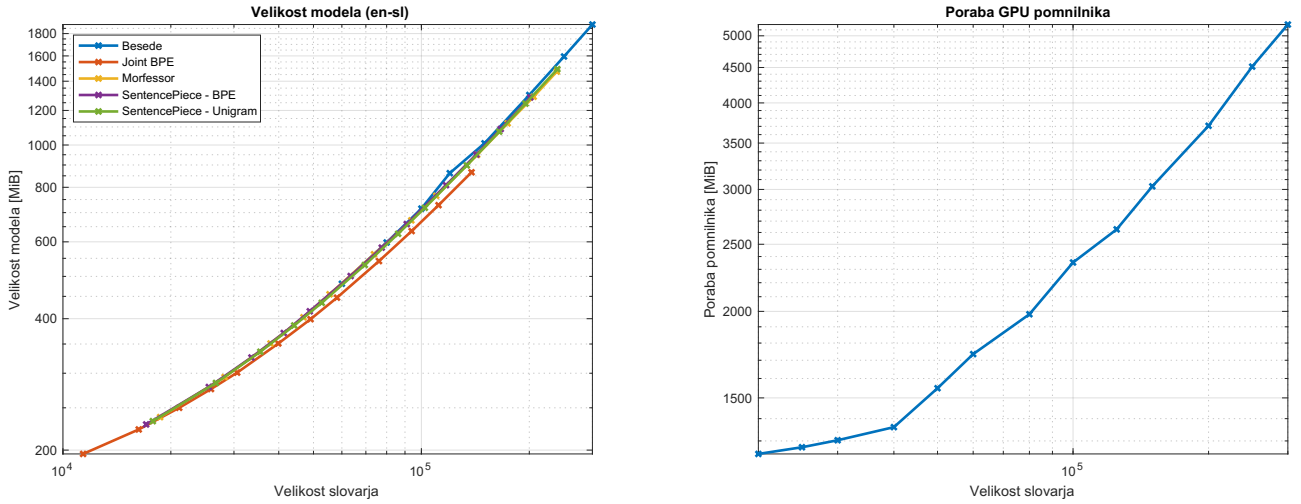
Na sliki 2 so prikazani rezultati uspešnosti prevajanja v odvisnosti od velikosti slovarja za vse sisteme. Na slikah so velikosti slovarjev ponazorjene v logaritemskem merilu.

Na splošno lahko opazimo, da uspešnost narašča z večanjem slovarjev, čeprav posamezni sistemi odstopajo od tega trenda, npr. prevajalnik iz slovenščine v angleščino s slovarjem 120.000 besed. Opazimo pa tudi, da uspešnost prevajanja pri uporabi besednih modelov naraščaja hitreje in se pri največjih slovarjih precej približa uspešnosti prevajalnikov z razcepljanjem.

Ko med sabo primerjamo sisteme, ki uporabljajo razcepljanje, vidimo manjše razlike. Kljub temu lahko opazimo, da pri prevajanju iz angleščine v slovenščino večinoma daje najboljše rezultate orodje Sentence Piece z razcepljanjem na podlagi unigramov (Sentence Piece – Unigram), v nasprotni smeri pa orodje Subword NMT s skupnim učenjem BPE (Joint BPE).

Slika 3 prikazuje hitrost učenja modela prevajalnika, slika 4 pa hitrost prevajanja pri njegovi uporabi. V vseh primerih uporabljamo kot merilo za hitrost število obdelanih segmentov besedila na sekundo, saj se zaradi različnih razcepljanj število pojavnic razlikuje med sistemi. Število besed na sekundo pri učenju dobimo, če upoštevamo, da je povprečno število besednih pojavnic (besed in ločil) na segment v angleškem besedilu je 18,7, v slovenskem besedilu pa 20,2.

Opazimo lahko, da se obe hitrosti zmanjšujeta z



Slika 5: Velikost izdelanega modela za vse modele ter poraba pomnilnika na grafični procesorski enoti.

večanjem slovarja. Hitrost pri besednih modelih je večja kot pa hitrost pri ostalih modelih, vendar se tudi tukaj razlika pri večjih slovarjih zmanjšuje. Vidimo pa sicer, da so najhitrejši modeli tisti, ki za razcepljanje korpusa uporabljajo orodje Morfessor. Lahko pa v teh modelih opazimo več točk, ki močno odstopajo od trendov. Predvidevamo, da so odstopanja nastala zaradi naključnih začetnih nastavitev nekaterih parametrov pri učenju, morebitnih odstopanj na uporabljeni strojni opremi, specifičnih lastnosti programske opreme za učenje modelov ali pa med prilagajanjem velikost mini serije pri različnih velikostih slovarja.

Prikazane hitrosti prevajanja ne upoštevajo predprocesiranja in postprocesiranja besedila.

Na sliki 5 so prikazane velikosti datotek za vse modele prevajanja in njihova poraba pomnilnika enote GPU za besedne modele. Velikosti datotek naraščajo skoraj linearno z velikostjo slovarja (na grafu sta obe osi v logaritemskem merilu). Vsakemu modelu sta pridruženi dve datoteki z obema slovarjema, ki pa sta bistveno manjši. Prikazane velikosti so za modele prevajanja iz angleščine v slovenščino. Modeli v nasprotni smeri imajo primerljive velikosti.

Desno je prikazana še poraba pomnilnika pri uporabi modelov pri prevajanju. Vidimo, da ima orodje osnovno porabo pomnilnika, kar se kaže v manjših spremembah porabe pri malih slovarjih. Pri večjih slovarjih pa poraba pomnilnika prav tako linearno narašča. Prikazana je poraba pomnilnika za besedne modele, ki je enaka v obeh smereh prevajanja. Porabe pomnilnika pri drugih modelih so primerljive glede na velikost slovarja. Pri preverjanju porabe pomnilnika je bila uporabljena mini serija velikosti 64. Pri učenju pa je poraba pomnilnika lahko tudi večja.

Pripomniti je potrebno, da bi bile velikosti drugačne v primeru drugih nastavitev hiperparametrov modelov.

5. Zaključek

V članku smo prikazali in primerjali nekatere najpogostejše podatkovno vodene metode za razcepljanje besed in njihovo uporabo na primeru nevronskega strojnega prevajalnika. Naši rezultati kažejo, da z razcepljanjem besed še vedno dosegamo boljše rezultate kot pa s prevajalniki brez razcepljanja, tudi v primerih, ko jim večamo slovarje. Trend pa kaže, da lahko z besednimi prevajalniki z nadaljnjim večanjem slovarja dohitimo modele z razcepljanjem besed. Ob trenutnem trendu razvoja in večanja pomnilniških zmogljivosti enot GPU, bo takšne modele v prihodnje možno naučiti in uporabljati.

Prikazani rezultati lahko služijo raziskovalcem in uporabnikom kot orientacija pri izbiri velikost slovarja za strojne prevajalnike, če želijo upoštevati uspešnost prevajanja, hitrost prevajanja in velikost modela. Slednja je lahko pomembna zaradi omejitev strojne opreme.

Za boljše razumevanje uporabnosti razcepljanja besed v strojnem prevajanju bi bilo potrebno izvesti še nadaljnje raziskave. V tem prispevku smo se omejili na stalne vrednosti hiperparametrov modelov. Izvedli smo le postopke podatkovno vodenega razcepljanja. V nadaljevanju lahko preučujemo tudi metode razcepljanja na podlagi slovničnega znanja ali pa kombiniranje komplementarnih metod. Pomemben prispevek h kakovosti prevajanja pri besednih modelih imata lahko tudi večanje učne množice in večanje hiperparametrov modela. Slednje pa sicer pomeni tudi povečanje velikosti modela in njegovo počasnejše delovanje. Nadaljnje raziskave lahko tudi vključujejo podrobnejšo analizo napak, ki se pojavljajo pri različnih metodah razcepljanja.

6. Zahvala

Raziskovalni program št. P2-0069, v okvirju katerega je nastala ta raziskava, je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Avtorji se zahvaljujejo konzorciju HPC RIVR (www.hpc-rivr.si) za sofinanciranje raziskave z uporabo zmogljivosti sistema HPC MAISTER na Univerzi v Mariboru (www.um.si).

Zahvaljujejo se tudi avtorjem vzporednega korpusa ParaCrawl za njegovo prosto dostopnost.

7. Literatura

Tamali Banerjee in Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level NMT. V: *Proceedings of the second workshop on subword/character level models*, str. 55–60.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins in Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. V: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, str. 4555–4567. Association for Computational Lin- guistics.

Mathias Creutz in Krista Lagus. 2002. Unsupervised discovery of morphemes. V: *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, str. 21–30, Philadelphia, Pennsylvania.

Rohit Gupta, Laurent Besacier, Marc Dymetman in Matthias Gallé. 2019. Character-based NMT with transformer. *arXiv:1911.04997*.

Georg Heigold, Stalin Varanasi, Günter Neumann in Josef van Genabith. 2018. How robust are character-based word embeddings in tagging and MT against wrod scramlbing or randdm nouse? V: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, str. 68–80, Boston, MA. Association for Machine Transla- tion in the Americas.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins in Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. V: *Proceedings of ACL 2018, System Demonstrations*, str. 116–121, Melbourne, Australia. Associa- tion for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin in Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. V: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, str. 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo in John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. V: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, str. 66–71, Brussels, Belgium. Association for Com- putational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. V: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, str. 66–75, Melbourne, Australia. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. V: *Proceedings of the Tenth Workshop on Statistical Machine Translation*, str. 392–395, Lisbon, Portugal. Association for Com- putational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. V: *Proceedings of the Third Conference on Machine Translation: Research Papers*, str. 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow in Alexandra Birch. 2016. Neural machine translation of rare words with subword units. V: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, str. 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Felix Stahlberg. 2020. Neural Machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos in Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline. Tehnično poročilo, Aalto University.

Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec[1], Kaja Dobrovoljc[3,1], Darja Fišer[4,3,1], Jan Jona Javoršek[1], Simon Krek[2,1], Taja Kuzman[1], Cyprian Laskowski[2], Nikola Ljubešić[1,2], Katja Meden[1]

[1] Institut »Jožef Stefan«
tomaz.erjavec@ijs.si, kaja.dobrovoljc@ijs.si, jan.javorsek@ijs.si, simon.krek@ijs.si,
taja.kuzman@ijs.si, nikola.ljubesic@ijs.si, katja.meden@ijs.si
[2] Center za jezikovne vire in tehnologije Univerze v Ljubljani
cyp@cjvt.si
[3] Filozofska fakulteta Univerze v Ljubljani
darja.fiser@ff.uni-lj.si
[4] Inštitut za novejšo zgodovino

Povzetek

Prispevek povzame storitve slovenske raziskovalne infrastrukture za jezikovne vire in tehnologije CLARIN.SI, ki je članica evropskega konzorcija raziskovalnih infrastruktur CLARIN ERIC. Najprej obravnavamo vodenje, organizacijo in tehnično infrastrukturo CLARIN.SI, nato pa njene spletne storitve, predvsem repozitorij digitalnih jezikovnih virov in orodij ter konkordančnike. Sledi pregled promocije področja jezikovnih tehnologij in digitalne humanistike v Sloveniji, kar vključuje storitve centra znanja za računalniško obdelavo južnoslovanskih jezikov CLASSLA, financiranje projektov in organizacijo, podporo ali sodelovanje na konferencah in delavnicah. Predstavimo tudi sodelovanje CLARIN.SI s CLARIN ERIC in s sorodnima slovenskima infrastrukturama DARIAH-SI in CESSDA/ADP ter vključenost v slovenske in evropske projekte.

The CLARIN.SI Research Infrastructure

The paper summarises the services offered by the Slovenian research infrastructure for language resources and technologies CLARIN.SI, which is a member of the European research infrastructure consortium CLARIN ERIC. We first present the governance, organisation and technical infrastructure of CLARIN.SI, followed by a description of its web applications with a focus on its repository and concordancers. Next comes an overview of support activities that CLARIN.SI offers to the fields of language technologies and digital humanities in Slovenia, which includes services of the knowledge centre for computational processing of South-Slavic languages CLASSLA, financial support of projects, and organisation or support of conferences and workshops. We also introduce the work of CLARIN.SI within CLARIN ERIC, its cooperation with its sister national infrastructures DARIAH-SI and CESSDA/ADP, and involvement in national and European projects.

1. Uvod

Raziskovalna infrastruktura (RI) CLARIN[1] (»Common Language Resources and Technology Infrastructure« oz. »Infrastruktura za skupne jezikovne vire in tehnologije«) zagotavlja digitalne jezikovne vire, orodja in storitve za podporo raziskovalcem na področju humanistike in družboslovja in drugih področij, ki se ukvarjajo z jezikom (Jong et al., 2018). CLARIN je bila ena od infrastruktur, ki so bile predvidene že v prvem načrtu Evropskega strateškega foruma za raziskovalne infrastrukture ESFRI (Váradi et al., 2008). Ustanovljena je bila leta 2012 in je bila ena prvih RI, ki je pridobila status evropske pravne osebe konzorcija raziskovalnih infrastruktur ERIC (European Research Infrastructure Consortium). CLARIN ERIC ima sedež na Nizozemskem in trenutno združuje RI 22 držav članic in 3 opazovalke. Zaposluje vodjo in podporno osebje za koordinacijo in centralne tehnične storitve, medtem ko imajo glavno vlogo pri zagotavljanju storitev nacionalni centri RI.

Glede na pomen slovenskega jezika za Slovenijo je sodelovanje v CLARIN ključnega pomena, saj spodbuja empirično podprto raziskovanje jezika ter razvoj jezikovnih virov in tehnologij, s čimer lahko slovenščina v informacijski družbi nastopa enakopravno z drugimi jeziki, tudi mnogo večjih skupnosti (Krek, 2022). Korist od RI imajo raziskovalci, učitelji in študenti slovenskega jezika ter drugih jezikoslovnih smeri, računalniškega jezikoslovja in umetne inteligence, pa tudi drugi raziskovalci s področja humanistike in družboslovja, ki pri svojem delu uporabljajo jezikovna gradiva. RI nudi podporo tudi slovaropiscem, prevajalcem in podjetjem, ki v svoje produkte vključujejo obdelavo slovenskega jezika, nenazadnje pa tudi laičnim uporabnikom za namene reševanja praktičnih vprašanj.

Slovenska RI CLARIN.SI je bila ustanovljena leta 2014, članica CLARIN ERIC pa je postala leta 2015, za kar je bilo potrebno, da je bil ustanovljen nacionalni konzorcij in da je Vlada Republike Slovenije podpisala memorandum, s katerim se je zavezala plačevati članarino za članstvo Slovenije v CLARIN ERIC. Do sedaj je bila edina publikacija, ki celostno predstavi CLARIN.SI, objavljena kmalu po njeni ustanovitvi (Erjavec et al., 2014), kjer smo predstavili prve korake RI in načrte za nadaljnje delo. Pričujoči prispevek povzema, kaj je bilo narejenega v minulih osmih letih: v razdelku 2. predstavimo organizacijsko strukturo in upravljanje infrastrukture, v 3. repozitorij jezikovnih virov in orodij, v 4. spletne storitve, v 5. podporne dejavnosti, v 6. vpetost CLARIN.SI v domače in evropske projekte in v aktivnosti CLARIN ERIC, v 7. pa podamo zaključke in načrte za nadaljnje delo.

[1] https://www.clarin.eu/

2. Organiziranost CLARIN.SI

Infrastruktura ima sedež na Institutu »Jožef Stefan« (IJS), kjer tudi domuje večina njene računalniške opreme in kjer se zagotavlja varnost, vzdrževanje in neprestano obratovanje spletnih storitev RI. Pri vodenju in tehničnem vzdrževanju sodelujejo tri organizacijske enote IJS, in sicer Odsek za tehnologije znanja E8, Laboratorij za umetno inteligenco E3 ter Center za mrežno infrastrukturo CMI.

CLARIN.SI je organiziran kot konzorcij, ki nima narave pravne osebe, v njem pa ima članstvo 12 partnerjev.V konzorciju so združene vse glavne institucije, ki se v Sloveniji ukvarjajo z razvojem ali uporabo jezikovnih virov in tehnologij, in sicer:

- Univerze: Univerza v Ljubljani, Univerza v Mariboru, Univerza v Novi Gorici in Univerza na Primorskem. Univerza v Ljubljani je sedež Centra za jezikovne vire in tehnologije (CJVT), ki koordinira delo na področju korpusnega jezikoslovja in jezikovnih tehnologij ter razvija in vzdržuje temeljne digitalne jezikovne vire in jezikovnotehnološka orodja za sodobni slovenski jezik.

- Raziskovalni inštituti: ZRC SAZU, Institut »Jožef Stefan« (IJS), Inštitut za novejšo zgodovino (INZ) in Znanstveno-raziskovalno središče Koper. Znotraj ZRC SAZU Inštitut za slovenski jezik Frana Ramovša zbira jezikovno gradivo in ga uporablja za izdelavo temeljnih del slovenskega jezikoslovja, predvsem slovarjev. IJS kot gostitelj raziskovalne infrastrukture CLARIN.SI koordinira delo infrastrukture, vzdržuje in nadgrajuje njen repozitorij in storitve ter razvija jezikovne vire in orodja.

- Društva oz. zavodi: Slovensko društvo za jezikovne tehnologije (SDJT), ki s konferenco »Jezikovne tehnologije in digitalna humanistika« (JTDH) promovira razvoj jezikovnih tehnologij za slovenski jezik, in Zavod za uporabno slovenistiko Trojina s svetovalno in podporno dejavnostjo ter izdelavo jezikovnih virov in orodij.

- Podjetji Alpineon in Amebis, med katerima prvo podjetje v infrastrukturo CLARIN.SI prispeva predvsem govorne tehnologije, drugo pa se ukvarja z izdelavo programske opreme s področja jezikovnih tehnologij in elektronskega založništva.

Odločitve o vodenju RI sprejema oz. potrjuje Upravni odbor (UO) CLARIN.SI, v katerem ima vsak partner po enega predstavnika in poljubno število namestnikov oz. namestnic. Komunikacija se odvija prek dopisnega seznama upravnega odbora, ki trenutno šteje 34 članov, enkrat letno pa organiziramo sestanek CLARIN.SI UO, na katerem se pogovorimo o delovanju RI v preteklem letu in naredimo načrte za naslednje.

Delovanje raziskovalne infrastrukture CLARIN v Sloveniji se tako kroji na podlagi potreb in konsenza vseh večjih akterjev na področju digitalnega jezikoslovja in jezikovnih tehnologij, kot tudi digitalne humanistike in družboslovja, saj CLARIN.SI tesno sodeluje z dvema sestrskima RI v slovenskem prostoru. To sta DARIAH-SI s sedežem na Inštitutu za novejšo zgodovino (INZ), ki predstavlja nacionalno vozlišče evropske RI za digitalno humanistiko, in CESSDA/ADP v Arhivu družboslovnih podatkov na Fakulteti za družbene vede Univerze v Ljubljani (ADP), ki je nacionalno vozlišče evropske RI za digitalno družboslovje CESSDA (»Consortium of European Social Science Data Archives«). CLARIN.SI je tudi ena od ustanovnih članic Slovenskega nacionalnega superračunalniškega omrežja SLING[2] in preko njega članica federacije računskih in podatkovnih virov EGI[3] ter Partnerstva za napredno računalništvo v Evropi PRACE[4].

CLARIN.SI vzdržuje dvojezično (slovenščina, angleščina) spletno stran,[5] na kateri je predstavljena RI kot tudi vse njene storitve. Spletno mesto nudi tudi kontaktne podatke, npr. e-poštni naslov, na katerega se lahko obrnejo uporabniki, ki si želijo pomoči ali nasvetov. Poleg tega spletno mesto vključuje z geslom zaščitene interne strani, do katerih imajo dostop člani oz. namestniki UO in ki vsebujejo ustanovne dokumente, zapisnike sestankov, relevantne zapisnike CLARIN ERIC itd.

Za dokumentiranje tehničnega vzdrževanja CLARIN.SI uporablja interno instalacijo platforme WordPress, na kateri dokumentiramo postopke vzdrževanja za vse spletne storitve CLARIN.SI, medtem ko se za zahtevke za reševanje odkritih problemov uporablja instalacijo platform Redmine.

Kritične spletne storitve CLARIN.SI so vedno instalirane tudi na razvojnem strežniku, kjer se najprej preveri delovanje vsake spremembe na programski opremi, na ponujenih jezikovnih virih ali v dokumentaciji. Delovanje spletnih storitev se preverja prek sistema NAGIOS, repozitorij pa tudi neodvisno s strani CLARIN ERIC. V primeru napak so tako skrbniki storitve nemudoma obveščeni in lahko takoj pristopijo k odpravljanju težave.

3. Repozitorij jezikovnih virov

Osnovna storitev CLARIN.SI je vzdrževanje repozitorija jezikovnih raziskovalnih podatkov oz. jezikovnih virov, kot so velike in bogato označene zbirke besedil (korpusi), računalniški leksikoni in modeli, pa tudi strojno berljivi slovarji in računalniška orodja. Računalniška platforma repozitorija je odprtodostopna CLARIN-DSpace,[6] ki so jo razvili posebej za namene CLARIN repozitorijev v okviru češke raziskovalne infrastrukture CLARIN (sedaj CLARIAH, ki je nastala po združitvi češke CLARIN in DARIAH) na Inštitutu za formalno in uporabno jezikoslovje na Karlovi Univerzi v Pragi. Platformo poleg Slovenije, in

[2] https://www.sling.si/
[3] https://www.egi.eu/
[4] https://prace-ri.eu/
[5] https://www.clarin.si/
[6] https://github.com/ufal/clarin-dspace

seveda Češke, uporablja še sedem drugih nacionalnih repozitorijev CLARIN, kar skupaj predstavlja 40 % vseh rednih članic CLARIN ERIC.

Repozitorij CLARIN.SI je poleg ADP edini v Sloveniji akreditiran s certifikatom »Core Trust Seal«,[7] torej kot zaupanja vreden podatkovni repozitorij. Repozitorij v skladu s strategijo CLARIN ERIC implementira načela FAIR[8,9](najdljivost, dostopnost, interoperabilnost in ponovna uporaba). Evropski agendi za odprto znanost in načelom FAIR CLARIN sledi avant la lettre (Jong et al., 2018), in sicer z naslednjimi instrumenti:

- Akademska avtentikacija AAI, ki deluje po sistemu SSO (»Single sign-on«), kjer ločimo ponudnike identitete (Arnes, univerze, druge akademske institucije) in ponudnike storitev (v našem primeru repozitorij), da uporabnikom ni potrebno ustvariti svojega računa na CLARIN.SI, pač pa se v repozitorij prijavijo prek svojega EduGain uporabniškega imena in gesla pri izbranem ponudniku identitete.
- Trajni identifikatorji vnosov po sistemu »handle«, kar omogoča pripis trajnega naslova URL vsakemu vnosu v repozitorij, ki je, enako kot DOI, neodvisen od specifičnega URL-ja tega vira v okviru repozitorija, in s tem tudi odporen na spremembe v platformi oz. lokaciji repozitorija.
- Vpetost v mednarodne spletne agregatorje metapodatkov, kot so OpenAIRE[10], Re3data[11], od 2022 pa tudi European Language Grid. Preko CLARIN ERIC je bil CLARIN.SI med prvimi RI vključen tudi v sistem ponudbe virov in storitev v okviru Evropskega odprtega znanstvenega oblaka EOSC[12] že vse od vzpostavitve portala EOSC leta 2018. V okviru RI CLARIN se za metapodatkovne zapise uporablja priporočila CMDI[13] (»Component MetaData Infrastructure«), izvoz oz. žetev metapodatkov pa je omogočena tudi v standardu Dublin Core.
- Bogata izbira licenc, od odprtih, kot so licence Creative Commons, do bolj omejenih, ki zahtevajo predhodno prijavo v repozitorij in digitalni podpis sporazuma o uporabi vira.
- Eksplicitni pogoji uporabe, ki določajo pravice in dolžnosti tako upravljalcev repozitorija kot uporabnikov.
- Navodila za deponiranje vnosov, ki opišejo postopek oddaje virov s posebnim poudarkom na zahtevanih metapodatkih in njihovi obliki, saj se pri CLARIN.SI trudimo vzdrževati čim bolj popolne in enotne metapodatkovne zapise.
- Navodila za kodiranje deponiranih podatkov, ki navajajo sprejemljive formate zapisa in načine označevanja podatkov, poleg tega pa zajemajo tudi splošna navodila za pripravo kvalitetnih in usklajenih podatkov. Po tem se repozitorij CLARIN.SI razlikuje od večine drugih repozitorijev CLARIN (Lenardič in Fišer, 2022), saj ti tipično ponujajo samo seznam sprejemljivih formatov, ne pa tudi bolj splošnih navodil za pripravo kakovostnih podatkov, kakršna so lahko zelo koristna za avtorje s področja humanističnih znanosti brez poglobljenega znanja računalniških veščin za pravilno pripravo podatkov.
- Seznam pogosto postavljenih vprašanj z odgovori in podobne vsebine.

Poleg prilagojenosti za opis jezikovnih virov je za razliko od splošnih repozitorijev za samoarhiviranje, kot je npr. Zenodo, pomembna odlika repozitorija CLARIN.SI zagotavljanje visoke kvalitete deponiranih jezikovnih virov in njihovih metapodatkov, saj vsak vnos pred objavo skrbno pregleda eden od urednikov repozitorija, ki preveri, ali vnos ustreza merilom CLARIN.SI. Če jim ne, urednik vnos zavrne z obrazložitvijo napak, v vnaprej dogovorjenih primerih pa tudi pomaga pri popravljanju vira.

V osmih letih, kolikor jih je minilo od prvega vnosa, je število deponiranih jezikovnih virov in orodij naraslo na več kot 300, ki so rezultat dela prek 700 avtorjev, pri čemer je v mnoge bilo vloženih več let dela. V letu 2021 je repozitorij beležil okoli 40.000 ogledov in 4.000 prenosov. V tem letu so bili najpogosteje preneseni viri zbirka 751 emodžijev z avtomatsko pripisanim sentimentom, ki je bil izračunan na podlagi 70.000 tvitov v 13 evropskih jezikih, označenih za sentiment in s strani 83 anotatorjev (Kralj Novak et al., 2015) ter jezikovni modeli (besedne vložitve) tipa BERT (Devlin et al., 2018) za slovenske besede (Ulčar in Robnik-Šikonja, 2021), ki so koristni za marsikatero nalogo obdelave slovenskega jezika.

S spodbujanjem deponiranja jezikovnih virov in pomočjo pri njihovem oblikovanju in opisu je CLARIN.SI bistveno pripomogel k uveljavljanju koncepta odprte, preverljive, ponovljive in odgovorne znanosti na področju jezikoslovnih raziskav v Sloveniji ter številne jezikovne vire, nastale kot rezultat slovenskih raziskovalnih projektov, obvaroval pred izginotjem in jim omogočil mednarodno vidnost in vpliv.

4. Spletne storitve

Poleg repozitorija CLARIN.SI trajno vzdržuje več spletnih storitev, od katerih so najpomembnejši konkordančniki, tj. orodja za analizo korpusov, in sicer ponuja CLARIN.SI uporabo konkordančnika KonText in dveh različic konkordančnika noSketch Engine (Crystal in Bonito). Vsi trije uporabljajo isti zaledni program, in sicer Manatee (Rychlý, 2007), ki omogoča

[7] https://www.coretrustseal.org/
[8] https://www.go-fair.org/fair-principles/
[9] https://www.clarin.eu/fair
[10] https://www.openaire.eu/
[11] https://www.re3data.org/
[12] https://eosc-portal.eu/
[13] https://www.clarin.eu/content/component-metadata

hitre poizvedbe po bogato označenih korpusih, vendar se razlikujejo v čelnem delu. NoSketch Engine je odprtokodna različica komercialnega konkordančnika Sketch Engine (Kilgarriff et al., 2014),[14] medtem ko je bil KonText razvit na oddelku Češkega nacionalnega korpusa Karlove univerze v Pragi (Machálek, 2020). Poleg izgleda konkordančnikov so glavne razlike med njimi v tem, da noSketch Engine ponuja nekaj več funkcionalnosti kot KonText (predvsem možnost izračuna ključnih besed korpusa oz. podkorpusa), medtem ko KonText podpira prijavo prek sistema AAI (enako kot repozitorij), kar nato omogoča personalizirane nastavitve zaslona, hranjenje zgodovine poizvedb, itd.

Vsi konkordančniki na CLARIN.SI ponujajo isti nabor korpusov, ki jih je sedaj že preko 40, od referenčnih do specializiranih, pa tudi govorjenih in večjezičnih. Tu izpostavimo novi korpus metaFida, ki združuje 34 obstoječih korpusov in vsebuje skupaj 4,5 milijarde pojavnic, s čimer je največji in najbolj raznovrsten korpus za slovenščino, po katerem je mogoče iskati s pomočjo konkordančnikov.

Konkordančniki CLARIN.SI se uporabljajo pri izvajanju študijskih programov na več univerzah, v sklopu jezikoslovnih raziskav ali pri različnih raziskovalnih projektih, kot tudi v prevajalskih podjetjih.

Naslednja spletna storitev, ki jo ponuja CLARIN.SI, je platforma za ročno označevanje korpusov WebAnno (Yimam et al., 2013), ki so jo razvili v okviru CLARIN-DE. V okviru CLARIN.SI smo razvili pretvorbo iz zapisa korpusov TEI v format TSV3, ki ga uporablja WebAnno, in združevanje izvornega korpusa TEI z ročnimi oznakami iz datoteke TSV, s čimer smo omogočili dodajanje oz. spreminjanje obstoječih oznak v TEI kodiranih korpusih z oznakami, ki so bile ročno vstavljene oz. popravljene na platformi WebAnno (Erjavec et al., 2016)[15]. Naša instalacija in pretvorba je bila do sedaj uporabljena pri prek 10 projektih, npr. za ročno označevanje normaliziranih besednih oblik, lem in oblikoslovnih oznak uporabniško generiranih vsebin v okviru projekta Janes »Jezikoslovna analiza nestandardne slovenščine« (Fišer et al., 2020),[16] za označevanje dvojezičnih terminov v okviru projekta KAS »Slovenska znanstvena besedila: viri in opis« (Erjavec et al., 2021)[17] ali za označevanje definicij terminov v besedilih v okviru projekta TermFrame »Terminologija in sheme znanja v medjezikovnem prostoru« (Vintar in Martinc, 2022).

Za kontrolirano in kolaborativno vzdrževanje je postala zelo popularna platforma Git, ki jo v okviru CLARIN.SI prav tako uporabljamo, ne samo za programsko opremo, temveč tudi za jezikovne vire. Za spletno dostopne repozitorije Git, ki vključujejo tudi množico drugih funkcij, kot so zahtevki in izvajanje programov, sta najbolj uporabljana GitHub in GitLab. Na GitHubu ima CLARIN.SI svojo virtualno organizacijo,[18] ki združuje sedaj že okoli 60 odprtokodnih projektov. Za razliko od GitHuba, ki obstaja samo kot spletna storitev v lasti podjetja Microsoft, je mogoče platformo GitLab tudi instalirati, kar ima to prednost, da so projekti locirani na lokalni računalniški opremi, dostopnost projektov pa je mogoče tudi omejiti, kar je v posameznih primerih potrebno, npr. zaradi avtorskih pravic nad besedili nekega jezikovnega vira, ki se ga razvija. Instalacija GitLab na CLARIN.SI[19] vsebuje okoli 20 projektov, tako javnih (kot npr. že omenjena pretvorba TEI za WebAnno) kot tudi zasebnih.

CLARIN.SI v okviru centra znanja CLASSLA, ki ga obravnavamo v naslednjem razdelku, ponuja tudi spletno storitev ReLDIanno za jezikoslovno označevanje besedil v slovenskem, hrvaškem in srbskem jeziku.[20] Storitev podpira oblikoskladenjsko označevanje, lematizacijo, označevanje imenskih entitet in skladenjsko razčlenjevanje, dostopna pa je tako prek spletnega vmesnika kot prek vtičnika API, pri čemer lahko rezultate prikaže na zaslonu ali pa označeno besedilo prenesemo na lastni računalnik.

5. Strokovna podpora in diseminacija

5.1. Središča znanja

CLARIN.SI je aktiven pri promociji in spodbujanju razvoja računalniškega jezikoslovja, ne le za slovenščino, ampak tudi za druge južnoslovanske jezike, kot so hrvaščina, srbščina, makedonščina in bolgarščina, s čimer si je RI bistveno povečala mednarodno odmevnost. CLARIN.SI namreč skupaj z bolgarsko raziskovalno infrastrukturo CLARIN CLADA-BG in hrvaškim Institutom za hrvaški jezik in jezikoslovje upravlja središče znanja CLARIN za južnoslovanske jezike CLASSLA, v okviru katerega ponuja strokovno podporo pri uporabi jezikovnih virov in tehnologij za južnoslovanske jezike. Središče znanja podpira raziskovalce z dokumentacijo o prosto dostopnih jezikovnih virih, orodjih za ustvarjanje in obdelavo besedilnih korpusov ter drugih jezikovnih tehnologijah. Poleg tega CLASSLA razvija lastne jezikovne tehnologije in korpuse, s katerimi pokriva velike potrebe južnoslovanskih jezikov, ki so tehnološko manj podprti. Tako je na primer v letu 2020 v sklopu projekta zbiranja korpusov besedil iz Wikipedije središče ustvarilo prvi jezikoslovno označeni makedonski korpus, CLASSLAWikimk (Ljubešić et al., 2021).

V 2021 je CLARIN.SI postal tudi član CLARIN centra znanja za obdelavo uporabniško generiranih vsebin CKCMC,[21] ki ga vodi Eurac Research, Bolzano.

5.2. Financiranje projektov

CLARIN.SI finančno podpira projekte, letno izbrane na odprtem razpisu za člane konzorcija, ki pripomorejo k uresničitvi strategije CLARIN.SI. Ta de-

[14] https://www.sketchengine.eu/
[15] https://gitlab.clarin.si/clarinsi/webanno_tei
[16] https://nl.ijs.si/janes/
[17] https://nl.ijs.si/kas/
[18] https://github.com/clarinsi
[19] https://gitlab.clarin.si/
[20] http://clarin.si/services/web/
[21] https://cmc-corpora.org/ckcmc/

javnost je bila zelo odmevna in je tudi pomembno doprinesla k zanimanju za raziskave in razvoj jezikovnih virov med mladimi. Od leta 2018, ko smo z iniciativo začeli, je bilo uspešno izvedenih 19 projektov, v sklopu katerih so med drugim nastali korpus parlamentarnih razprav Državnega zbora Republike Slovenije siParl (Pančur et al., 2020), nadgradnja korpusa akademske slovenščine KAS 2.0 (Žagar et al., 2022) in govornega korpusa Gos Videolectures (Verdonik et al., 2019), orodje za učinkovito analizo slovenskih korpusov LIST (Krsnik et al., 2019) in drugi jezikovni viri in programska oprema. Med drugim je CLARIN.SI financiral tudi projekt »Razvoj učnega gradiva na korpusu siParl 2.0: Korpusni pristop k raziskovanju parlamentarnega diskurza« (Fišer in de Maiti, 2021).

5.3. Organizacija dogodkov

CLARIN.SI sodeluje pri organizaciji in izvedbi dogodkov s področja računalniškega jezikoslovja in sorodnih tematik v Sloveniji, npr. »XVIII EURALEX Intl. Congress« (Ljubljana, 2018) ali »22nd Intl. Conf. on Text, Speech and Dialogue« (Ljubljana, 2019), predvsem pa glavne konference za to področje v Sloveniji, »Jezikovne tehnologije in digitalna humanistika«, ki ima prek 20-letno tradicijo in z organizacijo katere je začelo društvo SDJT. SDJT od leta 2005 organizira občasna predavanja JOTA (Jezikovnotehnološki abonma), kjer je CLARIN.SI podprl snemanje in arhiviranje 12 predavanj na VideoLectures.NET[22], do sedaj z 10.000 ogledi.

5.4. Obveščanje in promocija

Nenazadnje, delovanje CLARIN.SI in njegovih središč znanja redno predstavljamo na domačih in tujih delavnicah in konferencah, kot so konferenca Evropskega strateškega foruma za raziskovalne infrastrukture (ESFRI), konference CLARIN idr., ter na predavanjih v okviru študijskih programov slovenskih univerz.

CLARIN.SI organizira tudi delavnice o uporabi korpusov in jezikovnih tehnologij za namene znanstvenih raziskav. Tako smo npr. izvedli delavnice[23] za uporabo konkordančnika noSketch Engine, platform WebAnno in Git, središče znanja CLASSLA pa je sodelovalo pri delavnici o uporabi korpusov za analizo regionalne variacije spolne zaznamovanosti jezika[24].

O dejavnostih partnerjev konzorcija CLARIN.SI in njegovih središč znanja javnost obveščamo tudi prek ažurnih novic, objavljenih na spletni strani infrastrukture, poštnega seznama ter objav s profila CLARIN.SI na Twitterju. Delo CLARIN.SI in njegovega središča znanja CLASSLA je bilo izpostavljeno tudi v več publikacijah »CLARIN ERIC Tour de CLARIN« (Fišer et al., 2019).

[22] https://videolectures.net/jota/
[23] https://www.clarin.si/info/dogodki/
[24] https://www.clarin.si/info/k-center/delavnice/

6. Vpetost v projekte in infrastrukture

CLARIN.SI je vpet v domače in evropske projekte, s čimer zagotavlja večjo izkoriščenost in vidljivost ter seveda tudi dodaten dotok sredstev za svoje delovanje.

6.1. Sredstva Evropske kohezijske politike

V okviru projekta kohezijskih sredstev MIZŠ 2018–2021 so partnerji konzorcija IJS, UM in UL nadgradili strojno opremo, s čimer je omogočeno hitrejše in proti okvaram odpornejše delovanje spletnih storitev CLARIN.SI, pridobljena gruča GPU strežnikov na Univerzi v Mariboru pa služi za raziskave globokega strojnega učenja obdelave jezikovnih podatkov, npr. na področju govora. S temi nadgradnjami lahko CLARIN.SI slovenski raziskovalni skupnosti zagotavlja odlično raziskovalno infrastrukturo, ki mdr. pripomore k privlačnosti slovenskih partnerjev v mednarodnih raziskovalnih in inovacijskih projektih ter podpira doseganje znanstvene odličnosti in mednarodno vrhunskih rezultatov. Tako npr. projekt EU MaCoCuuporablja gručo računalnikov CLARIN.SI za zajem in obdelavo spletnih velepodatkov, v okviru projekta EU InTaviapa se jezikoslovno označuje Slovenski biografski leksikon z modeli, razvitimi na gruči GPU. Več velikih projektov EU, kot sta ELEXIS in EMBEDDIA, je deponiralo razvite jezikovne vire v repozitorij CLARIN.SI.

6.2. Vpetost v evropske projekte

Med evropskimi projekti posebej izpostavimo ELEXIS,[25] saj je bila za potrebe tega projekta v repozitoriju CLARIN.SI narejena nova zbirka CLARIN.SI ELEXIS, v kateri so zbrani metapodatki in povezave do spletnih vmesnikov 143 digitalnih slovarjev. Ob koncu projekta ELEXIS v okviru CLARIN.SI oz. IJS načrtujemo tudi vzpostavitev novega centra znanja CLARIN za digitalno leksikografijo.

6.3. Vpetost v domače projekte

Sodelujemo tudi v več domačih projektih. Največji je »Razvoj slovenščine v digitalnem okolju«[26], ki mu CLARIN.SI zagotavlja svoje storitve za pregled in deponiranje v projektu izdelanih jezikovnih virov ter definicijo shem za usklajeno označevanje jezikovnih virov slovenskega jezika. V načrtu je tudi izdelava seznamov kontroliranih besedišč za jezikoslovno označevanje slovenskih besedil na ravni oblikoskladnje, skladnje, imenskih entitet, udeleženskih vlog itd.

6.4. Sodelovanje z drugimi RI

CLARIN.SI sodeluje s slovenskimi centri sestrskih infrastruktur CESSDA/ADP in DARIAH-SI. V projektu »RDA Node Slovenia« (2019–2020), ki ga je koordiniral ADP (FDV UL), smo pregledali in analizirali slovenske repozitorije raziskovalnih podatkov (Meden in Erjavec, 2021). Z INZ oz. DARIAH-SI pa smo sodelovali na področju standardizacije zapisa in izdelave korpusov parlamentarnih podatkov.

[25] https://elex.is/
[26] https://www.cjvt.si/rsdo/

6.5. Sodelovanje v delu CLARIN ERIC

CLARIN.SI je ena od aktivnejših nacionalnih RI v CLARIN ERIC. Pridobili smo sredstva za dva manjša projekta, ki sta vključevala mednarodni delavnici, in sicer 2016 v Ljubljani in 2019 v Amersfoortu. Slednja je bila v sodelovanju z DARIAH-SI posvečena izdelavi priporočil za standardizirano kodiranje korpusov parlamentarnih razprav z imenom Parla-CLARIN[27] (Erjavec in Pančur, V tisku), ki je postala priljubljena izbira za kodiranje parlamentarnih korpusov. Na tej osnovi je CLARIN.SI pridobil ključno vlogo v dveh večjih »CLARIN Flagship« projektih, ParlaMint I (2020–2021) in ParlaMint II (2022–2023).

Namen projektov ParlaMint je ustvariti primerljive, interpretativne in enotno kodirane korpuse parlamentarnih razprav. V že zaključenem projektu ParlaMint I je CLARIN.SI vodil zbiranje in kodiranje 17 korpusov nacionalnih parlamentov (Erjavec et al., 2022), ki so odprto dostopni na repozitoriju CLARIN.SI, kot tudi na konkordančnikih RI. V okviru projekta ParlaMint II, katerega namen je razširitev in obogatitev obstoječih korpusov ter dodajanje korpusov novih partnerjev, prav tako pa tudi razvoj izobraževalnih gradiv in primerov dobrih praks uporabe parlamentarnih korpusov za raziskave v humanistiki in družboslovju, člani CLARIN.SI vodijo štiri izmed petih delovnih sklopov[28].

Člani UO CLARIN.SI sodelujejo v delu CLARIN delovnih teles za pravna vprašanja (Mateja Jemec Tomazin, ZRC SAZU), za standardizacijo (Tomaž Erjavec, IJS) in za uporabniška vprašanja (Jakob Lenarčič, FF UL) ter na letnih konferencah CLARIN (T. Erjavec je predsednik programskega odbora konference v 2022 v Pragi). J. Lenarčič je prejel CLARIN »Stewen Krawer award« za mladega raziskovalca leta 2019, mdr. za svoje delo (skupaj z Darjo Fišer) pri vzpostavitvi iniciative »CLARIN Resource Families«[29], T. Erjavec pa je prejel »Steven Krauwer Award for CLARIN Achievements 2021« za svoje delo na projektu ParlaMint. Darja Fišer in Kristina Pahor de Maiti (FF UL) sta leta 2021 prejeli nagrado »Teaching with CLARIN Award« za najboljši učni material, povezan z uporabo virov CLARIN. Kaja Dobrovoljc (FF UL) je predstavila RI CLARIN na konferenci ob 20. obletnici ESFRI v Parizu leta 2022[30]. Darja Fišer je bila med leti 2016 in 2020 direktorica področja za uporabniška vprašanja, z letom 2023 pa naj bi postala generalna direktorica CLARIN ERIC.

7. Zaključki

CLARIN.SI je izjemno uspešno vzpostavljena infrastruktura, ki pokriva široko interdisciplinarno področje od humanističnih in družboslovnih raziskav do razvoja sistemov in tehnologij znanja in umetne inteligence. Podpira temeljne in aplikativne raziskave ter razvoj aplikacij, informacijskih sistemov in orodij na vseh ravneh tehnološke pripravljenosti.

Z izjemnim nacionalnim in regionalnim pomenom, spodbujanjem področja, pritegovanjem mladih, povezovanjem z industrijo ter široko vključenostjo deležnikov, močno vlogo pri uvajanju načel odprte znanosti ter izjemno odmevno in uspešno ter tudi nagrajeno vlogo na ravni evropskega in mednarodnega sodelovanja CLARIN.SI s povezanimi projekti predstavlja zgled za vzpostavitev uspešne vrhunske sodobne interdisciplinarne znanstveno-raziskovalno tehnološke infrastrukture.

V naslednjem obdobju bo CLARIN.SI poleg vzdrževanja obstoječih storitev še bolj intenzivno spodbujal ponovno uporabo raziskovalnih podatkov, s čimer bo omogočal raziskovalcem na področju humanistike in družboslovja povečanje produktivnosti, in kar je še pomembneje, vzpostavljanje novih raziskovalnih smeri, ki obravnavajo eno ali več družbenih vlog jezika. Drug pomemben cilj je izvajanje smernic za zagotavljanje interoperabilnosti CLARIN ERIC[31], ki je ključni predpogoj za učinkovito podporo raziskovalnemu delu skozi interoperabilnost orodij, virov, metapodatkov, standardov za zapis, kot tudi na organizacijski ravni (Jong et al., 2020). Hkrati bo treba okrepiti podporo uporabnikom, saj univerze in agencije od raziskovalcev v doktorskih in raziskovalnih programih vse intenzivneje zahtevajo načrte za ravnanje z raziskovalnimi podatki in njihovo trajno hrambo.

ESFRI kažipot 2021[32] za RI poudarja pomen podatkov FAIR, pri čemer smo na tem področju v okviru repozitorija CLARIN.SI storili že več pomembnih korakov, se pa bomo vidikom FAIR posvečali tudi naprej. Tako je v povezavi z RDA Node Slovenia v načrtu priprava delavnice o certifikaciji CTS in načelih FAIR za slovenske repozitorije raziskovalnih podatkov. ESFRI kažipot 2021 poudarja tudi vedno večjo prisotnost velepodatkov in pomembnost infrastruktur, da jih ustrezno hranijo in obdelujejo. Zaradi vedno večje količine dostopnih besedil, premika s pisnih na govorne in vizualne jezikovne vire ter vse bogatejšega označevanja besedil tudi na področju jezikovnih virov prehajamo v obdobje velepodatkov, kot se že sedaj izkazuje v projektih ParlaMint II, RSDO in MaCoCu. Zato bo CLARIN.SI v naslednjem obdobju podprl uporabo strojne in programske kapacitete za hrambo in predvsem obdelavo velepodatkov. Kažipot tudi poudarja pomembnost raziskovalnih infrastruktur za zajem, hrambo in obdelavo podatkov z družbenih omrežij in spleta. CLARIN.SI je že sedaj posvečal posebno pozornost takšnim jezikovnim virom, v prihodnosti pa bo te aktivnosti še okrepil, ne samo za slovenski, temveč (v okviru centra znanja CLASSLA in projekta MaCoCu) tudi za druge južnoslovanske jezike.

Kažipot poudarja tudi instrumentalizacijo in dostopnost podatkov ter storitev, pomembnih za posamezne skupnosti. Konzorcij CLARIN.SI trenutno vključuje

[27] https://clarin-eric.github.io/parla-clarin/
[28] https://www.clarin.eu/parlamint
[29] https://www.clarin.eu/resource-families
[30] https://www.esfri.eu/esfri-events/esfri-20years-conference
[31] https://www.clarin.eu/content/interoperability
[32] https://www.esfri.eu/esfri-roadmap-2021

12 članic, s čimer sicer pokriva veliko večino slovenskih deležnikov, ki bodisi proizvajajo ali uporabljajo jezikovne vire in tehnologije, ne pa vseh. V naslednjem obdobju se bo CLARIN.SI trudil razširiti svoj konzorcij, s čimer bo pokril tudi skupnosti potencialnih uporabnikov infrastrukture, ki do sedaj še niso bili zajeti v njeno delovanje. CLARIN.SI prav tako načrtuje študijo potreb posameznih skupnosti (raziskovalci in predavatelji s področja humanistike in s področja računalniške lingvistike, slovaropisci, prevajalci, osebe s posebnimi potrebami) in izboljšanje svoje ponudbe v skladu z ugotovitvami.

Kažipot med drugim poudarja pomen izobraževanja, šolanja in podpore pri uporabi infrastruktur za obstoječe in bodoče uporabnike. V prvem obdobju obstoja je bil CLARIN.SI izrazito kadrovsko podhranjen, a smo kljub temu izvedli vrsto dogodkov na delavnicah po Sloveniji in v tujini, predvsem na različnih fakultetah, kjer smo infrastrukturo predstavili študentom. V naslednjem obdobju se bomo načrtno lotili teh aktivnosti z bolj proaktivnim pristopom k izvedbi predavanj in delavnic tako za študente kot tudi za raziskovalce in predavatelje ter razvoju in promociji izobraževalnih materialov.

Nenazadnje je za prihodnost CLARIN.SI pomemben tudi pred kratkim sprejet Načrt razvoja raziskovalne infrastrukture 2030 (NRRI 2030)[33] v Sloveniji, ki ima »v načrtu nadaljevanje in krepitev dejavnosti še v okviru mednarodnih projektov CLARIN« (str. 60), priznava dosedanje sodelovanje z RI DARIAH-SI in CESSDA/ADP, ob tem pa predvideva tudi povezovanje z novima RI, in sicer OPERAS (Odprta znanstvena komunikacija v evropskem raziskovalnem prostoru za družboslovne in humanistične vede)[34], ki jo v Sloveniji vodi ZRC SAZU, in PRACE (Partnerstvo za napredno računalništvo v Evropi)[35], ki jo vodi ARNES.

Zahvala

Delo predstavljeno v prispevku so podprli ARRS v okviru financiranja raziskovalnih infrastruktur ESFRI, Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvo v okviru projektov C3330-19-952059 »Razvoj raziskovalne infrastrukture za mednarodno konkurenčnost slovenskega RRI prostora / RI-SI-CLARIN« in OP20.06780 »Razvoj slovenščine v digitalnem okolju« ter projekti CLARIN ERIC.

Zahvaljujemo se tudi sodelavcem CLARIAH-CZ za pomoč pri nadgradnjah in vzdrževanju platforme repozitorija, sodelavcem Češkega nacionalnega korpusa, predvsem Tomášu Macháleku, za pomoč pri instalaciji konkordančnika KonText in sodelavcem podjetja Lexical Computing, predvsem Janu Bušti in Tomášu Svobodi, za pomoč pri instalaciji konkordančnika Sketch Engine Crystal.

[33] https://www.gov.si/assets/ministrstva/MIZS/Dokumenti/ZNANOST/Novice/NRRI-2030/NRRI-2030_SLO.pdf

[34] https://www.operas-eu.org

[35] https://www.prace-ri.eu

8. Literatura

Jacob Devlin, Ming-Wei Chang, Kenton Lee in Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805. https://doi.org/10.48550/arXiv.1810.04805.

Tomaž Erjavec, Jan Jona Javoršek in Simon Krek. 2014. Raziskovalna infrastruktura CLARIN.SI. V: *Zbornik Devete konference JEZIKOVNE TEHNOLOGIJE*, Ljubljana, 9. - 10. oktober 2014. Slovensko društvo za jezikovne tehnologije. https://nl.ijs.si/isjt14/proceedings/isjt2014_03.pdf.

Tomaž Erjavec, Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Darja Fišer, Cyprian Laskowski in Katja Zupan. 2016. Annotating CLARIN.SI TEI corpora with WebAnno. V: *Proceedings of the CLARIN annual conference*. https://www.clarin.eu/sites/default/files/erjavec-etal-CLARIN2016_paper_17.pdf.

Tomaž Erjavec, Darja Fišer in Nikola Ljubešić. 2021. The KAS corpus of Slovenian academic writing. *Language Resources and Evaluation*, 55(2):551–583. https://rdcu.be/b7GrB.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx in Darja Fišer. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0.

Tomaž Erjavec in Andrej Pančur. V tisku. The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. *Journal of the Text Encoding Initiative*. https://journals.openedition.org/jtei/index.html.

Darja Fišer in Kristina Pahor de Maiti. 2021. "Prvič, sem političarka in ne politik, drugič pa...". *Contributions to Contemporary History*, 61(1). https://doi.org/10.51663/pnz.61.1.07.

Darja Fišer, Jakob Lenardič, Ilze Auziņa, Nan Bernstein Ratner, Koenraad De Smedt, Kaja Dobrovoljc, Réka Dodé, Rickard Domeij, Helge Dyvik, Tomaž Erjavec, Olga Gerassimenko, Jan Hajič, Michal Křen, Nikola Ljubešić, Brian MacWhinney, Monica Monachini, Beatrice Nava, Costanza Navarreta, Aneta Nedyalkova, Klaus Nielsen, Marin Noémi VadászLaak, Susanne Nylund Skog, Lene Offersgaard, Petya Osenova, Valeria Quochi, Sanita Reinsone, Inguna Skadiņa, Kiril Simov, Ondřej Tichý, Noémi Vadász, Tamás Váradi in Kadri Vider. 2019. *Tour de CLARIN Volume Two*. Zenodo. https://doi.org/10.5281/zenodo.3754164.

Darja Fišer, Nikola Ljubešić in Tomaž Erjavec. 2020.

The Janes project: Language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, 54:223–246. https://rdcu.be/7RX4.

Franciska De Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer in Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. V: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7–12, 2018. European Language Resources Association (ELRA). https://aclanthology.org/L18-1515.

Franciska De Jong, Bente Maegaard, Darja Fišer, Dieter Van Uytvanck in Andreas Witt. 2020. Interoperability in an infrastructure enabling multidisciplinary research: The case of CLARIN. V: *Proceedings of the 12th Language Resources and Evaluation Conference*, str. 3406–3413. European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.417/.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý in Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography*, 1:7–36.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban in Igor Mozetič. 2015. *Emoji Sentiment Ranking 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1048.

Simon Krek. 2022. Delivrable D1.31: Report on the Slovenian Language. Tehnično poročilo, European Language Equality Project. https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_31__Language_Report_Slovenian_.pdf.

Luka Krsnik, Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Aleksander Ključevšek, Simon Krek in Marko Robnik-Šikonja. 2019. Corpus extraction tool LIST 1.2. http://hdl.handle.net/11356/1276.

Jakob Lenardič in Darja Fišer. 2022. CLARIN Depositing Guidelines: State of Affairs and Proposals for Improvement. V: *Proceedings of the CLARIN Annual Conference*, Prague, Czech Republic, October 10–12, 2022. https://www.clarin.eu/event/2022/clarin-annual-conference-2022.

Nikola Ljubešić, Filip Markoski, Elena Markoska in Tomaž Erjavec. 2021. *Comparable corpora of South-Slavic Wikipedias CLASSLA-Wikipedia 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1427.

Tomáš Machálek. 2020. KonText: Advanced and Flexible Corpus Query Interface. V: *Proceedings of the 12th Language Resources and Evaluation Conference*, str. 7003–7008, Marseille, France, May. European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.865.

Katja Meden in Tomaž Erjavec. 2021. Pregled slovenskih repozitorijev raziskovalnih podatkov. Tehnično poročilo, Jožef Stefan Institute. https://www.clarin.si/info/services/projects/#RDA_Node_Slovenia.

Andrej Pančur, Tomaž Erjavec, Mihael Ojsteršek, Mojca Šorn in Neja Blaj Hribar. 2020. Slovenian parliamentary corpus (1990-2018) siParl 2.0. http://hdl.handle.net/11356/1300.

Pavel Rychlý. 2007. Manatee/Bonito - A Modular Corpus Manager. V: *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, str. 65–70, Brno. Masarykova univerzita.

Matej Ulčar in Marko Robnik-Šikonja. 2021. *Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1397.

Tamás Váradi, Steven Krauwer, Peter Wittenburg, Martin Wynne in Kimmo Koskenniemi. 2008. CLARIN: Common language resources and technology infrastructure. V: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/317_paper.pdf.

Darinka Verdonik, Tomaž Potočnik, Mirjam Sepesy Maučec, Tomaž Erjavec, Simona Majhenič in Andrej Žgank. 2019. Spoken corpus Gos VideoLectures 4.0 (transcription). http://hdl.handle.net/11356/1223.

Špela Vintar in Matej Martinc. 2022. Framing karstology: From definitions to knowledge structures and automatic frame population. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 28(1):129–156.

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

Aleš Žagar, Matic Kavaš, Marko Robnik-Šikonja, Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Marko Ferme, Mladen Borovič, Borko Bošković, Milan Ojsteršek in Goran Hrovat. 2022. Corpus of academic Slovene KAS 2.0. http://hdl.handle.net/11356/1448.

ILiAD: An Interactive Corpus for Linguistic Annotated Data from Twitter Posts

Simon Gonzalez

The Australian National University,
Canberra, Australian Capital Territory, Australia
u1037706@anu.edu.au

Abstract

Social Media platforms have offered invaluable opportunities for linguistic research. The availability of up-to-date data, coming from any part in the world, and coming from natural contexts, has allowed researchers to study language in real time. One of the fields that has made great use of social media platforms is Corpus Linguistics. There is currently a wide range of projects which have been able to successfully create corpora from social media. In this paper, we present the development and deployment of a linguistic corpus from Twitter posts in English, coming from 26 news agencies and 27 individuals. The main goal was to create a fully annotated English corpus for linguistic analysis. We include information on morphology and syntax, as well as NLP features such as tokenization, lemmas, and n-grams. The information is presented through a range of powerful visualisations for users to explore linguistic patterns in the corpus. With this tool, we aim to contribute to the area of language technologies applied to linguistic research.

1. Introduction

In this current age, the use of social media platforms has permeated across all circles of society, from personal communication to government communications. Its impact is hard to be overstated. It is considered as a form of mass media, but distinctive from other forms such as television and radio, where the information is presented by a specific broadcasting mechanism (Page et al., 2014). In the case of social media, the content can be delivered by anyone, making it more personal and individual than other mass forms. The adopting of this technology in language research has been an organic and necessary process. This is because language research investigates the use of language in society, and since social media is a medium of language, we need to understand how we use language in this digital world.

One framework that has efficiently paved the way for linguists to examine social media language is Computer-Mediated Communication (CMC). This has been defined as a relatively new 'genre' of communication (Herring, 2001), which can involve a diversity of forms connecting the physical and the digital (Boyd and Heer, 2006). One of the focus of study in CMC research is on the intrinsic characteristics of digital language, e.g. stylistics, use of words, semantics, and other relevant linguistic features. This has been done for various CMC types, including social media.

But describing social media features is not a straight-forward task because it is not a homogenous genre. It has a diversity of types depending on the main shareable content (e.g., YouTube for videos, Twitter for texts[1]) and main format (e.g., Reddit as a discussion forum, Pinterest Product pins for products purchase), for example. But one common feature is that all platforms have an interactive component in which users can express ideas, comment, and reply to other people's perspectives. The inherent communicative aspect in this social interaction is one that has strong implications in linguistic research, which is that when we analyse language from social media, we look at how language is used in natural contexts, with concrete communicational purposes. What distinguishes then our approach as language researchers, from engineers and app developers, for example, is that we are interested to study how people use technology to communicate and describe what makes it a distinctive type of language (Page et al., 2014). In this sense, we are interested in identifying the language patterns as used in social media platforms, knowing that patterns found in social media are not necessarily representative of language patterns in other contexts. This has been demonstrated empirically by Grieve et al. (2019), where they compared Twitter data versus traditional survey data. They found that some patterns were observed more strongly in Twitter data than in the survey data. Results like these are evidence then that when we deal with social media language, we are examining a way of expression, which has features like other language forms, but at the same time it has its own distinctive characteristics. This is paramount to be considered when new language analysis technologies are developed.

1.1. Twitter and Corpus Linguistics

The combination of language research and social media is a complex endeavor, making people working with both apply skills that are necessary in this interdisciplinary undertaking. One area that reflects this complexity and that has efficiently adapted social media is Corpus Linguistics (CL). A strong characteristic of CL is that it is used to collect, store, and facilitate language analysis for large datasets (Szmrecsanyi, 2011; Grieve, 2015). And with the advantage of having more sophisticated tools available, such as in social media research, corpora are becoming larger and larger, with the only constraints being computational power and storage capacity.

Many social media platforms have been widely used for language and linguistic research (c.f. Liew and Hassan, 2021; Nagase et al., 2021; Trius and Papka, 2022; Wincana et al, 2022). Out of these platforms, Twitter stands out due to its world spread, and the option it gives to researchers when stratifying the demographics of user accounts, including the use of the geo-code and time-stamp

[1] The type of content of social media platforms is not restricted to only one. This is just an example on the main purpose for specific cases. For instance, YouTube allows users to write comments on videos and Twitter can embed videos on posts.

information of the posts[2] (Grieve et al., 2018). It is classified as a microblogging site (Chandra et al. 2021) where the content can be on opinions, news, arguments, and other types of sentences (Chaturvedi et al., 2018). Because of their wide-spread use, it has been used in the creation of numerous corpora created from Twitter posts (c.f. Dijkstra et al., 2021; Grieve et al., 2019, Tellez et al., 2021).

1.2. Current Project

In this paper, we present the development of a web-based corpus from Twitter posts, named *ILiAD: An Interactive Corpus for Linguistic Annotated Data*. In relation to our methodological approach, we propose that corpora built from social media helps study the patterns of language used in this context and capture their linguistic complexity. By doing this, we can have a better view of the multilayered nature of the corpus.

2. Goal of the Paper

The aim of the corpus is to capture the linguistic complexities used in Twitter language, and we chose two types of account users: news agencies and individuals. We explore the differences between their structures and patterns. The language of journalism is characterised based on its main purpose: exert influence on readers and convince them on a specific interpretation (Fer, 2018; Moschonas, 2014). This is achieved by three main stylistic features. The first one is language clarity, a feature that is strongly appropriate for journalism more than many other language styles. The second one is accuracy. This refers to the ability to convey ideas accurately and avoiding ambiguities in interpretation. The final one is the simplicity. This aims to convey messages without the use of complex words that may blur the intention of the message (Fer, 2018). The aim therefore is to prepare the corpus for further exploration, querying and analysis to understand the language used in Twitter.

The analysis can focus on many linguistic parameters and here we approach it in an integrated way. This can give users the opportunity to explore the corpus from different angles and linguistic perspectives.

3. Methodology

The stages of data collection, data processing, and app deployment were carried out in R (R Core Team, 2021), using *shiny* R (Chang et al., 2021) for the app development. Apps developed in shiny have three main advantages. The first one is its interactivity capability. With this, users can interact with the whole corpus, across a range of visualisation outputs and tables. The second one is its reactive power. With this, users modify parameters in the tables and visualisations, and the app changes outputs based on user inputs. The positive impact on corpus linguistics is invaluable. With these features, a corpus can be used to have a full understanding on the shape of the data as well as an exploration of patterns.

3.1. Data Collection

We applied four criteria to identify the Twitter accounts to be included in the corpus. The first criterion was that account users (news agencies and individuals) had to have English as the main language of communication. The second one was that accounts had to be active at the moment of extraction. The reason was to capture tweets that were synchronous and where topics and trends could be shared across accounts. The third criterion was that accounts had to have a large number of tweets, enough to reach over 3,000. This was done to make sure that enough posts were left after the filters were applied, which is explained below. The final criterion was to only include those users whose posts were not mainly retweets. This filter aimed to exclude those accounts that do not produce content but only retweet posts from other accounts. From this, we identified 29 news agencies, and 27 individual accounts. The percentages are shown in Table 1.

User Type	*Total Tweets*	*Percentage*
News Agency	84,354	54%
Individual	71,477	46%
Total	**155,831**	**100%**

Table 1: Total number of tweets in the corpus and their proportions per account type.

The data extraction was done through an R script developed by the main author. We used the *rTweet* (Kearney, 2019) package, which allows users to gather Twitter posts by the free Twitter API, giving a total of over 156,000 tweets.

Year	*Total Tweets*	*Percentage*
2009	139	0.1%
2010	178	0.1%
2011	497	0.3%
2012	2230	1.4%
2013	5097	3.3%
2014	3625	2.3%
2015	5159	3.3%
2016	6745	4.3%
2017	5508	3.5%
2018	6301	4.0%
2019	7847	5.0%
2020	18742	12.0%
2021	20697	13.3%
2022	73066	46.9%
Total	**155,831**	**100%**

Table 2: Total number of tweets in the corpus and their proportions per year.

3.2. Data Processing

From the collected data, we applied six filters to make sure that the corpus reflects comparable linguistic data for all account users. The first filter was to exclude tweets that were not in English (n=10,067; 6%). This was done by filtering out those tweets which did not have the English (*en*) assigned by Twitter's machine language detection,

[2] The geo-code information is optional in Twitter, and the user decides whether to show this or not. Other approaches include running algorithms that identify locations based on factors such as time zone and language features, which are used to infer locations.

which is annotated in the tweet's metadata. The second filter was to exclude re-tweets (n=23260; 15%). This restricts the data only to those posts that come from the given user and not from other accounts. The third filter was to exclude quote tweets (n=7,142; 5%). These are tweets that are re-tweeted with an added comment from the user. Keeping quote tweets in the data would add repeated tweets to the corpus and also would add patterns and word counts that do not correspond to a specified account. The fourth filter deleted repeated tweets (n=778; 0.5%). This targeted those cases in which account users write the same content and post it as a separate tweet, but not as a re-tweet. Similar to quote tweets, keeping repeated tweets would inflate the content of the corpus and it would not be representative. For the fifth filter, we excluded strings that were URL links, which do not have linguistic features[3] of interest in this paper (n=1,208; 0.8%). For the sixth and last filter, we first calculated the number of words for each tweet, which were split by white spaces to get the number of individual words. We then excluded those tweets that had a length of less than eight words (n=14,125; 9%). This filter targets those tweets which do not have linguistic content but only social media features such as hashtags or links.

With these filters, the final data contained 112,690 tweets. This is a loss of 28% (n = 43,919) of the original data exported from the Twitter API.

3.3. Text Processing

After data filtering, we implemented a wide range of *Natural Language Processing* (NLP) techniques for the data wrangling and analysis. We carried out the text processing using the *UDPipe* (Straka and Straková, 2017) package as the main tool for the NLP tasks. UDPipe is defined as single tool which contains a tokenizer, morphological analyzer, Parts-Of-Speech tagger, lemmatizer, and a dependency parser. It currently offers 77 language models, with some languages having more than one model available. We used the *EWT* English model available in the package. We selected the text column from the API output and made it the input for the main UDPipe function. The core purpose of the UDPipe package is to create a single-model tool for a given language which can be used to process raw text and convert it to a CoNLL-U-formatted text. This format stores tagged information for all words in dependency trees, including morphological and syntactic features (Straka and Straková, 2017). From this format, the UDPipe algorithm creates morphological taggers and dependency parsers. The main taggers are described below.

3.3.1. Tokenization

The tokenization tools are wrapped within a trainable tokenizer based on artificial neural networks, specifically, the bidirectional LSTM artificial neural network (Graves and Schmidhuber, 2005) and a gated linear unit – GRU (Cho et al., 2014). It works by comparing the words in the input text to the trained tokenizer and does not add any additional knowledge about the language. If a given word, or group of words, is not recognized, the tokenizer tries to reconstruct it by utilizing an additional raw text corpus.

3.3.2. Morphological Analysis

There are three main fields tagged in the data process:

1. Part-of-speech tagging
2. Morphological features
3. Lemma or stem

The parts-of-speech tagging uses *MorphoDiTa* (Straková et al., 2014). The tagging process exploits the rich linguistic features of inflective languages with large number of suffixes, where multiple forms can be related to a single lemma. From this, the tagger estimates common patterns on endings and creates morphological templates from the observed clusters. On Table 3, we show the top ten counts and proportions of Parts-Of-Speech tags in the current corpus, as output from UDPipe.

POS	*Corpus Count*	*Percentage*
NOUN	76,795	20.8%
VERB	62,537	16.9%
ADP	39,237	10.6%
PROPN	37,862	10.3%
PRON	37,399	10.1%
DET	31,284	8.5%
PUNCT	30,001	8.1%
ADJ	24,452	6.6%
ADV	16,425	4.4%
AUX	13,171	3.6%

Table 3: Total number of top ten Part-Of-Speech tags in the corpus and their proportions.

3.3.3. Classification Features

UDPipe uses two models that facilitate the tagging process and improve the overall accuracy by employing different classification feature sets. The first one the POS tagger, which disambiguates all available morphological fields in the data. The second model, a lemmatizer, disambiguates the lemmas tagged.

3.3.4. Dependency Parsing

Dependency parsers are part of the family of grammar formalisms called *dependency grammars* (Jurafsky and Martin, 2021). In these, the syntactic structure sentences are described on the grammatical relations between the words, shown as directed binary dependencies. All structures start at the root node of the tree, and then components and the dependencies are shown throughout the entire structure. Dependency parser trees can deal very efficiently with languages that are rich morphologically and also have a relatively free word order, for example Spanish, Czech, and English, with varying flexibility. Another important advantage of using dependency parsers is that they allow closer examination of semantic relationships between arguments in the sentence.

Summing up, the features, descriptions, and tagging done by the UDPipe framework, offer invaluable information relevant for linguistic analysis used in Corpus Linguistics. With these features extracted for all tweets, we have information available at different layers for linguistic

[3] URL Links are an important aspect of social media language. However, its analysis is beyond the scope of this paper.

analysis: morphological, syntactic, and even semantic, through the dependency parsers.

3.4. Data Filtering

After obtaining the output from the UDPipe package, we proceeded to filter the data. The motivation was to prepare it for the linguistic analysis within the corpus. This filtering process affects two dataset outputs which used for different purposes in the corpus. The first one is used for calculating n-grams and word frequencies. The second one is for showing Syntactic Dependencies.

3.4.1. Token Filtering

Identifying the right tokens in social media language is a difficult process. The correct practice in this step is crucial to achieve efficient outcomes. This filtering differs from the practice done on other language media such as the language in newspapers, television, and academic papers. Following O'Connor et al., (2010), we excluded tokens containing hashtags, URL links, @-replies, strings of punctuation, and emoticons[4]. Their proportions are shown in Table 4.

Content Excluded	*Total Count*	*Percentage*
Emoticons	1,556	0.4%
Hashtags	1,986	0.5%
URL Links	2,857	0.7%
@-replies	3,851	0.9%
Punctuation	30,001	7.3%

Table 4: Total number of social media content excluded and their proportions in the whole corpus.

3.4.2. Removing Stop Words

Following standard procedures, we removed stop words for calculating n-grams and word frequencies. An important observation is that removing stop words is a compromise for the corpus, since certain word combinations are affected, especially those which appear together with the words in the list. Future versions of this work aim to efficiently implement analysis considering the role of stop words in the corpus.

Here we removed stop words by following the steps below:

1. First, we selected a list of stops words from the *stopwords* (Benoit et al., 2021) package in R. We selected the ones used for English and it included 175 words (see Table 5 for the top 15).
2. Next, we filtered out the stop words in this data subset.
3. Finally, we filtered out stop words that are specific for Twitter, and that includes words such as *RT*, *follow*, *follows*, and *following*. In future versions, we aim to implement a disambiguation algorithm where a key word, such as follow, can be identified as a word used in social media context (e.g. *follow us on Twitter*), or in a more traditional one (e.g. *follow the road*).

Stop word	*Total Count*	*Percentage*
the	17,151	4.18%
to	11,543	2.81%
a	9,076	2.21%
be	8,480	2.07%
and	7,844	1.91%
of	7,001	1.71%
I	6,734	1.64%
in	6,429	1.57%
you	5,315	1.3%
have	4,083	1%
that	3,933	0.96%
it	3,803	0.93%
for	3,793	0.92%
on	3,552	0.87%
he	3,442	0.84%

Table 5: Top 15 stop words excluded in the data subset and their proportions in the corpus.

3.4.3. Sentence Structure Filtering

In this filter, we aimed to identify those posts which were not linguistic phrases or sentences, thus including only those structures that were classified into a sentence category. For each of the tweet breakdown done by UDPipe (as shown in Table 6), we looked at the PUNCT classification, where we identified three types of sentences: statements (ending with "."), questions (ending with "?") and exclamations (ending with "!"). Any unclassified sentence was deleted from the dataset. Deciding to keep sentences that follow the standard punctuation symbols has a strong impact in a corpus based on Twitter language, since sentences here can follow other rules, e.g. ending a sentence with emoticons or other use of punctuation symbols, such as !!! or :). However, an important number of sentences follow the most standard use of punctuation symbols, which is a reliable representation of the data collected. Finally, for each sentence, we checked whether there was a conjugated verb. For those sentences which had no conjugated verbs, the identified sentence was deleted from the dataset used for the Syntactic Dependency section. For this, we created a data subset that only contained sentences and their corresponding classification done in the previous step. This was the input for the Section explained in 4.1.2.

token	*upos*	*feats*	*dep_rel*
Senate	PROPN	Number=Sing	nsubj
needs	VERB	Mood=Ind	root
to	PART		mark
think	VERB	VerbForm=Inf	xcomp
and	CCONJ		cc
vote	VERB	VerbForm=Inf	conj

Table 6: Sample output from *UDPipe.*

3.5. Calculating N-Grams

By implementing NLP techniques, this brings more depth to the corpora analysis since it allows users to explore more areas in the data. In the current version of the app, we

[4] Emoticons entail rich linguistic information. However, their analysis is not included in this version of the tool.

use unigram and bigram explorations. The n-grams are calculated using the *tidytext* (Silge and Robinson, 2016) package. We followed the established approach of deleting stops words in English, using the *stopwords* package. After the filtering, the n-grams were calculated across all the data.

3.6. Entity Identification

A second group of NLP techniques implemented is the identification of entities in the corpus, and that includes mentions of people, physical locations, and established organisations. We used the *entity* (Rinker, 2017) package for this purpose. This package is a wrapper to simplify and extend the *NLP* (Hornik, 2020) package and the *openNLP* (Hornik, 2019) package named entity recognition. The advantage of this approach is that we can use it to detect important information, which is crucial especially in large datasets, that can be captured when identifying entities. This also has a strong impact on our understanding of linguistic features, since they are related to important elements in sentences, such as nouns and adjectives. By implementing this, the app brings more depth to the corpora analysis since it allows users to explore the main entities in the corpus.

3.7. Twitter Metrics

The final metrics measured and obtained aims to show information that is relevant when dealing with Twitter data. The motivation is to be able to contextualise the information in the corpus within the overall world of social media. The information presented here is extracted from the Twitter API output, which means that we display two features publicly available. The first one is the number of tweets across time. We also include a general summary of the main sour locations by country of the tweets contributing to the data. Previous studies (c.f. Grieve et al., 2019) have demonstrated that the use of geo-coding information is relevant for linguistic studies, but here, we only show the country of origin of all tweets without identifying individuals or linking linguistic features to any demographics.

4. App Infrastructure

The app was developed in RStudio, which has been widely used for corpus linguistics development and related tasks (Abeille and Godard, 2000; Gries, 2009), and the main framework was within shiny R. Shiny apps allow great interactivity and responsiveness. Interactivity allows users to explore visualizations in effective ways, and responsiveness allows users to navigate contents in real time, with the use of clicks and dropdown menus. Other libraries that we used for the creation of visuals were *ggplot2* (Wickham, 2016) and *echarts4r* (Coene, 2022). ggplot2 allows a great degree of flexibility when creating figures. This is relevant since there is a lot of complexity of the linguistic data that we present. But this allows complex ideas to be presented in a digestible way. Another advantage of this is that it allows users to see data points within the general context as well as being able to narrow down into more specific analysis. This creates a seamless navigation of linguistic data in an efficient way. The presentation of the app components was divided into two main sections. The first component gives users tools to explore linguistic features and the second one gives information on Twitter metrics. Due to the limitations on Twitter Terms of Service, the app cannot display the raw tweets as a database format nor give the option to download data. The interactive tool therefore focuses on the presentation of the linguistic features derived from the data.

4.1. Exploring Calculated Features

The linguistic features are the main backbone of the corpus. In this section, there are visualisation options that can be used to have both a broad understanding of patterns, as well as a deep exploration of linguistic features.

4.1.1. Parts of Speech

This section gives the overall statistics of the words classified into their POS, including distributions and proportions per year and sentence type. The exploration can be done in different levels: all corpus or by user type (news agencies or individuals). The input data in this section comes from the **Sentence Structure Filtering Section (3.4.3)**.

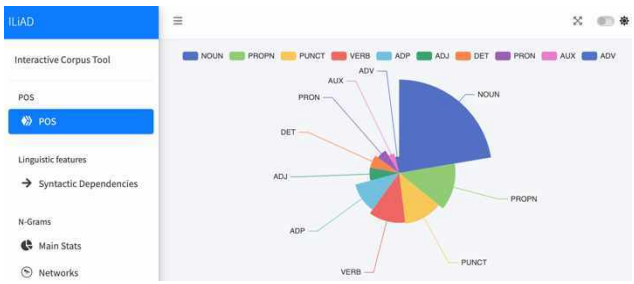


Figure 1: POS Distributions Tab.

4.1.2. Syntactic Dependencies

This section allows users to explore the syntactic dependencies of all the available sentences. Here we use a combination of the UDPipe output and the *textplot* (Wijffels et al., 2021) package, which creates the dependencies as in the figure below. Since users can select all available sentences, this is a powerful function than can be used to explore syntactic patterns across the corpus and facilitates the understanding of syntactic structures. The input data in this section comes from the **Sentence Structure Filtering Section (3.4.3).**

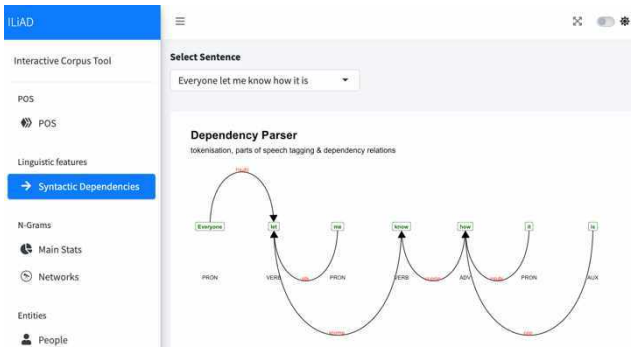


Figure 2: Syntactic Dependencies Tab.

4.1.3. Exploring N-Grams

N-grams are explored through visualisations, including connection networks. These networks are developed within the *Network Analysis* (NA) approach. The power of this analysis comes from its capability of observing

relationships between components. This technique has been implemented in other fields, such as psychology (Jones et al., 2021; Mullarkey at al., 2019), and social network research (Clifton and Webster, 2017; Würschinger, 2021). NA visualizations follow the assumption that if a relationship is meaningful within the whole network, it will stand out from other relationships by stronger connections than random or weaker relationships. In this analysis, the connections are based on the frequencies which connect n-grams. Here we use the functionality from the *visNetwork* (Almende, 2021) package.

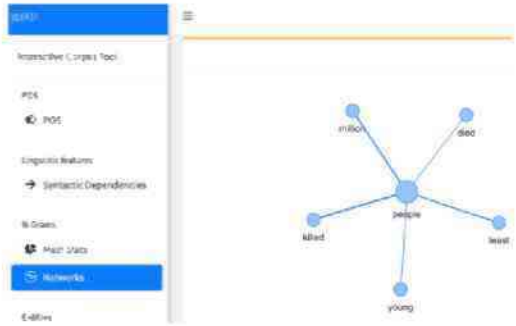


Figure 3: N-Grams Visualisation Tab showing a network relationships.

4.1.4. Exploring Entities

We use a different visualization approach for the entities captured in the corpus. We use bar plots and word clouds. The advantage of bar plots is that they show the frequencies in a way that we can see from the most frequent to the least frequent, organized from left (most frequent) to right (least frequent). Word clouds are an easy and user-friendly way to represent frequencies. Here, more frequent words are represented with larger fonts than less frequent words. An example for the organizations mentioned in the corpus is shown in the figure below. At the top, we see the bar plot and at the bottom the word cloud.

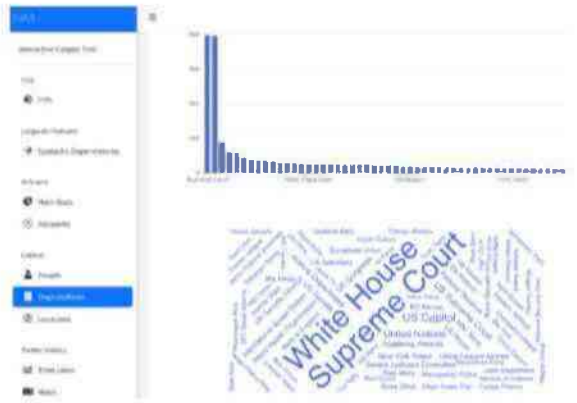


Figure 4: Named Entities Visualisation Tab.

4.2. Twitter Data Metrics

The final section shows relevant Twitter data metrics, for which we dedicate two sections. The first one is a timeline visualization using a combination of the *ggplot2* package and the *plotly* (Sievert, 2020) package. This combination gives ggplot2 plots interactive power. The timeline displays the number of posts across time, for all the data available in the corpus. This timeline can also be selected to observe by account type, giving more granularity of exploration. Another timeline visualization is applied to N-grams. This has been used to observe lexical innovations (c.f. Grieve et al., 2019), by looking at N-grams that increase in terms of frequency across time. This tool can facilitate this type of analysis.

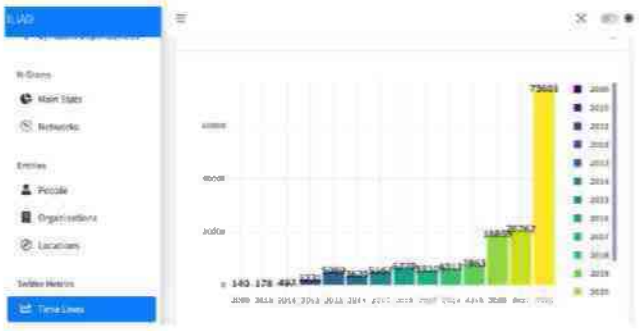


Figure 5: Twitter Timeseries Count Tab.

The second visualization implemented is a world map showing the region source information of tweets. The purpose is to visualize the main geographical areas from where the tweets come. We use the functionality from *echarts4r* package, which is very efficient at displaying this type of information, as well as being interactive in an online context.

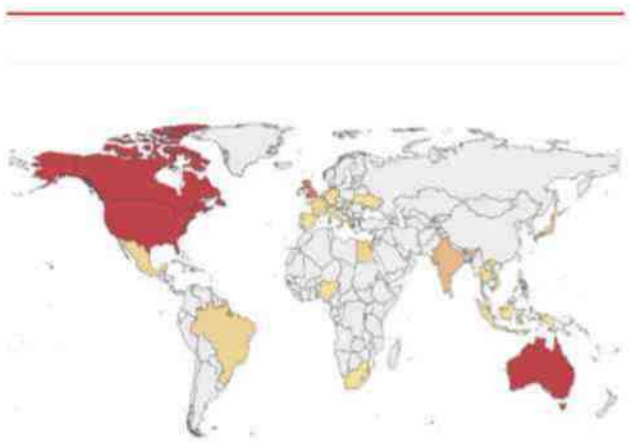


Figure 6: Twitter Map Tab.

5. Discussion

The app presents a wide range of visualizations and analyses from the Twitter corpus. The features capture different linguistic layers, including morphology, syntax, and n-grams. With the inclusion of Twitter metrics, this tool gives all exploration opportunities to understand the whole corpus. R and shiny R have proven to be an efficient combination to develop and deploy the corpus. For the text processing tasks, the use of the UDPipe and tidytext packages have been highly effective. The in-built functions have been used and we have created our custom-made functions to complete the tasks done throughout the whole process. For visualization tasks, the combination of ggplot2, plotly, visNetwork, and echarts4r has demonstrated efficient to represent complex linguistic features and relationship analysis. The app can be accessed through the following GitHub repository: https://github.com/simongonzalez/ILiAD.

6. Conclusion

In this paper, we have presented the development of a linguistic corpus based on the Twitter posts. It has been designed to be used by a diversity of audiences who are interested in exploring linguistic patterns from corpora based on social media language. Similar tools have been developed with invaluable contributions to the field of Corpus Linguistics. Our proposal, however, makes stronger integrations with a variety of visualization types that enhance the analysis in a holistic way. The tool also gives users interactive and reactive power throughout all the data, which not only offers a corpus to analyse, but a corpus to interact with and query in a more organic way, compared to more traditional approaches of presenting corpora. Finally, it has been developed within an open-source framework, making it freely available to any user interested in using and even expanding this tool.

7. Future Work

In the current version, we have selected a relatively small number of users in the corpus, as compared to other larger projects with similar goals. This is to allow the implementation of the interactive capability in the visualization methods, which requires a high level of computational power. We aim to add more data in future versions using more efficient processing algorithms. Finally, we see the value of adding linguistic analysis to emoticons. In a future version, we aim to include analysis on emoticons, as a distinctive component of social media language.

8. Acknowledgements

I want to thank the anonymous reviewers of this paper for their invaluable comments and insights in the shape and content of the final version. Their generosity and expertise have improved this paper in innumerable ways and saved me from many errors. Those that inevitably remain are entirely my own responsibility.

9. References

Anne Abeillé and Danièle Godard. 2000. French word order and lexical weight. In: R. Borsley, ed., *The Nature and Function of Syntactic Categories, Syntax and Semantics*, Academic Press, pages 325–358.

Benoit Thieurmel. 2021. *visNetwork: Network Visualization using 'vis.js' Library*. R Package Version 2.1.0.

Kenneth Benoit, David Muhr and Kohei Watanabe. 2021. *stopwords: Multilingual Stopword Lists*, (Version 2.2) [R package]. Retrieved from https://github.com/quanteda/stopwords

Danah Boyd and Jeffrey Heer. 2006. Profiles as Conversation: Networked Identity Performance on Friendster. In: *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, 2006, pages 59c–59c, doi: 10.1109/HICSS.2006.394.

Subhadip Chandra, Randrita Sarkar, Sayon Islam, Soham Nandi, Avishto Banerjee and Krishnendu Chatterjee. 2021. Sentiment Analysis on Twitter Data: A Comparative Approach. *International Journal of Computer Science and Mobile Applications*, 9(10):01–12.

Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges. 2019. *shiny: Web Application Framework for R* (Version 1.3.2) [R package]. Retrieved from https://CRAN.R-project.org/package=shiny

Snigdha Chaturvedi, Shashank Srivastava and Dan Roth. 2018. Where have I heard this story before? Identifying narrative similarity in movie remakes. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2, pages 673–678. New Orleans, Louisiana. Association for Computational Linguistics.

Allan Clifton and Gregory D. Webster. 2017. An introduction to social network analysis for personality and social psychologists. *Social Psychological and Personality Science*, 8(4):442–453.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

John Coene. 2022. *echarts4r: Create Interactive Graphs with 'Echarts JavaScript',* Version 5. https://echarts4r.john-coene.com/.

Jelske Dijkstra, Wilbert Heeringa, Lysbeth Jongbloed-Faber and Hans Van de Velde. 2021. Using Twitter Data for the Study of Language Change in Low-Resource Languages. A Panel Study of Relative Pronouns in Frisian. *Frontiers in Artificial Intelligence*, 4:644554.

Simona Fer. 2018. *The Language of Journalism: Particularities and Interpretation of Its Coexistence with Other Languages* (February 22, 2018). Available at SSRN: https://ssrn.com/abstract=3128134 or http://dx.doi.org/10.2139/ssrn.3128134

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, pages 5–6.

Stefan Th. Gries. 2009. *Quantitative Corpus Linguistics with R.* London and New York: Routledge.

Jack Grieve1, Chris Montgomery, Andrea Nini, Akira Murakami and Diansheng Guo. 2019. Mapping Lexical Dialect Variation in British English Using Twitter. *Frontiers in Artificial Intelligence*. 2(11). doi: 10.3389/frai.2019.00011.

Jack Grieve, Andrea Nini and Diansheng Guo. 2018. Mapping Lexical Innovation on American Social Media. *Journal of English Linguistics*, Vol. 46, pages 293–319.

Jack Grieve. 2015. Dialect Variation. In: Douglas Biber and Randi Reppen, eds., *The Cambridge Handbook of English Corpus Linguistics*. Cambridge University Press.

Susan C. Herring. 2001. Computer-mediated discourse. In: D. Schiffrin, D. Tannen and H. Hamilton, eds., *The Handbook of Discourse Analysis*, (Oxford: Blackwell Publishers), pages 612–634.

Kurt Hornik. 2019. *openNLP: Apache OpenNLP Tools Interface*, (Version 0.2-7) [R package].

Kurt Hornik. 2020. *NLP: Natural Language Processing Infrastructure*, (Version 0.2-1) [R package].

Payton J. Jones, Ruofan Ma and Richard J. McNally. 2021. Bridge Centrality: A Network Approach to Understanding Comorbidity. *Multivariate Behavioral*

Research, 56(2):353–367, DOI: 10.1080/00273171.2019.1614898 (2021).

Daniel Jurafsky and James H. Martin. 2021. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.* All rights reserved. Draft of June December 29, 2021.

Michael W. Kearney. 2019. rtweet: Collecting and analyzing Twitter data. *Journal of Open Source Software*, 4(42), 1829. doi: 10.21105/joss.01829, R package version 0.7.0, https://joss.theoj.org/papers/10.21105/joss.01829.

Tze Siew Liew and Hanita Hassan. 2021. The search for national identity in the discourse analysis of YouTube comments. *Journal of Language and Linguistic Studies*.

Spiros Moschonas. 2014. The Media On Media-Induced Language Change. In: J. Androutsopoulos, ed., *Mediatization and Sociolinguistic Change*, pages 395–426. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110346831.395.

Michael C. Mullarkey, Igor Marchetti and Christopher G. Beevers. 2019. Using Network Analysis to Identify Central Symptoms of Adolescent Depression. *Journal of Clinical Child and Adolescent Psychology*, 48(4):656–668, DOI: 10.1080/15374416.2018.1437735 (2019).

Ryotaro Nagase, Takahiro Fukumori and Yoichi Yamashita. 2021. Speech Emotion Recognition with Fusion of Acoustic- and Linguistic-Feature-Based Decisions. In: *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),* pages 725–730.

Brendan O'Connor, Michel Krieger and David Ahn. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. *ICWSM*.

Ruth Page, David Barton, Carmen Lee, Johann Wolfgang Unger and Michele Zappavigna. 2014. *Researching Language and Social Media* (1st ed.). Taylor and Francis. Retrieved from https://www.perlego.com/book/1559453/researching-language-and-social-media-pdf (Original work published 2014)

R Core Team 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Tyler Rinker. 2017. *entity: Named Entity Recognition*, (Version 0.1.0) [R package].

Carson Sievert. 2020. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. ISBN 9781138331457, https://plotly-r.com.

Julia Silge and David Robinson. 2016. tidytext: Text Mining and Analysis Using Tidy Data Principles. In: *JOSS*, 1(3). doi:10.21105/joss.00037, http://dx.doi.org/10.21105/joss.00037.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Jana Straková, Milan Straka and Jan Hajič. 2014. Open-source tools for morphology, lemmatization, pos tagging and named entity recognition. In: *Proceedings of 52nd An- nual Meeting of the Association for Computational Lin- guistics: System Demonstrations*, pages 13–18, Balti- more, Maryland, June. Association for Computational Linguistics.

Benedikt Szmrecsanyi. 2011. Corpus-based dialectometry: a methodological sketch. *Corpora*, 6(1):45–76. DOI: 10.3366/cor.2011.0004 | preprint pdf

Eric S. Tellez, Daniela Moctezuma, Sabino Miranda and Mario Graff. 2021. A large scale lexical and semantic analysis of Spanish language variations in Twitter. *ArXiv*, abs/2110.06128.

Lilia I. Trius and Nataliya V. Papka. 2022. Some Aspects of Online Discourse Manipulation on Social Media (the case of Instagram English Presentational Discourse of Pfizer Inc.). *Current Issues in Philology and Pedagogical Linguistics*.

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Jan Wijffels, Sacha Epskamp, Ingo Feinerer and Kurt Hornik. 2021. *textplot: Visualise complex relations in texts*, (Version 0.2.0) [R package]. Retrieved from https://github.com/bnosac/textplot

Gita Wincana, Wahyudi Rahmat and Ricci Gemarni Tatalia. 2022. Linguistic Tendencies of Anorexia Nervosa on Social Media Users Facebook (Pragmatic Study). *Journal of Pragmatics and Discourse Research*.

Quirin Würschinger. 2021. Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter. *Frontiers in Artificial Intelligence*. 4:648583. doi: 10.3389/frai.2021.648583. PMID: 34790894; PMCID: PMC8591557.

Raba *Kolokacijskega slovarja sodobne slovenščine* pri prevajanju kolokacij

Martin Anton Grad,* Nataša Hirci†

*,† Oddelek za prevajalstvo, Filozofska fakulteta Univerze v Ljubljani
Aškerčeva 2, 1000 Ljubljana
martin.grad@ff.uni-lj.si
natasa.hirci@ff.uni-lj.si

Povzetek

Prispevek povzema izsledke raziskave, ki se je ukvarjala z rabo *Kolokacijskega slovarja sodobne slovenščine* (KSSS) pri prevajanju kolokacij iz angleščine v slovenščino. Dodiplomski študenti Oddelka za prevajalstvo FF UL so prevajali besedilo, v katerem je bilo označenih deset kolokacij. Med procesom prevajanja se je dogajanje na zaslonu snemalo, tako da je bilo po koncu naloge mogoče analizirati tako prevodne rešitve posameznih kolokacij kot tudi prevajalski proces, pri čemer smo se osredotočali zlasti na rabo jezikovnih virov, med katerimi nas je najbolj zanimala raba KSSS. Rezultati so pokazali, da je vključevanje KSSS v pedagoški proces uspešno, saj so vsi študenti, ki so sodelovali v raziskavi, z njim seznanjeni in ga pri svojem prevajalskem delu tudi aktivno uporabljajo. Pri tem so se med študenti pokazale občutne razlike glede tega, kako dobro poznajo napredne funkcije, ki jih KSSS nudi, in posledično kako uspešni so pri iskanju ustreznih kolokacij. Raziskava je pokazala tudi, da raba jezikovnih virov ne vodi nujno do optimalne prevodne rešitve.

Using the *Collocations Dictionary of Modern Slovene* in the Process of Translating Collocations

The paper outlines the findings of the study on how the *Collocations Dictionary of Modern Slovene* (KSSS) is utilized when translating collocations from English into Slovene. Undergraduate students of the Department of Translation Studies at the Faculty of Arts in Ljubljana were requested to translate a text with ten selected collocations. During the translation process, their on-screen activities were recorded to allow for the analysis of both translation solutions, as well as the translation process, focusing on the use of language resources, in particular KSSS. The results have shown that the integration of KSSS into the training process is successful, as all the students participating in the study were familiar with this resource and actively use it in their translation work. However, the study has also exposed significant differences between students in terms of their familiarity with the advanced features of KSSS and, consequently, their success and efficiency in finding appropriate collocations. The results of the study have also highlighted that the use of language resources does not necessarily lead to an optimal translation solution.

1. Uvod

Kolokacije predstavljajo izjemno zanimiv jezikovni pojav, ki pa je hkrati tudi zelo izmuzljiv. Uvodoma povzemamo definicije Gantar et al. (2021), ki so osnova za vključitev kolokacij *v Kolokacijski slovar sodobne slovenščine* (KSSS), rabo katerega opisuje pričujoči prispevek (prim. Kosem et al., 2018). Pri definiranju kolokacij avtorji izpostavijo statistični, skladenjski in semantični vidik.

Atkins in Rundell kolokacije definirata kot "ponavljajoče se kombinacije besed, v katerih kaže določen leksikalni element (jedro) očitno tendenco sopojavljanja z drugim leksikalnim elementom (kolokatorjem), s frekvenco, ki je večja od naključne sopojavitve" (2008: 302). Gantar et al. pri tem izpostavijo problematiko kolokacij, katerih sestavni deli v besedilu navadno ne nastopajo skupaj oz. se mednje vrivajo drugi elementi, ki jih imenujejo "razširjene kolokacije" (2021: 19).

Poleg statističnega pa kolokacije definira tudi skladenjski vidik, saj med kolokacijskima elementoma obstaja hierarhično razmerje, v katerem baza določa kolokator (Hausmann, 1984: 401).

Gantar et al. (2021) predstavijo še tretji, najpomembnejši vidik, ki je hkrati tudi najbolj problematičen, in sicer pomenski vidik kolokacij. Ta je namreč tesno povezan s statističnim vidikom, ki kolokacije uvršča med pola, ki ju predstavljajo proste besedne zveze na eni in popolnoma ustaljene večbesedne enote na drugi strani, kar posledično vpliva na semantične spremembe in omejitve pri izbiri kolokacijskih elementov.

Pri definiranju kolokacij avtorji navajajo dva ključna pristopa, in sicer ožji, ki kolokacije prepoznava kot samostojen tip frazeoloških enot, ki so delno ali popolnoma (pomensko in skladenjsko) zamrznjene, in širši, ki med kolokacije šteje tudi frekventne besedne zveze, katerih notranja povezovalnost ni ozko zamejena ali celo izključujoča,

pač pa je lahko niz sopojavnic tudi razmeroma odprt (Gantar et al. 2021: 20-21).

Kljub temu, da kolokacije predstavljajo težavo, zlasti ko se z njimi srečamo v tujem jeziku, ker se pri razumevanju ne moremo zanašati na skladenjska ali pomenska razmerja maternega jezika, saj ni nujno, da uporabljana jezika določeno kolokacijo tvorita na enak način, pa so kolokacijski slovarji koristen pripomoček tudi za rojene govorce, zlasti ko gre za področje, ki jim ni najbolj domače. Dodana vrednost tovrstnih jezikovnih virov za prevajalce je v tem, da na enem mestu na pregleden način prikažejo širok nabor kolokacij, med katerimi je nato mogoče izbrati tisto, ki je tako s pomenskega kot tudi sobesedilnega vidika najustreznejša.

2. Namen raziskave in pregled področja

Zaradi svoje jezikovne in kulturne specifičnosti, so kolokacije izjemno zanimive z vidika prevajanja, s čimer so se raziskovalno ukvarjali tako domači (npr. Gabrovšek, 2014; Jurko 2014; Sicherl, 2004; Vrbinc, 2006) kot tuji avtorji (Kwong, 2020; McKeown in Radev, 2000 itd.). S prevajalskega vidika je pomembno, na kakšen način prevajalci rešujejo težave, ki se pojavljajo pri prevajanju kolokacij, saj le ustrezne prevajalske strategije lahko privedejo do ustreznih prevodnih rešitev. Pri iskanju možnih prevodnih rešitev in potrditvi prevodnih ustreznic si prevajalci lahko pomagajo z različnimi jezikovnimi viri. Eden od možnih raziskovalnih pristopov, ki nam lahko pomagajo osvetliti, kako prevajalci pridejo do ustreznih prevodnih rešitev in kakšne vire pri tem uporabljajo, so uporabniške študije, ki pa se seveda razlikujejo glede na potrebe ciljne skupine, (prim. Arhar Holdt et al., 2015; Arhar Holdt et al., 2016; Pori et al., 2020; Pori et al., 2021). Nekateri avtorji (Rozman, 2004; Stabej, 2009; Logar Berginc, 2009; Arhar Holdt, 2015) so pred časom sicer opozarjali na to, da v Sloveniji primanjkuje uporabniških študij, vendar pa se v zadnjem času na tem področju stanje spreminja. Izvedenih je bilo namreč kar nekaj uporabniških raziskav, in sicer tako med prevajalci, tolmači, študenti, učitelji slovenščine, lektorji in jezikoslovci, kot tudi drugimi, ki se poklicno ukvarjajo z jeziki (prim. Čibej et al., 2015; Gorjanc, 2014; Hirci, 2013; Pori et al., 2021; Mikolič, 2015; Šorli in Ledinek, 2017; Arhar Holdt et al., 2017).

V pričujočem prispevku je predstavljena uporabniška študija, ki se ukvarja z vprašanjem prevajanja kolokacij in rabo jezikovnih virov pri iskanju prevodnih rešitev. Prispevek predstavlja izsledke raziskave, ki je preučevala proces prevajanja kolokacij iz angleščine v slovenščino med študenti prevajalstva, s poudarkom na rabi KSSS v procesu prevajanja, s čimer se doslej raziskovalno ni ukvarjal še nihče.

Raziskovanje prevajalskega procesa ponuja dragocene informacije o strategijah, ki so potrebne, da se določeno besedilo prevede iz izvirnega v ciljni jezik. Poglobljen pregled ključnih raziskovalnih vprašanj, najpogosteje uporabljenih tehnologij in trendov razvoja tega področja ponuja Jakobsen (2017) (prim. Hansen, 2009; Hvelplund, 2019; idr.). V Sloveniji je področje preučevanja prevajalskega procesa manj raziskano (prim. Hirci, 2012).

Študenti na Oddelku za prevajalstvo Filozofske fakultete Univerze v Ljubljani se s KSSS seznanijo že v prvem letniku dodiplomskega študija. V raziskavi je sodelovalo 15 študentov in študentk, in sicer osem iz drugega (v nadaljevanju označeni z oznakami od II-1 do II-8) in sedem iz tretjega letnika (III-1 do III-7) dodiplomskega študija.

Prevajalska naloga je vsebovala članek s poljudnoznanstveno vsebino[1] in navodila za prevod. V 437 besed dolgem članku z astronomsko tematiko je bilo označenih 10 kolokacij. Kljub temu, da je bilo na voljo celotno besedilo, saj je pri prevajanju kolokacij sobesedilo ključnega pomena, so študenti morali prevesti zgolj tiste povedi, v katerih so bile označene kolokacije. Četudi označevanje kolokacij, ki so bile predmet analize, predstavlja odmik od avtentične prevajalske situacije, smo se za ta korak odločili, da bi ohranili konsistentnost analize, kar je bilo v luči majhnega števila sodelujočih pravzaprav nujno.

Da bi pridobili čim več podatkov o prevajalskem procesu, je delo potekalo na platformi Zoom, kjer so sodelujoči uporabili možnost deljenja zaslona, kar je omogočalo spremljanje procesa prevajalskega dela, celotno dogajanje na zaslonu pa se je za potrebe kasnejše analize tudi snemalo.

Prevajalska naloga je omogočila analizo z dveh različnih vidikov. Prvega predstavlja sam prevod oz. posamezne prevodne rešitve kolokacij in povedi, v katerih se nahajajo, drugega pa posnetek prevajalskega procesa, zlasti rabe jezikovnih virov, še posebej KSSS.

V prvem delu smo se pri analizi osredotočali na to, ali je prevod ustrezen (tj. ali prevodna rešitev predstavlja ustrezno, v ciljnem jeziku sprejemljivo kolokacijo in ali ta kolokacija tudi pomensko ustreza

[1] Povezava do članka: http://news.bbc.co.uk/2/hi/science/nature/1006305.stm

izvirniku). V primeru neprimernih oz. pogojno sprejemljivih rešitev smo pri vsaki kolokaciji dodali komentar vidika, ki se je zdel problematičen.
rešitve posamezne kolokacije je sledil pregled posnetka zaslona, pri čemer smo analizirali, na kakšen način so študenti uporabljali jezikovne vire, da bi prišli do prevodne ustreznice – katere, kako učinkovito (če je to iz posnetka razvidno) in kako so prišli do ugotovitve, da je določena rešitev najustreznejša.

Posnetek prevajalskega dela je omogočil vpogled v proces rabe jezikovnih virov. Pregledu prevodne

Kljub temu, da je prevajalska naloga vsebovala 10 kolokacij (Tabela 1), se v pričujočem prispevku zaradi prostorske omejitve osredotočamo zgolj na tri, ki pa ponujajo zanimiv vpogled v celotni razpon zahtevnosti, težav, iskanja prevodnih rešitev in načinov rabe KSSS ter drugih jezikovnih virov.

3. Rezultati

	Povedi z označeno kolokacijo
1	Astronomers say reports that the Earth could be struck by a small asteroid in 2030 are **wildly exaggerated**.
2, 3	Less than a day after **(2) sounding the alert** about asteroid 2000SG344, a **(3) revised analysis** of the space rock's orbit shows it will in fact miss the Earth by about five million kilometres.
4	Some scientists have criticised the way the **information was released** to the media before it had been thoroughly confirmed.
5	**Threat rating***
6	On Friday, the International Astronomical Union **issued an alert** saying that the object had about a 1-in-500 chance of striking the Earth on 21 September 2030.
7	Were it to strike our planet, the **results** would be **devastating**, with an explosion greater than the most powerful nuclear weapon.
8	The new orbit reveals **a slight risk** of a collision with the Earth about 2071, but it is thought that when the orbit is better known this risk will disappear as well.
9	If it is manmade and did strike Earth, the **effects** would be very **local** and **limited**.
10	Some scientists have criticised the IAU and Nasa for releasing warnings about the asteroid only for those **warnings** to be **rescinded** less than a day later.

*podnaslov

Tabela 1: Pregled povedi z označenimi kolokacijami.

Kolokacija št. 4 se je izkazala za najbolj problematično, saj so bile tri prevodne rešitve v celoti neustrezne, tri pa zgolj delno ustrezne. Kot v celoti neustrezne smo označili tiste, ki izkazujejo bodisi skladenjsko neustreznost v ciljnem jeziku bodisi gre za pomensko neustrezno prevodno rešitev, četudi je bila uporabljena slovenska kolokacija z visoko pogostostjo.

Kot zgolj delno ustrezne smo opredelili kolokacije, ki so se v prevodu pomensko preveč oddaljile od izvirnika ali pa je njihova skladenjska oblika netipična. V nadaljevanju so prikazane vse prevodne rešitve, ki so bile označene kot neustrezne oz. delno ustrezne, in jezikovni viri, ki so jih študenti pri iskanju prevodnih rešitev uporabili.

	Kolokacija št. 4	Viri
	[…] the way the information was released […]	
II-1	skritizirali način, kako je bila informacija […] deljena z mediji	brez virov
II-5	skritizirali način izdaje podatkov medijem	angleški kolokacijski slovar ozdic.com
II-7	kritizirali način, da so novico objavili	spletni an-sl slovar Pons, KSSS
III-2	način, ki je bil uporabljen za posredovanje informacij javnosti	veliki an-sl slovar (Amebisov pregledovalnik podatkovnih zbirk ASP32), KSSS
III-3	način, na katerega so bile informacije sporočene medijem	Evrokorpus, KSSS
III-5	dejstvo, da so mediji objavili informacijo	korpus Gigafida

Tabela 2: Prikaz iskanja prevodnih rešitev za kolokacijo št. 4.

Kolokaciji št. 2 in 6 predstavljamo skupaj, saj sta si tako pomensko kot skladenjsko zelo podobni, a je prva študentom povzročala precej težav, prevodne rešitve za drugo pa so bile z izjemo dveh v celoti ustrezne.

V nadaljevanju so prikazane vse prevodne rešitve, ki so bile označene kot neustrezne oz. delno ustrezne, in jezikovni viri, ki so jih študenti uporabili.

	Kolokacija št. 2	Viri	Kolokacija št. 6	Viri
	[…] after sounding the alert about […]		**[…] issued an alert saying […]**	
II-1	dan po tem [sic] ko so sprožili alarm	KSSS	izdala opozorilo	brez virov
II-3	dan po sproženem alarmu	an-sl spletni slovar Pons, KSSS	izdala opozorilo	brskalnik Google, enojezična spletna slovarja Collins in Cambridge, KSSS
II-6	dan po sprožitvi alarma	an-sl spletni slovar Pons, enojezični spletni slovar Merriam-Webster, KSSS	izdala opozorilo	KSSS
II-8	dan po sprožitvi alarma	an-sl spletni slovar Pons, portal Fran, KSSS, korpus Gigafida	izdala opozorilo	an-sl spletni slovar Pons, KSSS, korpus Gigafida
III-2	dan potem [sic] ko so sprožili alarm	brez virov	sprožila alarm	brskalnik Google, portal Fran, KSSS, korpus Gigafida
III-3	dan po sproženem alarmu	brskalnik Google, enojezični spletni slovar Merriam-Webster, KSSS	sprožil alarm	Evrokorpus, EUR+Lex, KSSS

Tabela 3: Prikaz iskanja prevodnih rešitev za kolokaciji št. 2 in 6.

4. Diskusija

4.1. Kolokacija št. 4

Z vidika prevajanja se je za najbolj zahtevno izkazala kolokacija št. 4 (“release information”). Pri tem je treba izpostaviti, da je že sam izvirnik nekoliko problematičen, saj je samostalniška besedna zveza ”the way” v funkciji premega predmeta, ki sledi glagolu ”criticise”, najpogosteje uporabljena v smislu “kritizirati način, na katerega /…/”. Vendar pa je v analiziranem primeru mišljeno drugače: znanstveniki so kritizirali dejstvo, da je bila ta informacija sploh posredovana medijem, ne pa načina, kako je bilo to storjeno. Zdi se, da so študenti, katerih prevodne rešitve so bile označene kot neustrezne oz. zgolj delno ustrezne, izvirnik obravnavali preveč dobesedno, pri čemer niso upoštevali širšega konteksta, ki bi jim omogočil pravilno interpretacijo, čeprav so imeli na voljo celotno besedilo.

II-1: Četudi je kolokacija glagola “deliti” in samostalnika “informacija” ustaljena in v danem kontekstu tudi ustrezna, se redkeje pojavlja v trpni obliki. Glede na dejstvo, da študent ni uporabil nobenih jezikovnih virov, lahko zgolj domnevamo, da bi se sicer na podlagi zgledov v tvornem načinu morda odločil za drugačno rešitev.

II-5: Prevodna rešitev “način izdaje podatkov medijem” je problematična z vidika dobesednosti, obenem pa je tudi kolokacijsko vprašljiva, saj se glagolnik “izdaja” s samostalnikom v rodilniku večinoma pojavlja v pomenu izida tiskanega dela (npr. knjige, revije itd.) ali dajanja (delnic, denarja itd.) v obtok. Študentka je pri iskanju prevodne rešitve uporabila angleški kolokacijski slovar ozdic.com z iskalnima nizoma “information” in “released information”, vendar pa temu ni sledila korekcija izbrane prevodne rešitve.

II-7: Pri prevodu “kritizirali način, da so novico objavili” gre za pomensko napako, ki v veliki meri temelji na dobesednem razumevanju izvirnika. Čeprav prevodna rešitev “kritizirati način” s kolokacijskega vidika ni sporna, pa je bolj problematičen odvisnik, ki sledi. Samostalniku “način“, kadar ta sledi glagolu “kritizirati”, navadno sledi prilastkov odvisnik in ne predmetni. Študent je uporabil dva spletna jezikovna vira, in sicer angleško-slovenski slovar Pons za iskanje ustreznic besed “release”, “some” in “thoroughly”, v KSSS pa je iskal kolokacije s samostalnikom “novica”, pri čemer ni uporabil nobenih filtrov.

III-2: Prevodna rešitev “način, ki je bil uporabljen za posredovanje informacij javnosti” je kolokacijsko ustrezna, vendar se pomensko oddalji od izvirnika. Študentka je uporabila dva jezikovna vira – veliki angleško-slovenski slovar (Amebisov pregledovalnik podatkovnih zbirk ASP32) za glagol “criticise” in KSSS za samostalnik “informacija”, pri čemer ni uporabila nobenih filtrov.

III-3: Prevodna rešitev "način, na katerega so bile informacije sporočene medijem" je v smislu pomenskega odklona zelo podobna primeru III-2, in je kolokacijsko ustrezna, slogovno je vprašljiva zgolj trpna oblika. Študentka je kot vir uporabila Evrokorpus z iskalnima nizoma "released information" in "information released" ter KSSS z iskalnim nizom "izdaja podatkov", za katerega pa ni bilo zadetkov. Če bi namesto tega izbrala zgolj glagolnik "izdaja", bi z izbiro filtra "s samostalniki/2-rodilnik" lahko hitro prišla do ustrezne prevodne rešitve.

III-5: Prevodna rešitev "dejstvo, da so mediji objavili informacijo" je slogovno in kolokacijsko ustrezna, pomensko pa se tudi ta oddalji od izvirnika, saj so po tej interpretaciji predmet kritike mediji. Izvirnik v resnici govori o tem, da so znanstveniki kritizirali dejstvo, da so mediji te podatke sploh dobili oz. so posredno kritizirali svoje cehovske kolege, ne pa samih medijev. Študentka je pri iskanju prevodne rešitve uporabila zgolj korpus Gigafida z iskalnim nizom "objaviti informacijo".

4.2. Kolokaciji št. 2 in št. 6

Pri kolokaciji "sound the alert" je treba izpostaviti, da je v angleščini precej bolj pogosta pomensko podobna kolokacija "sound the alarm". To je najverjetneje tudi razlog, da so se študenti v kar šestih primerih v prevodu namesto za "opozorilo" odločili za samostalnik "alarm", saj so v jezikovnih virih bodisi našli prevodno ustreznico ("alert"→"alarm") bodisi dobili potrditev, da je angleška različica "sound the alarm" bolj pogosta. Od tu naprej so v jezikovnih virih iskali zgolj kolokacije za samostalnik "alarm". Vsi so izbrali sicer pravilno, a zelo dobesedno kolokacijo "sprožiti alarm".

II-1: Študent se je odločil za prevodno rešitev "dan po tem ko so sprožili alarm", pri čemer je kot vir uporabil KSSS, kjer je najprej iskal kolokacije za glagol "sprožiti", nato pa še za samostalnik "alarm". Tu je tudi našel kolokacijo "sprožiti alarm". Pri kolokaciji št. 6 se je odločil za drugačno prevodno rešitev, in sicer "izdati opozorilo", pri čemer pa ni uporabil nobenih jezikovnih virov.

II-3: Študentka je do prevodne rešitve "dan po sproženem alarmu" prišla s pomočjo angleško-slovenskega spletnega slovarja Pons z iskalnim nizom "sounding". Tu je našla kolokacijo "to sound the alarm", čemur je sledilo iskanje v KSSS, in sicer "alarm", kjer je našla kolokacijo "sprožiti alarm". Pri kolokaciji št. 6 se je tudi ona odločila za drugačno prevodno rešitev, in sicer "izdati opozorilo", do katere je prišla s pomočjo brskalnika Google (iskalni niz "issue an alert"); sledila je povezavi do spletnega enojezičnega slovarja Collins, ker pa tam ni našla tega niza, je isto kolokacijo vpisala še v spletni slovar Cambridge. Ker tudi tam ni našla zadetka, se je vrnila na brskalnik Google. Iz dogajanja na zaslonu ni razvidno zakaj, vendar je, ne da bi kliknila na katerega izmed zadetkov, uporabila KSSS z iskalnim nizom "opozorilo", kjer je nato kljub temu, da ni uporabila nobenega filtra, na prvi strani najpogostejših kolokacij našla "izdati opozorilo". Tej ugotovitvi pa ni sledila korekcija kolokacije št. 2.

II-6: Tudi študentka II-6 se je odločila za podobno prevodno rešitev, in sicer "dan po sprožitvi alarma". Prvi jezikovni vir je bil angleško-slovenski slovar Pons za iskalni niz "alert", kjer je našla prevodno ustreznico "alarm". Temu je sledilo iskanje po enojezičnem slovarju Merriam-Webster z geslom "sound", temu pa iskanje kolokacij za samostalnik "alarm" v KSSS. Tu je pregledala štiri kolokacije v sobesedilu, pri čemer so bili trije zadetki za "sprožiti lažni alarm" in en zadetek za "sprožiti požarni alarm", kar pa je ni omajalo v prepričanju, da je to prava rešitev. Ne glede na to odločitev se je pri kolokaciji št. 6 tudi ona odločila za drugačno prevodno rešitev, in sicer "izdati opozorilo", do katere je prišla neposredno s pomočjo KSSS z iskalnim nizom "opozorilo". Tudi v tem primeru tej ugotovitvi ni sledila korekcija kolokacije št. 2.

II-8: Prevodna rešitev "dan po sprožitvi alarma" je identična kot pri študentki II-6, podobni so tudi jezikovni viri, ki jih je uporabljala pri iskanju prevodnih rešitev. Prvi je bil angleško-slovenski slovar Pons za iskalni niz "alert", čemur je sledilo iskanje na jezikovnem portalu Fran z geslom "alarm". Nato je v KSSS iskala kolokacije za samostalnik "alarm", pri čemer ni uporabila nobenega filtra. Temu je sledilo preverjanje kolokacije "sprožiti alarm" v korpusu Gigafida z iskalnim nizom "alarm", kjer se je med konkordancami zadržala pri omenjeni kolokaciji, ki pa jo je nato v prevodni rešitvi spremenila v "sprožitev alarma". Pri kolokaciji št. 6 se je odločila za drugačno prevodno rešitev, in sicer za kolokacijo "izdati opozorilo". Najprej je uporabila angleško-slovenski slovar Pons ("issue" in "alert"), čemur je sledilo iskanje kolokacij za samostalnik "opozorilo" v KSSS (najprej brez filtrov, nato s filtrom "z glagoli", vendar pri pregledovanju zadetkov ni vztrajala). Sledilo je iskanje po korpusu Gigafida ("opozorilo"), kjer pa med konkordancami ni našla nobenega zadetka, ki bi se ji zdel primeren. Po tem je ponovno uporabila KSSS, kjer je med rezultati (spet brez filtra) našla končno prevodno rešitev "izdati opozorilo".

III-2: Do tako slovnično kot vsebinsko problematične prevodne rešitve "dan potem [sic] ko so sprožili alarm" je študentka prišla brez rabe jezikovnih virov. Morda je to tudi razlog, da se je za enako prevodno rešitev odločila pri kolokaciji št. 6, kjer pa je uporabila kar nekaj jezikovnih virov. Svoje iskanje je

začela v brskalniku Google z iskalnim nizom "izdati rdeč [sic] alarm", pregledala nekaj zadetkov (večinoma časopisnih naslovov) in nadaljevala iskanje na portalu Fran z iskalnim nizom "alarm". Isto geslo je nato uporabila še pri KSSS. Sledilo je iskanje po korpusu Gigafida (na portalu CJVT) z naprednimi funkcijami "okolica" in "levo 1", "desno 0". Med besednimi vrstami na levi je nato uporabila filter "glagol", kjer je po pogostosti izstopala kolokacija, ki jo je študentka nato izbrala kot prevodno rešitev. Omenjeni primer uporabe naprednih funkcij se zdi zelo poveden, saj dokazuje, da zgolj tehnična podkovanost pri rabi jezikovnih virov ne zagotavlja nujno tudi ustrezne prevodne rešitve. Čeprav omogoča časovno učinkovitost pri iskanju možnih kolokacij, mora prevajalec še vedno sprejeti končno odločitev o tem, katera izmed ponujenih možnosti predstavlja najustreznejšo prevodno rešitev, pri čemer vedno igra pomembno vlogo tudi prefinjen občutek za jezik.

III-3: Tudi ta študentka se je pri kolokaciji št. 2 odločila za prevodno rešitev "dan po sproženem alarmu". Začela je z brskalnikom Google (iskalni niz "souding [sic] the alert"), kjer je izbrala povezavo do spletnega slovarja Merriam-Webster za "raise/sound the alarm". Nato je v KSSS iskala kolokacijo za samostalnik "alarm", med rezultati opazila kolokacijo "sprožen alarm" in jo tudi uporabila. Pri iskanju prevodne rešitve kolokacije št. 6 je najprej uporabila Evrokorpus ("issue an alert"), od koder je sledila predlagani povezavi na EUR+Lex za isti iskalni niz, nazadnje pa je v KSSS ponovno iskala kolokacije za samostalnik "alarm", opazila že znano kolokacijo "sprožiti alarm" in jo nato tudi uporabila.

Na koncu je treba poudariti, da je v predstavljeni uporabniški raziskavi šlo za vodeno nalogo, ki realno situacijo prevajanja tovrstnega besedila nekoliko popači, zato je pri posploševanju opažanj potrebna previdnost.

Razlogi za to so poleg omejenega vzorca, tako z vidika števila sodelujočih kot tudi dejstva, da so v raziskavi sodelovali zgolj študenti dveh letnikov dodiplomskega študija, vsaj trije. Prvi je ta, da so bili študenti seznanjeni s tem, da se raziskava ukvarja z rabo jezikovnih virov, kar je morda vplivalo na to, katere vire so uporabljali in kako. Drugi razlog je, da so vedeli, da je poudarek naloge na prevajanju kolokacij, zaradi česar je mogoče, da so se posledično na to prvino bolj osredotočali. Tretji razlog pa je ta, da se je za potrebe kasnejše analize proces prevajanja snemal, zaradi česar se študenti bolj zavedajo vsake svoje poteze. Ali je – in če, do kakšne mere – to v dani situaciji vplivalo na sam proces prevajanja in končni izdelek je težko soditi, vsekakor pa je pri interpretaciji rezultatov in izpeljevanju zaključkov treba imeti v mislih tudi ta vidik.

Zlasti pri prevodnih strategijah kolokacije št. 4 se je pokazalo, da jezikovni viri, kljub temu, da prevajalcu lahko pomagajo priti do prevodnih rešitev, ki so parcialno pravilne, ne morejo vedno preprečiti napačne interpretacije izvirnika in posledično neposrečenih prevodnih rešitev, ki so rezultat združevanja besed, besednih zvez in kolokacij, ki so same zase ustrezne, kot celota pa preprosto ne funkcionirajo. Pri prepoznavanju in reševanju tovrstnih situacij zagotovo ključno vlogo odigra tudi dobro znanje materinščine.

Izpostaviti je treba tudi dejstvo, da se je od študentov pričakovalo, da bodo prevedli povedi, v katerih so bile kolokacije vnaprej označene. Vendar pa je povsem mogoče predvideti tudi scenarij, kjer bi se prevajalec odločil za prevodno rešitev, v kateri izvirna kolokacija ne bi bila prevedena v obliki kolokacije, a bi bila s pomenskega in sobesedilnega vidika rešitev prav tako ustrezna.

Kljub temu, da je KSSS enojezični jezikovni vir, se je v procesu prevajanja izkazal za zelo uporabnega. Pri tem se zdi, da ne gre toliko za možnost preverjanja, ali določena besedna kombinacija sploh predstavlja kolokacijo, temveč bolj za vpogled v širši nabor možnih kolokacij, ki jih KSSS ponuja, izmed katerih nato prevajalec lahko izbere tisto, ki je pomensko in sobesedilno najprimernejša.

Četudi se študenti prevajalstva že v prvem letniku dodiplomskega študija seznanijo s KSSS in ga pri prevajalskih nalogah tudi uporabljajo, se je izkazalo, da njegovega ustroja ne poznajo vsi enako dobro. Posledično se niti ne zavedajo možnosti, ki jih nudi, zato je v nekaterih primerih njihovo iskanje ustreznih kolokacij manj učinkovito, kar v skrajnem primeru lahko privede tudi do tega, da kljub pravilno vnesenemu iskalnemu nizu ne najdejo ustrezne kolokacije.

Ena izmed možnosti, ki jih slovar ponuja, je napredno iskanje s pomočjo filtrov, kjer je mogoče izbrati ustrezno kategorijo kolokatorja (npr. samostalnik, pridevnik, glagol itd.), v nekaterih primerih pa tudi njegovo podkategorijo (npr. sklon samostalnika, sklon pridevnika, predlog itd.). Pred študenti, ki te funkcije niso poznali oz. je niso uporabili, se je v primerih iskanja na podlagi baze, ki tvori kolokacije s številnimi kolokatorji, tako odprl nepregleden seznam kolokacij, deljen glede na različne kolokacijske vzorce, večina katerih je bila v danem primeru neuporabnih. Študenti so s pregledovanjem seznama po nepotrebnem izgubljali čas, ki bi ga sicer lahko bolje izkoristili.

5. Zaključek

Uporabniška raziskava med študenti prevajalstva o rabi jezikovnih virov pri iskanju prevodnih rešitev

kolokacij ali kolokacijskih parov je postregla s številnimi zanimivimi izsledki in omogočila vpogled v praktično rabo jezikovnih virov v prevajalskem procesu.

Obenem se je izkazalo, da so nekateri študenti tehnično dobro podkovani in vešči uporabe naprednih funkcij, ki jih posamezni jezikovni viri nudijo, a se to ne odraža vedno tudi v kakovosti njihovih prevodnih rešitev. Pri tem ne gre nujno za jezikovno šibkost v maternem jeziku, kot morda bolj za prekomerno zaupanje v jezikovni vir, pri čemer ne vzamejo v obzir možnih razlik med jezikovno materijo, ki jo prevajajo, in primeri, ki jih ponujajo jezikovni viri.

Raziskava je tako izpostavila tudi zelo konkretno pomanjkljivost, ki bi jo v pedagoškem procesu v prihodnje veljalo bolje nasloviti. Med raziskavo so se prav tako odprla številna vprašanja, povezana z rabo jezikovnih virov v procesu prevajanja, ki bi jih bilo smiselno nasloviti v prihodnjih raziskavah.

Raziskovalni program št. P6-0215 (Slovenski jezik – bazične, kontrastivne in aplikativne raziskave) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

6. Literatura

Špela Arhar Holdt. 2015. Uporabniške raziskave za potrebe slovenskega slovaropisja: prvi koraki. V: V. Gorjanc et al., ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 136-148. Ljubljana: Znanstvena založba Filozofske fakultete.

Špela Arhar Holdt, Jaka Čibej in Ana Zwitter Vitez. 2017. Value of language-related questions and comments in digital media for lexicographical user research. *International journal of lexicography*, 30(3), str. 285–308. Oxford: OUP.

Špela Arhar Holdt, Iztok Kosem, in Polona Gantar. 2016. Dictionary user typology: the Slovenian case. V: T. Margalitadze in G. Meladze, ur., *Lexicography and linguistic diversity. Proceedings of the XVII EURALEX International Congress, 6–10 September, 2016*, str. 179–187. Tbilisi: Ivane Javakhishvili Tbilisi State University.

Beryl T. Sue Atkins in Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

Jaka Čibej, Vojko Gorjanc in Damjan Popič. 2015. Vloga jezikovnih vprašanj prevajalcev pri načrtovanju novega enojezičnega slovarja. V: V. Gorjanc et al., ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 168-181. Ljubljana: Znanstvena založba Filozofske fakultete.

Dušan Gabrovšek. 2014. Extending Binary Collocations: (Lexicographical) Implications of Going beyond the Prototypical a – b. *ELOPE: 11*(2), str. 7–20. Ljubljana: Slovensko društvo za angleške študije.

Polona Gantar, Simon Krek in Iztok Kosem. 2021. Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. V: I. Kosem, ur., *Kolokacije v slovenščini*, str. 15–41. Ljubljana: Znanstvena založba Filozofske fakultete.

Vojko Gorjanc. 2014. Slovar slovenskega jezika v digitalni dobi. Irena Grahek in Simona Bergoč, ur., *E-zbornik Posveta o novem slovarju slovenskega jezika na Ministrstvu za kulturo*. Ljubljana: Ministrstvo za kulturo RS.

Gyde Hansen. 2009. Some thoughts about the evaluation of translation products in translation process research. *Copenhagen Studies in Language 38*, str. 389–402. Copenhagen: Samfundslitteratur.

Franz Josef Hausmann. 1984. Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. V: *Praxis des neusprachlichen Unterrichts*, 31, str. 395–406. Dortmund: Lensing.

Nataša Hirci. 2012. Electronic Reference Resources for Translators. *The Interpreter and Translator Trainer (6)* 2, str. 219–36. London: Taylor & Francis.

Nataša Hirci. 2013. Changing trends in the use of translation resources: the case of trainee translators in Slovenia. *ELOPE* 10, str. 149–165. Ljubljana: Slovensko društvo za angleške študije.

Kristian Tangsgaard Hvelplund. 2019. Digital resources in the translation process – attention, cognitive effort and processing flow. *Perspectives 27* (4), str. 510–24. London: Taylor & Francis.

Arnt Lykke Jakobsen. 2017. Translation process research. V John W. Schwieter in Aline Ferreira, ur., *The handbook of translation and cognition*, str. 19–49. Hoboken: Wiley.

Primož Jurko. 2014. Target language corpus as an encoding tool: collocations in Slovene-English translator training. *ELOPE 11* (1), str. 153–64. Ljubljana: Slovensko društvo za angleške študije

Iztok Kosem, Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej in Cyprian Laskowski. 2018. Kolokacijski slovar sodobne slovenščine. V: D. Fišer in Andrej Pančur, ur., *Zbornik konference Jezikovne tehnologije in digitalna humanistika / Proceedings of the conference on Language Technologies & Digital Humanities*, 20-21, str. 133-139. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani.

Nataša Logar Berginc. 2009. Slovenski splošni in terminološki slovarji: za koga? V: M. Stabej, ur., *Infrastruktura slovenščine in slovenistike*. Obdobja 28, str. 225–231. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Oi Yee Kwong. 2020. Translating Collocations: The Need for Task-driven Word Associations. V: *Proceedings of the Workshop on the Cognitive Aspects*

of the Lexicon, str. 112–16. Association for Computational Linguistics.

Kathleen R. McKeown in Dragomir R. Radev. 2000. Collocations. V Robert Dale et al., ur., *Handbook of Natural Language Processing,* str. 1–3. New York: Marcel Dekker.

Vesna Mikolič. 2015. Slovarski uporabniki – ustvarjalci: ustvarjati v jeziku in z jezikom. V: V. Gorjanc et al., ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 182-195. Ljubljana: Znanstvena založba Filozofske fakultete.

Eva Pori, Jaka Čibej, Iztok Kosem in Špela Arhar Holdt. 2020. The attitude of dictionary users towards automatically extracted collocation data: A user study. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research 8* (2), str. 168-201.

Eva Pori, Iztok Kosem, Jaka Čibej in Špela Arhar Holdt. 2021. Evalvacija uporabniškega vmesnika Kolokacijskega slovarja sodobne slovenščine. V I. Kosem, ur., *Kolokacije v slovenščini*, str. 235–268. Ljubljana: Znanstvena založba Filozofske fakultete.

Tadeja Rozman. 2004. Upoštevanje ciljnih uporabnikov pri izdelavi enojezičnega slovarja za tujce. *Jezik in slovstvo* 49 (3/4), str. 63-75. Ljubljana: Slavistično društvo Slovenije.

Eva Sicherl. 2004. On the Content of Prepositions in Prepositional Collocations. *ELOPE 1*(1-2), str. 37–46. Ljubljana: Slovensko društvo za angleške študije

Marko Stabej. 2009. Slovarji in govorci: kot pes in mačka? *Jezik in slovstvo* 54 (3−4), str. 115–138. Ljubljana: Slavistično društvo Slovenije.

Mojca Šorli in Nina Ledinek. 2017. Language policy in Slovenia: language users' needs with a special focus on lexicography and translation tools. V: I. Kosem et al., ur., *Electronic lexicography in the 21st century: proceedings of eLex 2017 Conference, 19–21 September 2017, Leiden, The Netherlands,* str. 377–394. Brno: Lexical Computing.

Marjeta Vrbinc. 2005. Native speakers of Slovene and their translation of collocations from Slovene into English: a Slovene-English empirical study. *Erfurt Electronic Studies in English*. Erfurt: Institut für Anglistik/Amerikanistik Erfurt.

Akustično modeliranje z različnimi osnovnimi enotami za avtomatsko razpoznavanje slovenskega govora

Lucija Gril,* Simon Dobrišek, ‡ Andrej Žgank*

* Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška cesta 46, 2000 Maribor
lucija.gril@um.si, andrej.zgank@um.si
‡ Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1000 Ljubljana
simon.dobrisek@fe.uni-lj.si

Povzetek

V članku je predstavljen sistem avtomatskega razpoznavanja govora za slovenski jezik. Za graditev akustičnih modelov smo uporabili dva različna jezikovna vira in dve različni osnovni akustični enoti pri zapisu slovarjev. Testiranje je potekalo na testni množici, ki je nastala znotraj projekta Razvoj slovenščine v digitalnem okolju in vsebuje malo manj kot 5 ur zvočnih posnetkov. Za graditev jezikovnih modelov smo uporabili hibridni pristop HMM-DNN. Za nevronske mreže smo uporabili dva tipa mrež, in sicer TDNN in LSTM. Najboljši rezultat WER je znašal 24,95 % in smo ga dosegli z arhitekturo TDNN in grafemskim slovarjem.

Acoustic modeling with various basic units for Slovenian automatic speech recognition

The article presents the automatic speech recognition system for the Slovenian language. We used two different language sources and lexicons based on two basic acoustic units. The system was tested by the test set containing a little less than 5 hours of sound recordings that developed by the RSDO project. We used the hybrid HMM-DNN approach to build language models. Two types of networks were used for neural networks, namely TDNN and LSTM. The best WER score was 24.95% and we achieved it with TDNN architecture and grapheme lexicon.

1. Uvod

Dandanes nas pametna okolja spremljajo že na vsakem koraku. Pametni telefoni, tablice, televizijski sprejemniki, ročne ure, naprave v gospodinjstvu itd. Vse te naprave so nagnjene k temu, da nam nudijo boljšo in preprostejšo uporabniško izkušnjo. Storitev, ki jih nudijo, je veliko in za vse je potrebno skrbno načrtovanje tako strojne kot tudi programske opreme. Ena izmed takšnih storitev je tudi avtomatsko razpoznavanje govora (angl. Automatic speech recognition – ASR). Če želimo razpoznavati govor, se je treba zavedati, da lahko programska oprema deluje brezhibno, vendar na uspešnost njenega delovanja vpliva še veliko drugih dejavnikov. Eden izmed njih je lahko na primer slab mikrofon, ki zajame veliko šuma in popači zvok ter tako degradira razpoznavanje govora. To posledično vodi tudi do slabše uporabniške izkušnje. Prav tako lahko do poslabšanja rezultatov pride, če je razpoznavalnik tekočega govora slabše zasnovan in nima optimalnih karakteristik. Zato je pomembno, da z eksperimenti preverjamo različne arhitekture in zasnove modelov avtomatskega razpoznavalnika govora.

Za razvoj razpoznavalnika govora potrebujemo veliko količino učnega gradiva za posamezen jezik. Za jezike z veliko govorci je takšnega dosegljivega gradiva praviloma veliko. Za jezike z manjšim številom govorcev, kamor lahko uvrščamo tudi slovenščino, pa takšnih virov ni dovolj za uporabo naprednih metod umetne inteligence, kot je na primer enovito učenje (ang. end to end) s konvolucijskimi mrežami. V zadnjem obdobju se pogosto uporablja tudi učenje s prenosom znanja, vendar za obe našteti metodi velja, da omogočata slabši nadzor nad modeliranjem v primerjavi s hibridnim pristopom, ki smo ga uporabili v tem članku. Praviloma hibridni pristop tudi dosega nekoliko boljše rezultate, kot pa druga dva pristopa. Za avtomatski razpoznavalnik govora potrebujemo govorne posnetke, ki jih spremljajo datoteke s transkripcijo, v katerih je zapis izgovorjenih besed. Hkrati potrebujemo besedilni korpus in slovar, s katerima se lahko razpoznavalnik govora nauči značilnosti besed in tudi njihovega kontekstnega uvrščanja.

Izgovorjene besede lahko v avtomatskem razpoznavalniku govora predstavimo z dvema različnima akustičnima enotama – s fonemi ali z grafemi. Fonemi so glasovne enote, ki predstavljajo izgovorjavo glasov v besedi. Fonemski zapis slovenske izgovorjave se v večini primerov razlikuje od grafemskega. Grafemi in fonemi se med seboj razlikujejo tudi v številu osnovnih enot. Grafem zapišemo z eno osnovno enoto, ki ustreza črki v besedi. Po drugi strani se lahko ista črka slika v več različnih fonemov, odvisno od konteksta, naglasa in mesta v besedi. Prav tako se lahko črka preslika v zaporedje dveh fonemov. Z mislijo na to lahko pri razpoznavalniku tvorimo slovarje, ki vsebujejo izgovorjene besede na dva načina, in sicer s fonemi ali grafemi. Izbira vrste slovarja razpoznavalnika govora neposredno vpliva na to, kakšna bo osnovna akustična enota. Izbira akustične enote vpliva tudi na zahtevnost in način priprave slovarjev, kompleksnost akustičnih modelov in prek tega na potreben pomnilnik in procesorske zmogljivosti za učenje in delovanje avtomatskega razpoznavalnika govora. Tvorjenje slovarja s fonemi je odvisno od jezika, ki ga želimo uporabiti. Za slovenski jezik je ta naloga razmeroma zapletena in kompleksa. Slovar se lahko tvori ročno, kar praviloma počnejo fonetiki ali slovenisti, ali avtomatsko. Pri avtomatskih postopkih pa se lahko zgodi, da je zapis besede fonetično napačen, kar se v kasnejših korakih odraža na neoptimalnem učenju in razpoznavanju govora. Priprava slovarja z grafemi je lažja, saj je pretvorba trivialna. Kateri pristop je primernejši, je odvisno tudi od količine učnih podatkov, ki jih uporabljamo, saj je pri slovarjih s fonemi, ki so sestavljeni iz več osnovnih enot, večji poudarek na

številski porazdelitvi glede na kategorijo. V okviru projekta Razvoj slovenščine v digitalnem okolju (RSDO, b. d.) trenutno vzporedno poteka graditev govorne baze in pa razvoj prvih verzij avtomatskega razpoznavalnika govora. Zato imamo trenutno še vedno na voljo dokaj omejeno količino transkribiranega slovenskega govora, kar je bil povod za uporabo grafemske akustične enote. Že v preteklosti se je tako za slovenski jezik (Žgank in Kačič, 2006) kot tudi za druge jezike (Killer et at., 2003) pokazalo, da je lahko v takšnih primerih uporaba grafemskih akustičnih enot dobra rešitev. Tako smo za cilj članka postavili primerjavo med fonemskimi in grafemskimi akustičnimi osnovnimi enotami v povezavi s trenutno razpoložljivimi govornimi viri.

V nadaljevanju članka najprej pregledamo, kaj je na področju modeliranja akustičnih osnovnih enot avtomatskega razpoznavanja govora že bilo izvedenega za slovenski jezik. V tretjem poglavju predstavimo, katere govorne in jezikovne vire smo uporabili pri zasnovi eksperimentov. Tvorjenje slovarjev in samodejno grafemsko-fonemsko pretvorbo na osnovi pravil predstavimo v četrtem poglavju. V petem poglavju predstavimo modeliranje akustičnih in jezikovnih modelov avtomatskega razpoznavalnika govora. Rezultati so predstavljeni in komentirani v šestem poglavju, ki mu sledi še zaključek.

2. Pregled sorodnih člankov

Avtomatski razpoznavalniki govora so kot svojo privzeto akustično osnovno enoto uporabljali foneme in njihove izpeljanke v obliki kontekstnega podaljševanja. Izhodišče je bilo, da gre pri avtomatskem razpoznavanju govora za pretvorbo iz govorjene v besedilno obliko, kar se tako sklada z izbiro osnovne akustične enote. Leta 2000 so Schillo in sodelavci predstavili prvi grafemski avtomatski razpoznavalnik govora, ki je z izbiro drugačne osnovne akustične enote kršil navedeno predpostavko. Sistem je za nemški jezik sicer dosegel slabše rezultate razpoznavanja govora kot fonemski sistem, vendar so bili naučeni grafemski modeli manjši.

Grafemi kot osnovne akustične enote postanejo hitro zanimivi tudi za večjezično in križnojezično razpoznavanje govora (Killer et at., 2003). V takšnih primerih je namreč možno združevati jezike brez podrobnega poznavanja fonetike vključenih jezikov. Osnovo pač predpostavlja zapisana črka. Uporabnost takšnega pristopa pride še dodatno do izraza pri križnojezičnem razpoznavanju govora, kjer so v ciljnem jeziku na voljo omejeni govorni viri. Uspešnost metode je v določeni meri odvisna tudi akustično-fonetične podobnosti med vključenimi jeziki.

Prve raziskave o uporabi grafemov kot osnovne akustične enote za križnojezično razpoznavanje slovenskega govora so predstavili Žgank in sodelavci (2005). Sledila je še uporaba grafemov za običajno enojezično avtomatsko razpoznavanje govora (Žgank in Kačič, 2006). Grafemi kot osnovne akustične enote so tako postali del standardne izbire za razpoznavanje slovenskega govora, še posebej v domeni dnevnoinformativnih oddaj (Gril et al., 2021). V kombinaciji s slovenskimi razpoznavalniki govora, ki so zasnovani na HMM akustičnih modelih ali na hibridni zasnovi HMM/DNN in imajo za učenje na voljo nekaj 10 ur transkribiranih govornih posnetkov, praviloma dosežejo boljše rezultate razpoznavanja govora. Predpostavimo sicer lahko, da se bo ta razlika manjšala, ko bo za slovenski jezik na voljo več ur transkribiranega govora. Z večanjem količine posnetkov namreč pridobimo na posamezno osnovno enoto tudi več vzorcev, kar izboljša možnost modeliranja akustičnih značilnosti in izboljša robustnost na potencialne napake, ki se lahko zgodijo zaradi avtomatske grafemsko-fonemske pretvorbe.

3. Govorni in jezikovni viri

Govorni in jezikovni viri so pri razpoznavalnikih govora ključna komponenta. Za govorne posnetke smo uporabili korpuse Gos 1.0 (Zwitter Vitez et al., 2013), Sofes (Dobrišek et al., 2017) in delovno različico testnega seta nastajajoče govorne baze RSDO (trenutna delovna različica je 2.0, ki ne vsebuje več črkovanja). Korpusa Gos in Sofes smo uporabili za učno in razvojno množico, medtem ko smo testni korpus 2.0 projekta RSDO uporabili za vrednotenje rezultatov. Za slovarje smo uporabili prostodostopni vir Sloleks 2.0 (Dobrovoljc et al., 2019) in trenutno verzijo slovarja, ki je nastala v projektu RSDO. Za besedilni korpus smo uporabili prostodostopni besedilni vir ccGigafida 1.0 (Logar et al., 2013).

Korpus Gos vsebuje 120 ur posnetkov. Posnetki zajemajo različne zvrsti, npr. televizijske oddaje, predavanja, pouk, zasebne pogovore … Ves govor je transkribiran v dveh različicah, in sicer v pogovorni in standardizirani različici. Posnetki zajemajo 1526 različnih govorcev. Govorni korpus Sofes vsebuje 9 ur in 52 minut posnetkov, ki zajemajo govor 134 različnih govorcev. Posnetki vsebujejo poizvedovanja po letalskih informacijah v slovenskem jeziku. Pri korpusu Sofes najdemo transkripcije v fonetičnem zapisu in standardiziranem zapisu govora. V testnem setu 2.0 RSDO je za 4 ure in 47 minut posnetkov. Korpus se od različice 1.0 razlikuje po tem, da ne vsebuje posnetkov črkovanja, kar znaša okoli 19 minut govora. Črkovanje smo iz splošnega testnega nabora izločili, saj je za njegovo učinkovito razpoznavanje treba uporabiti drugačne pristope. Testna množica RSDO zajema bran, javni, nejavni govor in posnetke državnega zbora. V posnetkih se pojavi 63 različnih govorcev. Tudi pri korpusu RSDO imamo dva različna zapisa govora, ki sta nastala na osnovi enakih priporočil kot v korpusu Gos.

Vir Sloleks 2.0 je leksikon, ki vsebuje okoli 2.792.000 posameznih besednih oblik. Vsak vnos vsebuje podatke o besedi (v katero besedno vrsto sodi in kakšne so njene slovnične lastnosti). Zapisane so tudi vse pregibne oblike za posamezno besedo. Slovenščina je pregiben jezik in zato je takšnih oblik zelo veliko. V različici 2.0 je označeno tudi mesto naglasa in zapis v mednarodni fonetični pisavi (IPA).

V našem primeru smo Sloleks 2.0 uporabili za tvorjenje fonetičnega slovarja avtomatskega razpoznavalnika govora. V takšnem slovarju potrebujemo besede in njihovo izgovorjavo s fonemi. Sloleks 2.0 smo s pomočjo postopka, ki so ga uporabili Ulčar in drugi (2019), pretvorili v obliko, ki je ustrezna za avtomatski razpoznavalnik govora.

Besedilni korpus CcGigafida vsebuje nekaj čez 103.000.000 besed in je javno dostopni del korpusa Gigafida, ki ga je možno uporabljati pod licenco Creative Commons. Besedilo vsebuje informacije o virih časopisov, revij, leta izdaj, vrsti besedil, naslovih, o avtorjih besedil. Korpus je označen z morfoskladenjskimi opisi in lemami.

Besedilni korpus ccGigafida smo uporabili za jezikovno modeliranje avtomatskega razpoznavalnika

govora. Zaradi pravilne obdelave smo iz korpusa izbrisali prazne vrstice in večkratne presledke. Odstranili smo tudi ločila, da je bilo besedilo v skladu z običajno obliko v sistemu za razpoznavanje govora.

4. Tvorjenje slovarjev za razpoznavalnik govora

Tvorjenje fonetičnih slovarjev, ki so potrebni za graditev hibridnih arhitektur avtomatskih razpoznavalnikov govora, temelji tako na uporabi obstoječih razpoložljivih leksikonov, ki so navadno ročno preverjeni in že vsebujejo fonetične prepise besed, kot tudi na uporabi samodejnih grafemsko-fonemskih pretvornikov, ki se uporabljajo za t. i. izvenslovarske besede, ki jih predvideva jezikovni model razpoznavalnika govora, niso pa še vključene v obstoječe leksikone.

Tvorjenje slovarja za prvo različico avtomatskega razpoznavalnika govora (»Rezultat R2.2.7: Orodje za grafemsko fonemsko pretvorbo – verzija 2«, 2022) , ki je bil razvit v okviru projekta RSDO in bo predstavljen v naslednjih poglavjih, je primarno temeljilo na uporabi že omenjenega leksikona Sloleks 2.0 ter ročno urejenega in preverjenega slovarja izgovorjav, ki je vključen v govorni korpus Sofes. Za vse besede, ki se pojavljajo v normiranih besednih prepisih vseh zvočnih govornih posnetkov, ki so se uporabili za tvorjenje akustičnega modela razpoznavalnika govora, ter za vse besede, ki se pojavljajo v normiranem besedilnem korpusu, ki se je uporabil za tvorjenje njegovega jezikovnega modela, smo najprej pogledali v leksikon Sloleks 2.0 in ročno urejen slovar govornega korpusa Sofes, če ta morda vsebujeta obravnavano besedo. Če je bila beseda v tem leksikonu oziroma slovarju vsebovana, se je njen fonetični prepis samo prenesel v slovar razpoznavalnika govora. Če obravnavana beseda v leksikonu Sloleks 2.0 oziroma slovarju Sofes ni bila vsebovana, pa se je njen fonetični prepis pridobilo z uporabo prve različice samodejnega grafemsko-fonemskega pretvornika, ki je bil razvit v okviru projekta RSDO in je v grobem opisan v nadaljevanju. Pri tvorjenju slovarja za predstavljeni razpoznavalnik govora se je samodejno moralo pretvoriti kar več kot 58 odstotkov vseh besed, ki so bile predvidene za razpoznavalnik govora. Pravilnost samodejne pretvorbe pa pri prvi različici razpoznavalnika govora še ni bila natančno preverjena in ovrednotena.

4.1. Samodejna grafemsko-fonemska pretvorba na osnovi pravil

Prva različica samodejnega grafemsko-fonemskega pretvornika, ki je bil razvit v okviru projekta RSDO in se je uporabil za tvorjenje slovarja razpoznavalnika govora, je temeljila na uporabi množice kontekstno odvisnih fonetičnih pravil, ki so bila določena na osnovi statističnih analiz in obstoječega jezikoslovnega in glasoslovnega poznavanja fonetičnih značilnosti slovenskega govorjenega jezika. Upoštevana kontekstno odvisna pravila so temeljila predvsem na upoštevanju mesta naglasa v danih besedah.

Mesto naglasa v besedi na splošno določa zlog, na katerem ima beseda jakostno ali tonsko izraženo slušno zaznavno izrazitost (Toporišič, 1992). Pomembna značilnost slovenskega jezika je, da se mesto naglasa pojavlja na prvem, zadnjem, predzadnjem ali tudi predpredzadnjem zlogu. Poleg tega pa lahko imajo posamezne besede tudi več mest naglasa. Mesto naglasa je določeno za vsako besedo posebej in se ga je med različnimi generacijami govorcev slovenskega govorjenega jezika zgodovinsko prenašalo z učenjem jezika in govornim sporazumevanjem. Kljub različnim mestom naglasa, ki so se z razvojem jezika in v različnih narečnih jezikovnih skupinah tudi spreminjala, pa je vendarle možno opredeliti določena pravila, ki v pretežni meri določajo mesto naglasa v besedah (Toporišič, 1991). Ta pravila so bila v glavnem upoštevana in uporabljena za samodejno določanje mesta naglasa v danih besedah. Pravila temeljijo na upoštevanju seznamov predpon, pripon, začetnic in končnic, ki se pojavljajo v slovenskih besedah in značilno določajo mesto naglasa v dani besedi. Pravila so bil določena na podoben način, kot je bilo to izvedeno pri razvoju sistema za samodejno tvorjenje umetnega slovenskega govora (Gros, 1997).

Uporabljena pravila sicer ne pokrijejo vseh trenutno uporabljanih slovenskih besed. Zato se je na osnovi dodatne statistične analize mest naglasov pri najbolj pogostih slovenskih besedah določilo še dodatna pravila za določitev najbolj verjetnega mesta naglasa v danih besedah. Ta pristop je do določene mere možno tolmačiti tudi kot izvajanje strojnega učenja iz podatkov.

Grafemski zapisi vhodnih besed se v razvitem pretvorniku z uporabo pravil pretvarjajo po vrsti, od leve proti desni. Pravila se v pretvorniku preverjajo in upoštevajo po danem vrstnem redu, zato si morajo slediti tako, da so na začetku seznama pravil za posamezen grafem najprej tista, ki opisujejo posebne primere pretvorb, sledijo pa jim bolj splošna pravila.

Razviti grafemsko-fonemski pretvornik na svojem vhodu predvideva besede, ki so že podane v normirani polni besedni obliki. Števila, števniki, denarne enote, okrajšave in drugi posebni zapisi morajo tako biti podani v polni besedni obliki. Za to je bilo poskrbljeno z normalizacijo besednih prepisov govornih posnetkov, ki so se uporabljali za tvorjenje akustičnega modela razpoznavalnika govora, in tudi besedil iz besedilnega korpusa, ki so se uporabljala za tvorjenje jezikovnega modela razpoznavalnika govora.

Izhodni nabor fonemskih različic je glede na samodejno določanje in upoštevanje mesta naglasa omogočal tudi ločevanje med dolgimi in kratkimi samoglasniki. Pri tvorjenju slovarja za razpoznavalnik govora pa se to ločevanje ni upoštevalo, ker se pri tvorjenju akustičnih modelov razpoznavalnikov govora samoglasnikov navadno ne ločuje na kratke in dolge, ker dolžina samoglasnikov nima osnovne pomensko razločevalne vloge pri razpoznavanju besed (Ulčar, 2019).

5. Arhitektura avtomatskega razpoznavalnika govora

Glede na razpoložljivo količino akustičnega učnega materiala, je bilo smiselno uporabiti hibridno arhitekturo avtomatskega razpoznavalnika govora, ki je v takšnih primerih praviloma učinkovitejša, kot so pa pristopi E2E.

Pri hibridnih sistemih avtomatskega razpoznavalnika govora lahko arhitekturno sestavo grobo razdelimo na dva dela, in sicer na akustični model in jezikovni model. Akustični model naučimo na osnovi vzorcev iz zvočnih posnetkov in njihovih ustreznih prepisov, jezikovni model pa glede na besedilni korpus. V nadaljevanju članka bomo podrobneje predstavili oba modela, za graditev katerih smo uporabili prostodostopno orodje Kaldi (Povey et al., 2011).

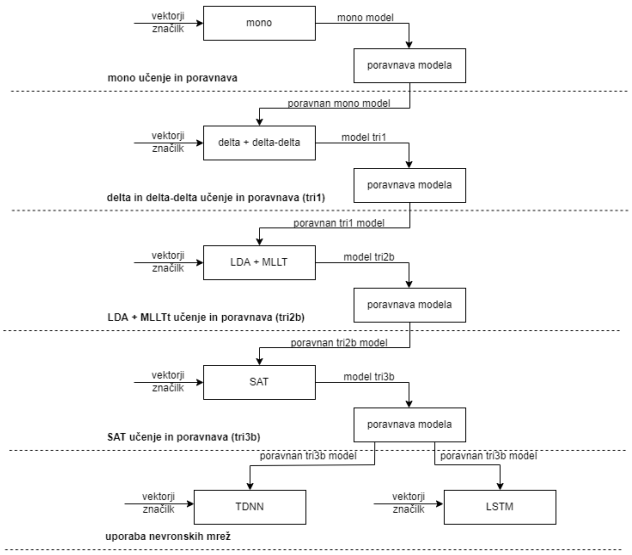
Za pripravo spremljajočih datotek, ki jih potrebujemo za graditev modela v orodju Kaldi, smo uporabili transkripcije govornih korpusov, ki so zapisane v obliki standardiziranega zapisa govora.

5.1. Akustično modeliranje

Za akustično modeliranje smo uporabili govorne baze Gos, Sofes in testno množico projekta RSDO. Zvočni posnetki govornih baz Gos in Sofes so bili v mono formatu in so bili zapisani v 16-bitnem zapisu. Frekvenca vzorčenja je bila 16 kHz. Posnetki testne množice projekta RSDO so imeli frekvenco vzorčenja 44,1 kHz, bitna hitrost in format pa sta bila enaka posnetkom v bazah Gos in Sofes. Orodje Kaldi za svoje delo potrebuje mono zvočne posnetke s frekvenco vzorčenja 16 kHz in 16-bitnim zapisom. Da lahko posnetke v orodju Kaldi procesiramo, moramo posnetke pretvoriti v ustrezni format. S prostodostopnim orodjem SoX smo posnetke pretvorili v mono zvočne posnetke, s frekvenco vzorčenja 16 kHz in 16-bitnim zapisom. Pretvarjanje posnetkov smo vključili v proces priprave potrebnih datotek za procesiranje v orodju Kaldi. S tem korakom smo se ognili ročnemu pretvarjanju vseh posnetkov.

Zvočne posnetke, ki so del govorne baze, smo spremenili v vektorje značilk. Na začetku posnetke razdelimo na okna dolžine 25 ms in jih nato transformiramo, da dobimo značilke MFCC (mel-frekvenčne kepstralne koeficiente). Za nadaljnje delo smo uporabili 12 MFCC koeficientov in energijo, nad katerimi smo izračunali še prvi in drugi časovni odvod. S prvim odvodom dobimo delta in z drugim delta-delta značilke. Nadaljevali smo z akustičnim modeliranjem, kjer smo v več fazah izvajali učenje novih akustičnih modelov in njihove poravnave.

Osnova akustičnega modeliranja avtomatskega razpoznavalnika govora so prikriti modeli Markova (angl. Hidden Markov Model – HMM). Z modeli HMM na osnovi vhodnih vektorjev značilk ocenjujemo verjetnosti hipotez izgovorjenega govora. Za to moramo poznati zapis fonemov v vsaki besedi. Takšne zapise imamo vnesene v fonetičnem slovarju, kjer je vsaka beseda predstavljena z nizom fonemov izgovorjene besede. Pri tem imamo lahko za posamezno besedo na voljo več izgovorjav, kar je odvisno od vključenega slovarja. Pri HMM modelih foneme predstavimo s skritimi stanji, model pa nato izračuna opazovana stanja s pomočjo Gaussovih porazdelitev, ki tvorijo hipoteze izgovorjene besede.



Slika 1: Postopek učenja akustičnega modela avtomatskega razpoznavalnika govora.

V naslednji fazi smo uporabili linearno diskriminančno analizo (angl. Linear discriminant analysis – LDA), s katero poiščemo linearno kombinacijo stanj. LDA vzame vektorje značilk in zgradi HMM stanja, vendar z manjšim prostorom značilke za vse podatke. LDA smo uporabili v kombinaciji z linearno transformacijo z največjo verjetnostjo (angl. Maximum Likelihood Linear Transform – MLLT), ki poenostavi računanje Gaussovih porazdelitev (Gales, 1999). MLLT vzame značilke iz LDA in izpelje edinstveno transformacijo za vsakega govorca. MLLT je prvi korak k normalizaciji govorcev, saj minimalizira razlike med govorci. Pri LDA in MLLT se uporabi prvih 13 značilk MFCC in vsako razdeli na 4 predhodna okna na levi in 4 naslednja okna na desni. To pomeni, da imamo končno dimenzijo značilk 117. Nato z LDA dimenzijo značilke omejimo na 40.

Za večjo natančnost avtomatskega razpoznavanja govora smo uporabili učenje s prilagajanjem govorcu (angl. Speaker Adaptive Training – SAT), ki za vsakega posameznega govorca izračuna parametre adaptacij glede na učne podatke tega govorca (Anastasakos et at., 1996).

Učenje akustičnega modela smo začeli z monofonskim akustičnim modelom in nadaljevali s trifonskim akustičnim modelom z delta in delta-delta (tri1) značilkami, trifonskim akustičnim modelom z LDA in MLLT (tri2b) ter na koncu še s trifonskimi akustičnimi modeli s SAT (tri3b). Postopek učenja je prikazan tudi s pomočjo diagrama, ki ga lahko vidimo na sliki 1.

V drugem delu graditve akustičnih modelov sledi prehod na globoke nevronske mreže. Nevronske mreže so sistemi, kjer algoritmi posnemajo delovanje nevronov v možganih. Sistem je sestavljen iz vhodnih, skritih in izhodnih plasti, ki so sestavljene iz enega ali več nevronov. Nevroni so med seboj povezani z relacijami, ki lahko potekajo naprej, nazaj ali naprej in nazaj. Na relacijah se uporabljajo uteži, s katerimi se izračunajo nova stanja.

Uporabili smo dva različna tipa nevronskih mrež, in sicer časovno zakasnjene nevronske mreže (angl. Time Delayed Neural Networks – TDNN) in nevronske mreže z

dolgim kratkoročnim spominom (angl. Long Short Term Memory – LSTM).

TDNN so nevronske mreže (Waibel, 1989), ki imajo več plasti. Začetne plasti se transformacije učijo bolj ozko, kasnejše pa imajo širši časovni kontekst. Za kontekstno modeliranje je treba zagotoviti, da vsaka nevronska celica poleg vhodne vrednosti, ki jo pridobi od aktivacijske funkcije oziroma iz nižje plasti, pridobi tudi informacijo o vzorcu izhodnih vrednosti in njihovega konteksta. Kar v primeru s časovnim signalom pomeni, da dobi vsaka nevronska celica na vhod informacijo o aktivacijskem vzorcu skozi čas od nižje ležečih plasti.

Nevronske mreže LSTM (Povey, 2018) vključujejo spominsko celico, ki ohrani informacijo dalj časa. Celica ima troje različnih vrat, in sicer vhodna, izhodna ter pozabljiva. Vhodna vrata uravnavajo količino podatkov prejšnjega vzorca, ki se bo shranila. Izhodna vrata določajo količino podatkov, ki se bo prenesla na naslednjo plast. Pozabljiva vrata pa regulirajo hitrost izgubljanja informacij v celici. Zaradi shranjevanja informacij so sistemi LSTM primerni za delo s časovnimi signali, saj se lahko pomembni dogodki zamaknejo. Modelu LSTM lahko rečemo tudi izboljšana ponavljajoča se nevronska mreža (angl. Recurrent Neural Network – RNN), saj je bila tako odpravljena težava izginjajočega gradienta (Hochreiter, 1991).

Arhitektura TDNN je sestavljena iz vhodnega nivoja, skritih nivojev in izhodnega nivoja. Vhodni nivo je dimenzije 40. Prvi skriti nivo mreže TDNN je bila mreža LDA z dimenzijo 40 in je bila polno povezana. Mreži LDA je sledilo še 8 polno povezanih mrež TDNN dimenzij 512. Na 8 nivojih mrež TDNN je bilo uporabljeno izpuščanje nevronov (angl. dropout). Mrežam TDNN sledita še dve vzporedni veji nivojev, in sicer verižna veja in veja xent. Verižna in xent veji sta sestavljeni iz dveh nivojev. Prva vzporedna nivoja tvorita mreži ReLU dimenzije 512. Mreži sta polno povezani in enako kakor mreže TDNN uporabljata izpuščanje nevronov. Mrežama ReLU sledita izhodna nivoja. Veji se razlikujeta po funkciji izgube. Verižna veja uporablja funkcijo logaritem verjetnosti pravilne sekvence fonemov oziroma grafemov, medtem ko veja xent za funkcijo izgube uporablja križno entropijo. Mreža TDNN je tako sestavljena iz 10 nivojev, pri katerih pa smo uporabili tudi časovno združevanje, kjer se na teh nivojih združijo informacije iz želenih časovnih oken glede na vhod. Časovno združevanje smo uporabili na nivoju LDA in 2., 4., 6., 7. ter 8. nivoju TDNN.

Učenje modelov TDNN je potekalo 7 epoh. Začetno efektivno stopnjo učenja (angl. initial effective lrate) smo nastavili na 0,0001 in končno (angl. final effective lrate) na 0,00001. Ostale vrednosti parametrov smo ohranili na privzetih vrednostih.

Tako kot arhitektura TDNN tudi LSTM vsebuje tri vrste nivojev. Prvi je vhodni in je enak vhodnemu nivoju arhitekture TDNN. Prav tako je tudi prvi skriti nivo arhitekture LSTM enak nivoju LSA, ki sestavlja arhitekturo TDNN. Naslednji štirje skriti nivoji so mreže LSTMP (angl. Long Short-Term Memory Projection) velikosti 1024. LSTMP je mreža LSTM, ki dodatno vsebuje še projekcijski nivo. V naši konfiguraciji arhitekture smo dimenzijo projekcijskega nivoja nastavili na 256. Skritim nivojem sledita dve veji izhodnih nivojev. Tudi tukaj se veji razlikujeta glede na funkcijo izgube tako kot pri arhitekturi TDNN.

Akustične modele LSTM za razpoznavanje govora smo učili s 4 epohami. Ostale vrednosti smo ohranili na privzetih, vključno z začetno in končno efektivno stopnjo učenja, ki sta bili nastavljeni na 0,001 in 0,0001.

V naslednjem poglavju bomo predstavili rezultate sistemov LSTM in TDNN za razpoznavanje govora. Ker je sistem TDNN dosegel boljše rezultate, smo del eksperimentov opazovali samo na sistemu TDNN.

5.2. Jezikovno modeliranje

Kot povezovalni člen med akustičnim in jezikovnim prostorom smo uporabili dva različna tipa slovarjev avtomatskega razpoznavalnika govora. Prvi tip uporabljenih slovarjev je bil fonemski slovar, kjer so besede zapisane s fonemi, in drugi tip, kjer smo namesto zapisa izgovorjene besede s fonemi uporabili zapis z grafemi. V tabeli 1 smo predstavili lastnosti slovarjev. Ena izmed lastnosti je tudi delež besede izven slovarja (angl. out of vocabulary – OOV), ki ga izračunamo kot:

$$OOV = \frac{\text{št. besed izven slovarja v testni množici}}{\text{št. vseh besed v slovarju}} \cdot 100 \quad (2)$$

Slovarji, ki smo jih uporabili, so večji, kakor tisti, ki so se uporabljali v prejšnjih razpoznavalnikih informativnih oddaj (Gril et at., 2021). Vrednosti OOV so zelo nizke in jih lahko enostavno zanemarimo.

Slovar	Tip slovarja	Št. besed	OOV [%]
Sloleks 2.0	fonemski	1.129.144	0,054
Sloleks 2.0	grafemski	931.848	0,065
RSDO	fonemski	1.440.070	0,008
RSDO	grafemski	1.440.070	0,008

Tabela 1: Lastnosti uporabljenih slovarjev.

Jezikovni model avtomatskega razpoznavalnika govora naučimo z besedilnim korpusom. Takšen model je sposoben predvidevati besedo, ki sledi, glede na predhodne besede v nizu. Jezikovni model ima tudi zmožnost kontekstnega uvrščanja, saj bo med besedami, ki imajo podobno izgovorjavo, izbral tisto, ki bo bolj smiselna glede na kontekst predhodno opazovanega zaporedja besed.

Jezikovni model smo naučili z uporabo orodja n-gram count, ki je del paketa SRILM (Stolcke, 2002). N-grami so v našem primeru nizi *n* besed v stavku. N-gram count glede na besedilni korpus generira n-grame in z njimi ocenjuje napovedne verjetnosti jezikovnega modela. Pri n-gram countu je treba določiti, do kakšne velikosti n-gramov želimo zgraditi model. Tako smo zgradili jezikovni model, ki je vseboval 1-grame, 2-grame in 3-grame.

6. Rezultati avtomatskega razpoznavanja govora

Uspešnost različnih verzij avtomatskega razpoznavalnika govora smo ovrednotili na testni množici 2.0 projekta RSDO. Za vrednotenje smo uporabili delež napačno razpoznanih besed (angl. Word Error Rate – WER). WER smo izračunali kot razmerje med številom vrinjenih, izbrisanih ter zamenjanih besed in med številom besed, ki so v referenčnem besedilu. To lahko zapišemo kot:

$$WER = \frac{(I + D + S)}{N} \cdot 100 \quad (1)$$

Kjer je *I* število vrinjenih besed (angl. insertions), *D* število izbrisanih besed (angl. deletions) in *S* število zamenjanih besed (angl. substitutions). Z *N* označimo število vseh besed v referenčnem besedilu testne množice. Razmerje nato pomnožimo s 100, saj WER praviloma podajamo v odstotkih.

Arhitektura	Slovar	Tip slovarja	WER [%]
LSTM	Sloleks 2.0	fonemski	38,70
TDNN	Sloleks 2.0	fonemski	27,19
TDNN	RSDO	fonemski	25,31
TDNN	Sloleks 2.0	grafemski	26,97
TDNN	RSDO	grafemski	24,95

Tabela 2: Rezultati razpoznavanja govora z različnimi vrstami vključenih metod in modelov.

Najprej poglejmo rezultate, ki smo jih dobili, ko smo vrednotili različna tipa arhitektur akustičnih modelov. Predstavljeni so v tabeli 2. Sistem LSTM se je izkazal za slabšega, saj je bil rezultat WER kar za 11,51 % slabši kot v primeru, ko smo uporabili sistem TDNN. Na osnovi tega rezultata smo kot nadaljnjo arhitekturo akustičnih modelov izbrali TDNN. Izhodiščni WER je bil 27,19 %. Ulčar in drugi (2019) so na podobnem sistemu dosegli malo slabši rezultat, vendar rezultati niso neposredno primerljivi, saj se je vrednotenje preverjalo na drugi testni množici. Primerjava s predhodnim podobnim ASR (Gril et at., 2021) kaže razliko v rezultatih. Avtorji so takrat dosegli 15,17 % WER, vendar z uporabo drugačnih govornih virov. Domena virov je bila v prejšnjem primeru omejena izključno na televizijske oddaje, res pa je, da so te lahko v nekaterih primerih, kot je na primer glasbeno ozadje govora, tudi dokaj kompleksne za avtomatsko razpoznavanje govora.

Za nadaljevanje razvoja sistema za razpoznavanje govora smo uporabili dva različna slovarja, in sicer slovar, ki je bil narejen na osnovi Sloleksa, in slovar, ki je bil pripravljen v sklopu projekta RSDO. V tabeli 2 lahko vidimo, da se rezultat vrednotenja z uporabo slovarja, ki je bil pripravljen pri projektu RSDO, izboljša za 1,88 %.

V zadnjem koraku smo primerjali med seboj še avtomatske razpoznavalnike govora, pri katerih smo z uporabo različnih tipov slovarja fonemsko osnovno akustično enoto zamenjali z grafemsko. Za avtomatski razpoznavalnik govora, pri katerem smo uporabili za osnovo Sloleks, je zamenjava fonemov z grafemi izboljšala rezultat za 0,22 %. Pri uporabi slovarja, ki je bil izdelan v okviru projektu RSDO, pa je zamenjava fonemov z grafemi WER izboljšala za 0,36 %. Rezultat s tem modelom in enotami je hkrati najboljši rezultat razpoznavanja govora, ki smo ga dosegli s predstavljenimi eksperimenti. Rezultat z grafemi je verjetno boljši zaradi omejene količine učnih podatkov in s tem tudi števila vzorcev na posamezno akustično enoto. Sklepamo lahko, da je teh bilo premalo za razpoznavo specifičnih akustičnih enot, ki so redkejše. Tako je razpoznavanje z grafemi, ki imajo manj akustičnih osnovnih enot, ker ne razlikujejo podvariant, delovalo bolje. Čeprav izboljšanje z grafemskim slovarjem ni posebej veliko, lahko pri tem tipu slovarja opozorimo na to, da je postopek priprave veliko preprostejši. Prednosti ima tudi pri uporabi, saj po velikosti zasede nekoliko manj pomnilniškega prostora, kar je posebej pomembno pri avtomatskih razpoznavalnikih govora z velikimi slovarji (angl. Large-Vocabulary Continuous Speech Recognition – LVCSR), kjer pri velikih datotekah hitro nastane ozko grlo. Dodatna prednost grafemskih akustičnih enot je tudi v tem, da lahko v praktični uporabi slovar avtomatskega razpoznavalnika govora nadgrajuje tudi laik.

7. Zaključek

V članku smo predstavili sistem za razpoznavo slovenskega govora. Za akustični model smo uporabili hibridni pristop HMM-DNN. Za napovedovanje skritih stanj v HMM smo uporabili dva tipa nevronskih mrež. Časovno zakasnjene nevronske mreže so se izkazale za boljši pristop kakor nevronske mreže z dolgim kratkoročnim spominom. Za tvorjenje slovarja smo uporabili dve osnovni akustični enoti. Grafemski modeli so v našem primeru dali boljše rezultate kakor fonemski. Uporabili smo novo testno množico, ki je nastala pri projektu RSDO. Najboljši delež napačno razpoznanih besed je bil 24,95 %. Rezultat je primerljiv tudi z rezultati drugih sistemov razpoznavanja govora. K dobremu rezultatu razpoznave prispeva velik slovar, ki je večji kakor pri primerljivih sistemih, in uporaba grafemov kot osnovne akustične enote. Sistemi z grafemi omogočajo enostavnejše tvorjenje slovarjev, enostavnejše je tudi nadgrajevanje takšnih slovarjev. Uporaba grafemov ima pozitivni učinek tudi pri uporabi sistemov, saj takšni modeli zavzemajo nekoliko manj pomnilniškega prostora.

Zahvala

Zahvaljujemo se avtorjem korpusa Gos 1.0, ki so nam omogočili njegovo uporabo za razvoj avtomatskega razpoznavalnika govora.

Raziskovalno delo je bilo delno opravljeno v okviru projekta RSDO – Razvoj slovenščine v digitalnem okolju. Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020.

8. Literatura

Tasos Anastasakos, John McDonough, Richard Schwartz in John Makhoul. 1996. A compact model for speaker-adaptive training. V: Proceedings ICSLP, str. 113–1140.

Simon Dobrišek, Jerneja Žganec Gros, Janez Žibert, France Mihelič in Nikola Pavešić. 2017. Speech Database of Spoken Flight Information Enquiries SOFES 1.0. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1125

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Jaka Čibej, Luka Krsnik in Marko Robnik-Šikonja. 2019. Morphological lexicon Sloleks 2.0. Slovenian language resource repository CLARIN.SI, ISSN 2820–4042, http://hdl.handle.net/11356/1230

Mark J. Gales. 1999. Semi-tied covariance matrices for hidden Markov models. *IEEE transactions on speech and audio processing*, 7(3): 272–281.

Jerneja Gros. 1997. *Samodejno tvorjenje govora iz besedil*. Doktorska disertacija. Fakulteta za elektrotehniko, Univerza v Ljubljani.

Sepp Hochreiter. 1991. Untersuchungen zu dynamischen neuronalen Netzen. Dostopno na:

https://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf (16. 5. 2022)

Mirjam Killer, Sebastian Stüker and Tanja Schultz. 2003. Grapheme based speech recognition. *Interspeech.*

Nataša Logar, Tomaž Erjavec, Simon Krek, Miha Grčar in Peter Holozan. 2013. Written corpus ccGigafida 1.0. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1035

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motliček, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer in Karel Vesely. 2011. The Kaldi speech recognition toolkit. V: *IEEE ASRU 2011 Workshop on automatic speech recognition and understanding. IEEE Signal Processing Society*.

Daniel Povey, Hossein Hadian, Pegah Ghahremani, Ke Li in Sanjeev Khudanpur, 2018. A Time-Restricted Self-Attention Layer for ASR. V: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), str. 5*874–5878.

RSDO. (b. d.). Dostopno na: https://www.cjvt.si/rsdo/.

Razvoj slovenščine v digitalnem okolju – RSDO: Rezultat R2.2.7: Orodje za grafemsko fonemsko pretvorbo – verzija 2, Poročilo projekta, 2022.

Christoph Schillo, Gernot A. Fink in Franz Kummert. 2000. Grapheme based speech recognition for large vocabularies. *Sixth International Conference on Spoken Language Processing*.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. V: *Seventh international conference on spoken language processing*.

Jože Toporišič. 1992. *Enciklopedija slovenskega jezika*. Cankarjeva založba. Ljubljana.

Jože Toporišič. 1991. *Slovenska slovnica*. Založba Obzorja. Maribor.

Matej Ulčar, Simon Dobrišek, Marko Robnik Šikonja. 2019. Razpoznavanje slovenskega govora z metodami globokih nevronskih mrež. *Uporabna informatika*, 27 (3). Dostopno na: https://uporabna-informatika.si/index.php/ui/article/view/53 (8. 11. 2021)

Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano and Kevin J. Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3): 328–339.

Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej in Tomaž Erjavec. 2013. Spoken corpus Gos 1.0. Slovenian language resource repository CLARIN.SI, ISSN 2820–4042, http://hdl.handle.net/11356/1040

Andrej Žgank, Zdravko Kačič, Frank Diehl, Jožef Juhar, Slavomir Lihan, Klara Vicsi in Gyorgy Szaszak. 2005. Graphemes as Basic Units for CrosslingualSpeech Recognition. V: *COST278 Final Workshop and ITRW on Applied Spoken Language Interaction in Distributed Environments*.

Andrej Žgank in Zdravko Kačič. 2006. Conversion from phoneme based to grapheme based acoustic models for speech recognition. *Interspeech*.

What works for Slovenian? A comparative study of different keyword extraction systems

Boshko Koloski, Senja Pollak, Matej Martinc

Jožef Stefan Institute, Jožef Stefan International Postgraduate School
Jamova cesta 39, Ljubljana, Slovenia
{boshko.koloski,senja.pollak,matej.martinc}@ijs.si

Abstract

Identifying and retrieving keywords from a given document is one of the fundamental problems of natural language processing. In this paper, we conduct a thorough comparative analysis of several distinct approaches for keyword identification on a new benchmark Slovenian keyword extraction corpus, SentiNews. The first group of methods is based on a supervised methodology, where previously annotated data is required for the models to learn. We evaluate two such approaches, *TNT-KID* and *BERT* . The other paradigm relies on unsupervised approaches, where no previously annotated data for training is needed. We evaluate five different unsupervised approaches, covering three main types of unsupervised systems: statistical, graph-based and embedding-based. The results show that supervised models perform significantly better than unsupervised approaches. By applying the *TNT-KID* method on the Slovenian corpus for the first time, we also advance the state-of-the-art on the SentiNews corpus.

1. Introduction

Identifying and retrieving keywords from a given document represents one of the crucial tasks for organization of textual resources. It is employed extensively in media organizations with large daily article production that needs to be categorized in a fast and efficient manner. While some media houses use keywords to link articles and produce networks based on keywords, journalists use keywords to search for news stories related to newly produced articles and also to summarize new articles with a handful of words. Manual categorization and tagging of these articles is a burdensome and time demanding task, therefore development of algorithms capable of tackling keyword extraction automatically, and therefore allowing the journalists to spend more time on more important investigative assignments, has become a necessity.

The approaches for automatic detection of keywords can be divided based on their need for annotated data prior to learning. One paradigm of keyword extraction focuses on extracting keywords without prior training (i.e. unsupervised approaches), while the other focuses on learning to identify keyphrases from an annotated data-set (i.e. supervised approaches). While unsupervised approaches can be easily applied for domains and languages that have low to no amount of labeled data, they nevertheless tend to offer non-competitive performance when compared to supervised approaches (Martinc et al., 2020), since they can not be adapted to the specific language and domain through training. On the other hand, supervised state-of-the-art approaches based on the transformer architecture (Vaswani et al., 2017) have become very effective in solving the task, but they do usually require substantial amounts of labeled data which is hard to obtain for some low-resource domains and languages.

In this research, we focus on one of the low-resource languages, Slovenian, for which not a lot of manually labeled data that could be leveraged for training of keyword extractors, is available. We systematically evaluate several distinct strategies for keyword extraction on Slovenian, among them also some, which have not been tested before on Slovenian. We show that the employment of the TNT-KID model (Martinc et al., 2020), a model specifically adapted for the monolingual low-resource scenario, leads to advance in state-of-the-art on the Slovenian SentiNews keyword extraction benchmark dataset (Bučar, 2017). To summarize, the main contributions of this work include:

- A systematical analysis of a keyword extraction dataset of Slovenian news.
- Thorough comparison of several supervised and unsupervised keyword extraction strategies on the Slovenian data-set. Supervised methods include the monolingual TNT-KID method, which has not been employed for Slovenian before, and an application of the multilingual BERT model (Devlin et al., 2019), same as in Koloski et al. (2022b). We also cover several unsupervised methods in this study, including statistical, graph-based and embedding based models.
- The advancement in state-of-the-art on the Slovenian keyword extraction dataset from SentiNews
- Release of a dockerized pretrained model of the best performing system TNT-KID-Slovene in terms of F1-score.

The paper is organized in the following manner: Section 2. describes the related work in the field, followed by the description of data and the exploratory data analysis in Section 3. Section 4. describes the experimental setting considered in this study and in Section 5., we discuss the results. Finally, Section 6. presents the conclusions of the study and proposes further work.

2. Related work

Keyword extraction approaches are either supervised or unsupervised.

2.1. Unsupervised methods

Modern supervised learning approaches are very successful in keyword extraction, but they are data intensive and time consuming. Unsupervised keyword detectors can address both problems and typically require much less computational resources and no training data, but this comes with the price of lower overall performance. Unsupervised methods can be divided into four main categories:

- *statistical* - methods that belong to this family are based on calculating various text statistics to capture keywords, such as frequency of appearance, position in the text, *etc*. KPMiner (El-Beltagy and Rafea, 2009) is one of the oldest methods and focuses on the frequency and position of a given keyphrase. After calculating several frequency based statistics, the method uses post-processing filtering to remove some keyphrases that are too rare or that are not positioned within the first k characters of the document. YAKE (Campos et al., 2018) represents one of the latest upgrades of the statistical approaches, and includes the simpler features proposed by the KPMiner. The main novelty is that it also considers the relatedness of term candidates to general document context, dispersion, and casing of a specific term candidate.

- *graph-based* - methods focus on creating graphs from a given document and then exploit graph properties in order to rank words and phrases. In the first, graph creation step, authors usually consider two adjacent words as two adjacent nodes in a graph *G*. Usually before the graph-creation step some form of word normalization is performed - either stemming or lemmatisation. Since keyword phrases can consist of multiple words, the methods consider the use of a sliding windows to obtain n-grams up to specific value of n, and using obtained n-grams as nodes. Text Rank (Mihalcea and Tarau, 2004) is one of the first such methods. In the second, keyword ranking step, it leverages Google's PageRank (Page et al., 1999) algorithm to rank the nodes according to their importance within the graph *G*. While TextRank is a robust method, it does not account for the position of a given term in the document. This was improved in the PositionRank (Florescu and Caragea, 2017) method that leverages PageRank on one side, and the position of a given term on the other side. An upgrade to the graph-creation step was introduced in Boudin (2018), where they consider encoding the potential keywords into a multipartite[1] graph structure. The method in addition also considers topic information. Similarly to TextRank it leverages PageRank (Page et al., 1999) to rank the nodes. RaKUn (Škrlj et al., 2019) is one of the most recent additions to the family of graph based keyword extractors. The main contribution of this method is that it introduces an intermediate step, that constructs meta-nodes from the initial nodes of the graph via aggregation of the existing nodes. After the construction of the meta-graph, it applies the *load centrality* metric for the term ranking, and also relies on multiple graph redundancy measures.

- *embedding-based* methods are gaining traction with the recent introduction of various off-the shelf pretrained embeddings such as FastText (Bojanowski et al., 2016) or transformer - BERT (Devlin et al., 2019) based embeddings. Key2Vec (Mahata et al., 2018) represents the pioneer of this type of methods, followed by the EmbedRank (Bennani-Smires et al., 2018) method. The aforementioned methods consider the semantic information captured by the distributed word and sentence embedding representations. KeyBERT (Grootendorst, 2020) is currently the state-of-the-art method of the type. The foundation of this method are pre-trained sentence-BERT (Reimers and Gurevych, 2019) based representations. The method considers embedding n-grams of a given size and compares them to the embedding of the entire document. The n-grams closely matching the representation of an entire document (i.e. keywords most representative of an entire document) are retrieved as keywords that best describe the overall document content. In order to diversify the results, the method also introduces the *Max Sum Similarity* metric with which the model selects the candidate phrases with the highest rank that are least similar to each other.

- *language model-based* - methods use language model derived statistics to extract keywords from text. Tomokiyo and Hurst (2003) considered multiple language models and measured the Kullback-Leibler Divergence (Joyce, 2011) for ranking both phrasesness and the informativeness of candidate terms.

2.2. Supervised methods

Supervised methods require manually annotated data for training. The methods can be divided into neural and non-neural.

2.2.1. Non-neural

The first methods that proposed a solution in a supervised manner, considered keyword extraction as a classification task. The KEA method (Witten et al., 1999) treats each word or phrase as a potential keyword, and uses TF-IDF (Sammut and Webb, 2010) metric and word position for representation, and Naive Bayes for classification of a given term as a keyword or not.

2.2.2. Neural

With the recent-gain in computing power and introduction of more modern deep architectures, the field of keyword extraction was taken by storm of neural architectures. The neural approaches can be divided are two groups: one that treat the task as a sequence-to-sequence generation and the one that model the task as sequence-labelling.

Meng et al. (2017) first proposed the idea of keyword extraction as a sequence-to-sequence generation task. In their work they proposed a recurrent generative model with an attention and a copying mechanism (Gu et al., 2016)

[1]Family of graphs where the nodes can be split into multiple disjoint sets.

based on the positional information. An additional strongpoint of this model is that is able to find keywords that do not appear in the text due to it's generative nature.

The first representative of the sequence-labelling method is the approach by Luan et al. (2017), where the authors consider bidirectional Long Short-Term Memory (BiLSTM) layer and a conditional random field (CRF) layer for classification. The more recent approaches of this type utilize the transformer architecture (Vaswani et al., 2017) in their models. An upgrade of the approach by Luan et al. (2017) was proposed by Sahrawat et al. (2020), where contextual embeddings generated by BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and GPT-2 (Radford et al., 2019) were fed into the BiLSTM network. Currently, the state-of-the-art model based on the transformer architecture is the one proposed by Martinc et al. (2020). They employ the tactic of not relying on the massive language model pretraining but rather on the language model pretraining on the much smaller domain specific corpora. This makes the approach more easily transferable to less resourced domains and languages.

Most keyword recognition studies still focus on English. Nevertheless, several multilingual and cross-lingual studies have been conducted recently, also including low-resource languages. One of them is the study by Koloski et al. (2021), which compared the performance of two supervised transformer-based models, a multilingual BERT with a BiLSTM-CRF classification head (Sahrawat et al., 2020) and TNT-KID, in a multilingual setting with Estonian, Latvian, Croatian and Russian news corpora. The authors also investigated whether combining the results of the supervised models with the results of the unsupervised models can improve the recall of the system. In Koloski et al. (2022b), an extensive study was conducted to compare the performance of supervised zero-shot cross-lingual approaches with unsupervised approaches. The study was conducted for six languages - Slovenian, English, Estonian, Latvian, Croatian, and Russian. The authors show that models fine-tuned to extract keywords on a combination of languages *outperform* the unsupervised models, when evaluated on a new previously unseen language not included in the training dataset.

3. Data

We conduct our experiments on the Slovenian SentiNews dataset (Bučar, 2017), which was originally used for news sentiment analysis, but nevertheless does contain manually labeled keywords and was therefore identified as suitable for keyword extraction (Koloski et al., 2022a). Before feeding the datasets to the models, they are lowercased. We split the dataset into three different splits: *train*, *validation* and *test*.

3.1. Exploratory data analysis

Next, we preform exploratory data analysis (*EDA*) on the given dataset. There are total of *7514* documents, *4796 (64%)* for training, *1199 (16%)* for validation and *1519 (20%)* for testing, which makes the dataset relatively small in comparison to some English keyword extraction datasets, such as for example KPTimes (Gallina et al., 2019), containing more than 200,000 documents. We benchmark all of our models on the same test split that was already used in the study by (Koloski et al., 2022b), in order to make our results directly comparable to the ones in the related work.

The documents have a similar structure in all of the three splits, having on average *370* words (*370.10* words in the train split, *366.89* words in the validation split and *377.46* words in the test split) and on average around *15* sentences (*15.419* sentences in the train split, *15.203* sentences in the validation split and *15.662* sentences in the test split).

	Split		
Property	Train	Valid	Test
Document statistics			
# of documents	4796	1199	1519
avg. # of sentences	15.419	15.2026	15.6622
avg. # of words	370.10	366.89	377.46
Keywords statistics			
# of keywords	19429	4773	5903
# of unique keywords	4414	1854	2049
# of unique keywords per document	0.9203	1.5462	1.3489
# of keywords per document	4.0052	4.1643	3.8861
keywords present in the document	59.91 %	60.54 %	59.95 %
Keyword composition statistics			
Proportion of 1-word terms	92.77%	93.17 %	92.68 %
Proportion of 2-word terms	5.88 %	5.61 %	5.98 %
Proportion of 3-word terms	0.62 %	0.57 %	0.58 %
Proportion of more than 3-word terms	0.74 %	0.65 %	0.76 %

Table 1: Dataset statistics. We conducted three different statistical analyses. The first one was on the document level and it considered counting the word and sentence tokens. The second focused on the keyword level statistics, such as total number of keywords, number of unique keywords, and the proportion of all versus unique keywords per document. Finally, we explored the composition of keywords, i.e. how many of them were composed of single words, two words, three words or more words.

There are in total *30,105* keywords in the dataset, with *8,317* of them being unique. On average there are *4* keywords per document in the training split, *4.16* keywords per document in the validation split and *3.8861* keywords per document in the test split. In regards to the unique keywords per split, there are *0.92* unique keywords per document in the training split, *1.55* in the validation split and *1.35* keywords per document in the test split. Since the keyword extractors used in this study are only able to extract keywords that are present in the data, we also calculated the share of keywords that are present in the document. In the training set, there were 59.91% of the keywords present, in the validation set 60.54% and in the testing set 59.95%.

Finally, we conducted a study on the composition of keywords in which we explored how many words constitute a specific keyphrase. In all of the splits, more than 92% of the keywords contained only a single words, 2-word terms represented about 5% of the keywords, while 3 or more word terms represented around 3% of all keywords. The most common keyword was *gospodarstvo* with *2,350* occurrences (representing roughly 12% of all keyword occurences), followed by *ekonomija* with *1315 (6.76%)* oc-

currences, followed by *banka* with *147 (0.08%)* occurrences.

These keywords suggest that most of the articles come from the economic and financial domain. In order to explore the structure and content of the dataset in more detail, we do additional network science analysis on the graph of 100 most-frequent terms. We construct a graph G_{100} in the following manner: we create links among every pair of keywords that accompany a given article in the training split. We repeat the step for every article in the training split.

We next focus on community detection in the constructed graph. For that purpose, we use the Louvain algorithm (Blondel et al., 2008). The algorithm detects four distinct communities. The first one colored *green* is the most central community - the community with the highest amount of shared links with the three other detected communities. It contains general terms like *family*, *declaration*, *NKB(a bank)*, *sod*. Next one is *purple* and it talks about the trend of rising *taxes*, new *laws* and the petrochemical industry. The community colored in *blue* represents the economic news about *infrastructure* and *construction* industries. The last is the *yellow* community that talks about *financial help* from the government and the *European union*, accompanied by the *unemployment* and the slow rise of *GDP*. The graph and its detected communities are presented in Figure 1.

4. Methods

In our experiments, we follow the experimental setting proposed in Koloski et al. (2021) and Koloski et al. (2022b). The methods and the hyperparameters used are described below.

4.1. Unsupervised approaches

We evaluate three types of unsupervised keyword extraction methods, statistical, graph-based, and embedding-based, described in Section 2. Note that these models were already evaluated on the same corpus in Koloski et al. (2022b).

4.1.1. Statistical methods

- **YAKE** (Campos et al., 2018): We consider n-grams with $n \in \{1, 2, 3\}$ as potential keywords.
- **KPMiner** (El-Beltagy and Rafea, 2009): We apply least allowable seen frequency of 3, while we set the *cutoff* to 400.

4.1.2. Embedding-based methods

- **KeyBERT** (Grootendorst, 2020): For document embedding generation we employ sentence-transformers (Reimers and Gurevych, 2019), more specifically the *distiluse-base-multilingual-cased-v2* model available in the Huggingface library[2]. Initially, we tested two different KeyBERT configurations: one with n-grams of size 1 and another with n-grams ranging from 1 to 3, with *MMR*=$false$ and with *MaxSum*=$false$. The unigram model outscored the model that considered n-grams of sizes 1 to 3 as keyword candidates for all languages, therefore in the final report we show only the results for the unigram model.

4.1.3. Graph-based methods

- **MultipartiteRank** (Boudin, 2018): We set the minimum similarity threshold for clustering at 74%.
- **RaKUn** (Škrlj et al., 2019): We use edit distance for calculating distance between nodes, and remove stopwords (using the *stopwords-iso* library[3]), a *bigram-count_threshold* of 2 and a *distance_threshold* of 2. An example graph of the RaKUn document representation and its predicted keywords are presented in Figure 2.

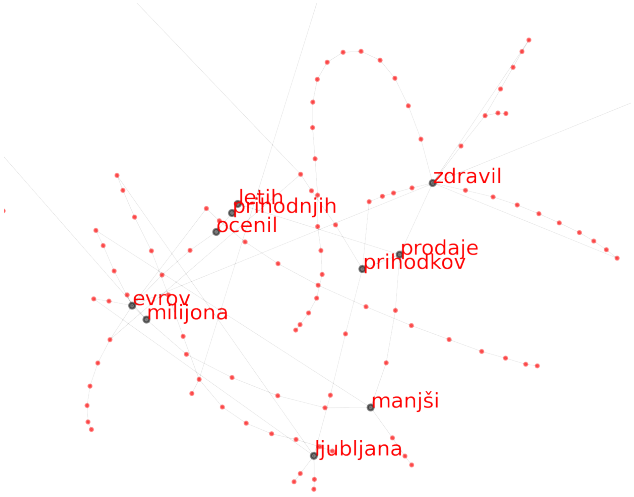


Figure 2: Visualization of one training example as it was seen by the RaKUn method. The visualization is generated via the Py3Plex (**?**) library. Top three extracted tokens here are *Ljubljana*, *Prihodki* and *Zdravil* - depicting that the article is about purchase of *medicine*.

We use the PKE (Boudin, 2016) implementations of *YAKE*, *KPMiner* and *MultiPartiteRank*. We use the official implementation for the RaKUn (Škrlj et al., 2019) and for the KeyBERT model (Grootendorst, 2020). For unsupervised models, the number of returned keywords need to be set in advance. Since we employ F1@10 as the main evaluation measure (see Section 4.3.), we set the number of returned keywords to 10 for all models.

4.2. Supervised approaches

We test two distinct state-of-the-art transformer-base models, BERT (Devlin et al., 2019) and TNT-KID (Martinc et al., 2020).

4.2.1. BERT sequence labelling

As a strong baseline, we utilize the transformer-based BERT model (Devlin et al., 2019) with a token-classification head consisting of a simple linear layer for

[2] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

[3] https://github.com/stopwords-iso/stopwords-iso

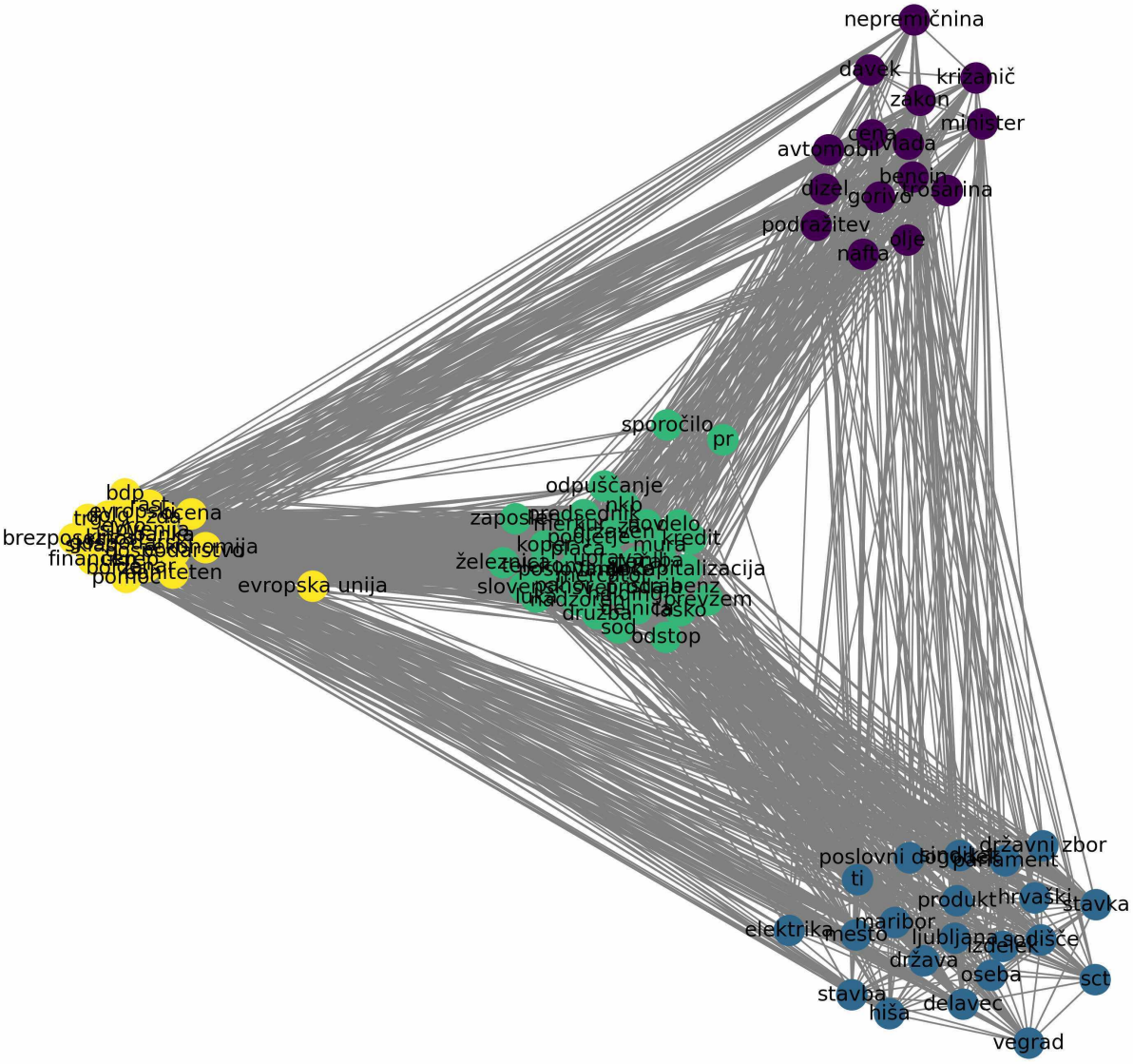


Figure 1: Visualization of the derived communities of the co-occurrence graph.

all our supervised approaches. We treat the keyword extraction task as a sequence classification task. We follow the approach proposed in Martinc et al. (2020) and predict binary labels (1 for 'keywords' and 0 for 'not keywords') for all words in the sequence. The sequence of two or more sequential keyword labels predicted by the model is always interpreted as a multi-word keyword. More specifically, we employ the *bert-uncased-multilingual* model from the HuggingFace library (Wolf et al., 2019) and fine-tune it on the SentiNews train split using an adaptive learning rate (starting with the learning rate of $3 \cdot 10^{-5}$), for up to 10 *epochs* with a batch-size of 8. Note that we chose this model since it is the best performing model on the Slovenian SentiNews dataset according to the study by Koloski et al. (2022b).

4.2.2. TNT-KID sequence labelling

Same as for BERT, we follow the approach proposed in Martinc et al. (2020) and predict binary labels (1 for 'keywords' and 0 for 'not keywords') for all words in the sequence. Again, the sequence of two or more sequential keyword labels predicted by the model is always interpreted as a multi-word keyword. We first pretrain TNT-KID as an autoregressive language model on the domain specific news corpus containing *884,407* news articles crawled from websites of several Slovenian news outlets. The model was trained for 10 epochs. After that, the model was fine-tuned on the SentiNews train set for the keyword extraction task, again for up to 10 epochs. Sequence length was set to 256, embedding size to 512 and batch size to 8, and we employ the same preprocessing as in the original study (Martinc et

al., 2020).

4.3. Evaluation setting

To evaluate the models, we compute F1, Recall, and Precision on 10 retrieved words. We next formally represent the Recall@10 metric:

$$Recall@10 = \frac{(\#\text{ of recommended relevant items @ 10})}{(\text{total }\#\text{ of relevant items}).}$$

and Precision@10 metric:

$$Precision@10 = \frac{(\#\text{ of recommended relevant items @ 10})}{(\#\text{ of recommended items @10})}$$

We omit the documents in which there are no keywords or which do not contain keywords. We do this because we only use approaches that extract words (or multi-word expressions) from the given document and cannot process keywords that do not appear in the text. All approaches are evaluated on the same monolingual test splits, which are not used for training the supervised models. Lower case and lemmatization are performed during the evaluation for both the gold standard and the extracted keywords (keyphrases).

5. Results

In this section we examine the results of the evaluation of the proposed models. We first study the results of the unsupervised methods and later the results of the supervised models.

5.1. Unsupervised methods

In this study we evaluate 5 different unsupervised methods: 2 statistical, 1 embedding-based and 2 graph-based methods. Comparing the two statistical methods, *KPMiner* outscored the *YAKE* method in terms of f1-score and precision. The embedding based *KeyBERT* method achieved the best results when compared to other unsupervised methods. From the graph-based methods, *RaKUn* performed the best in comparison with the *MPRU* method, achieving nearly 100% relative improvement. Table 2 presents the results for all systems and evaluation metrics in detail.

5.2. Supervised methods

We use two different supervised methods based on the sequence labeling paradigm. BERT based model outperforms TNT-KID in terms of recall by about 5 percentage points, achieving the best recall out of all models. In terms of precision, TNT-KID outscores the BERT model by 9.04 percentage points and achieves the best precision@10 score - 38.58%. We believe this is due to the extensive language-model pretraining on a large domain specific Slovenian news corpus and the frequency of common co-occurrence patterns in the data, that TNT-KID has learned to exploit successfully.

Model	precision@10	recall@10	f1-score@10
Statistical			
KPMiner	*12.80*	7.44	9.41
YAKE	5.91	*12.13*	7.94
Embedding-based			
KeyBert	12.13	12.00	11.53
Graph-based			
RaKUn	*6.72*	*12.52*	*8.75*
MPRU	3.39	6.96	4.55
Sequence-labelling			
BERT	29.54	**47.81**	32.59
TNT-KID	***38.58***	42.81	***40.59***

Table 2: Comparison of the evaluation of the proposed approaches. We report on the precision@10, recall@10 and f1-score@10. The scores of the best performing system of a specific type (i.e. statistical, embedding-based, graph-based or sequence-labelling based) are written in italic. The scores for the overall best-preforming model according to each metric are written in bold and presented in percents.

The final comparison of both the unsupervised and supervised models is presented in Table 2. The *TNT-KID* model performed the best in terms of precision and F1-score while *BERT* model performed the best out of all models in terms of recall. The supervised models outscored the unsupervised models by a large margin on the given task. The ranking of the models in terms of various metrics is given in Figure 3.

6. Conclusion and further work

In this study, we compared the performance of supervised and unsupervised keyword extraction methods on the new public benchmark for keyword extraction, derived from Slovenian SentiNews corpus. We have compared 8 different models, among them also TNT-KID, which has not been tested on Slovenian dataset yet. Five unsupervised approaches can be further divided into two graph-based, two statistical and one embedding-based approach. The embedding-based method *KeyBERT* showcased superior performance to the other unsupervised methods in terms of F1-score at 10 retrieved keywords.

When it comes to supervised approaches, we experimented with two transformer based models - one leveraging multilingual BERT and the other the TNT-KID method - that model keyword extraction as a sequence labelling task. The TNT-KID approach outperformed BERT-based approach (and all unsupervised models) in terms of precision and F1-score. These results therefore support the claims of the original study by (Martinc et al., 2020) that TNT-KID can be easily adapted for employment on less-resource languages, such as Slovenian, by domain specific unsupervised language model pretraining. By employing TNT-KID on the SentiNews dataset, we have advanced the state-of-the-art on the benchmark Slovenian keyword extraction dataset.

For further work, we plan to explore how potentially we can improve the results by constructing ensembles of keyword extractors. We will also propose testing several

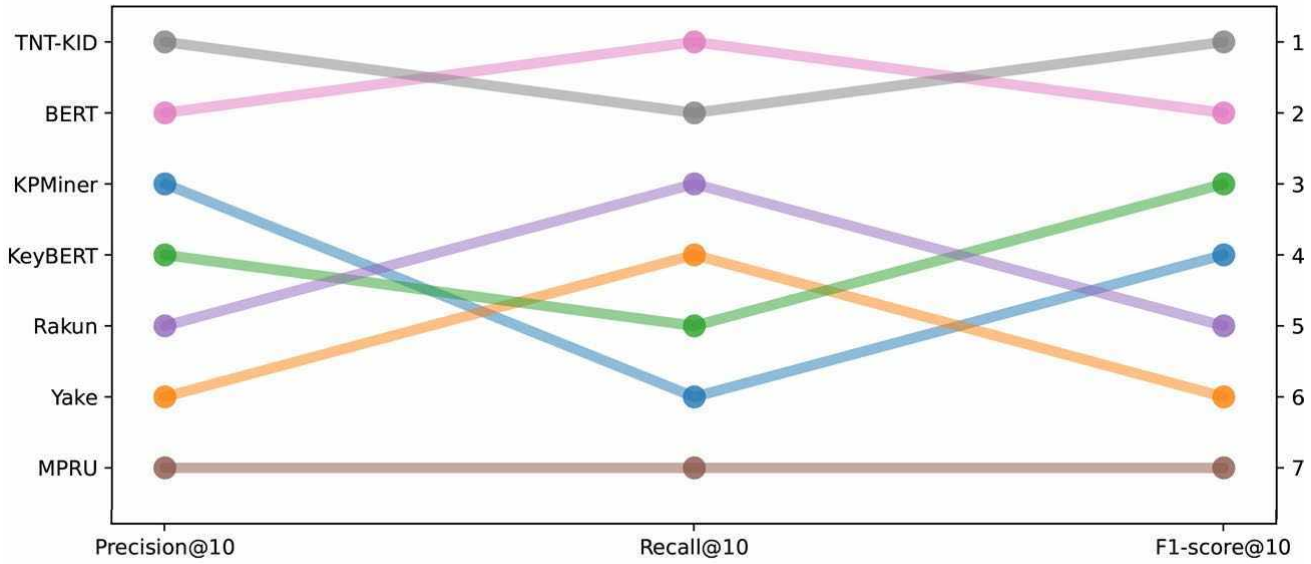


Figure 3: Comparison of the models ranking with respect to Precision@10, Recall@10 and F1-score@10.

different data splitting strategies, in order to study the possible effect of different splitting strategies on performance of different models and to establish the best possible split strategy. We also hypothesize that a possible improvement can be introduced by taking into account the co-occurrence of various pairs of keywords. Finally, in the future we plan to expand our experiments to also include the recently introduced monolingual massively pretrained model for Slovenian, SloBERTa (Ulčar and Robnik-Šikonja, 2020). We plan to fine-tune this model for the keyword extraction task and compare it to the TNT-KID, to check whether state-of-the-art can be advanced even further.

7. Availability

The best-performing *TNT-KID* based model is available as a docker model on the following link `https://gitlab.com/boshko.koloski/tnt_kid_app_slo`.

8. Acknowledgements

The authors acknowledge the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103) and the project Computer-assisted multilingual news discourse analysis with contextual embeddings (CANDAS, J6-2581).

9. References

Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. 2018. Simple unsupervised keyphrase extraction using sentence embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 221–229, Brussels, Belgium, October. Association for Computational Linguistics.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Florian Boudin. 2016. PKE: an open source Python-based keyphrase extraction toolkit. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 69–73, Osaka, Japan, December.

Florian Boudin. 2018. Unsupervised keyphrase extraction with multipartite graphs. *CoRR*, abs/1803.08721.

Jože Bučar. 2017. Manually sentiment annotated slovenian news corpus SentiNews 1.0. Slovenian language resource repository CLARIN.SI.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. Yake! collection-independent automatic keyword extractor. In *European Conference on Information Retrieval*, pages 806–810. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding.

Samhaa R El-Beltagy and Ahmed Rafea. 2009. KP-Miner: A keyphrase extraction system for English and Arabic documents. *Information systems*, 34(1):132–144.

Corina Florescu and Cornelia Caragea. 2017. PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115, Vancouver, Canada, July. Association for Computational Linguistics.

Ygor Gallina, Florian Boudin, and Béatrice Daille. 2019. Kptimes: A large-scale dataset for keyphrase generation on news documents. In *Proceedings of the 12th Inter-*

national Conference on Natural Language Generation, pages 130–135.

Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with bert.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany, August. Association for Computational Linguistics.

James M. Joyce, 2011. *Kullback-Leibler Divergence*, pages 720–722. Springer Berlin Heidelberg, Berlin, Heidelberg.

Boshko Koloski, Senja Pollak, Blaž Škrlj, and Matej Martinc. 2021. Extending neural keyword extraction with TF-IDF tagset matching. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 22–29, Online, April. Association for Computational Linguistics.

Boshko Koloski, Matej Martinc, Ilija Tavchioski, Blaž Škrlj, and Senja Pollak. 2022a. Slovenian keyword extraction dataset from SentiNews 1.0. Slovenian language resource repository CLARIN.SI.

Boshko Koloski, Senja Pollak, BlaÅ¾ Å krlj, and Matej Martinc. 2022b. Out of thin air: Is zero-shot cross-lingual keyword detection better than unsupervised? In *Proceedings of the Language Resources and Evaluation Conference*, pages 400–409, Marseille, France, June. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2641–2651, Copenhagen, Denmark, September. Association for Computational Linguistics.

Debanjan Mahata, John Kuriakose, Rajiv Ratn Shah, and Roger Zimmermann. 2018. Key2Vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 634–639, New Orleans, Louisiana, USA, June. Association for Computational Linguistics.

Matej Martinc, Blaž Škrlj, and Senja Pollak. 2020. Tnt-kid: Transformer-based neural tagger for keyword identification. *Natural Language Engineering*, pages 1–40.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592, Vancouver, Canada, July. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November. Previous number = SIDL-WP-1999-0120.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAi.

Nils Reimers and Iryna Gurevych. 2019. 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, 11.

Dhruva Sahrawat, Debanjan Mahata, Mayank Kulkarni, Haimin Zhang, Rakesh Gosangi, Amanda Stent, Agniv Sharma, Yaman Kumar, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Keyphrase extraction from scholarly articles as sequence labeling using contextualized embeddings. In *Proceedings of European Conference on Information Retrieval (ECIR 2020)*, pages 328–335, Lisbon, Portugal. Springer.

Claude Sammut and Geoffrey I. Webb, editors, 2010. *TF–IDF*, pages 986–987. Springer US, Boston, MA.

Takashi Tomokiyo and Matthew Hurst. 2003. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, page 33–40, Sapporo, Japan. Association for Computational Linguistics.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Slovenian roBERTa contextual embeddings model: SloBERTa 1.0.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Vancouver, Canada. Curran Associates, Inc.

Blaž Škrlj, Andraz Repar, and Senja Pollak. 2019. Rakun: Rank-based keyword extraction via unsupervised learning and meta vertex aggregation. *CoRR*, abs/1907.06458.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Kea: Practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries*, DL '99, page 254–255, Berkeley, California, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Spremljevalni korpus Trendi: metode, vsebina in kategorizacija besedil

Iztok Kosem,‡* Jaka Čibej,‡ Kaja Dobrovoljc,‡* Nikola Ljubešić‡

‡ Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana
iztok.kosem@ijs.si, jaka.cibej@ijs.si, kaja.dobrovoljc@ijs.si, nikola.ljubesic@ijs.si
* Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana

Povzetek

V prispevku opisujemo postopek gradnje korpusa Trendi – prvega spremljevalnega korpusa za slovenščino. Prva različica korpusa, imenovana Trendi 2022-05, vsebuje več kot 565 milijonov pojavnic iz več kot 1,4 milijona besedil. Namen korpusa je, da tako strokovni kot nestrokovni javnosti ponudi podatke o aktualni jezikovni rabi in omogoči spremljanje pojavljanja novih besed ter upadanja ali naraščanja rabe že obstoječih. Predstavimo metodologijo izdelave in vsebino korpusa ter prve korake pri načrtovani strojni klasifikaciji korpusnih besedil v kategorije (npr. gospodarstvo, okolje), s katerimi bo mogoče v korpusu spremljati jezikovno rabo tudi po tematskih področjih. Predstavimo tudi rezultate ankete, s katero smo preverili uporabniška pričakovanja o jezikovnem viru za spremljanje jezikovne rabe.

The *Trendi* Monitor Corpus of Slovene: Methods, Content, and Text Categorization

In the paper, we present the compilation of the Trendi corpus – the first monitor corpus of Slovene. The first version of the corpus, named Trendi 2022-05, contains over 565 million tokens coming from more than 1.4 million different texts. The purpose of the corpus is to provide both experts and non-experts with data on contemporary language use and enable the monitoring of the appearance of new words or the increase/decrease in the use of existing words. We present the methodology of corpus compilation, its content, and the first steps for the automatic classification of corpus texts into categories (such as economics and environment), which will enable the monitoring of language use by thematic areas. We also describe the results of a survey, the goal of which was to collect feedback on user expectations from a language monitoring resource

1. Uvod

Jezik se nenehno spreminja, pojavljajo se nove besede, obstoječe besede in besedne zveze dobivajo nove pomene, določene besede ali njihovi pomeni se prenehajo uporabljati ipd. V zadnjem času, tudi zaradi epidemije covida-19, ki je prinesla veliko novega izrazoslovja, je še posebej veliko pozornosti deležno področje neologije, tako leksikalne (nove besede) kot semantične (novi pomeni).

Za spremljanje sprememb v jeziku se tipično uporabljajo spremljevalni korpusi, ki vsebujejo najnovejša besedila v jeziku. Spremljevalni korpusi zapolnjujejo manko referenčnih korpusov, katerih izdelava zaradi raznovrstnosti besedil in njihovih formatov ter obsega traja dlje časa. V času tehnološkega napredka in ob dejstvu, da je zdaj zelo veliko besedil dostopnih na spletu, je izdelava spremljevalnih korpusov postala enostavnejša; kar je objavljeno danes, je lahko že jutri vključeno v korpus.

Za slovenščino kljub bogati opremljenosti na področju korpusov do zdaj nismo imeli spremljevalnega korpusa, čeprav se je med različnimi deležniki kazala jasna potreba po njem. Naslavljanja tega manka smo se lotili v okviru projekta *Spremljevalni korpus in spremljajoči podatkovni viri* (SLED),[1] ki poteka od oktobra 2021 do novembra 2022 in ga sofinancira Ministrstvo za kulturo Republike Slovenije. Cilj projekta ni samo izdelati spremljevalni korpus, temveč tudi pripraviti infrastrukturo za njegovo redno posodabljanje.

V prispevku najprej ponujamo pregled nekaterih pomembnejših tujih spremljevalnih korpusov, nato pa predstavimo metodologijo in vsebino spremljevalnega korpusa Trendi. Sledi predstavitev klasifikacije tematskih kategorij, ki smo jo izdelali za pripravo modela za avtomatsko kategorizacijo besedil. V zadnjem delu predstavimo anketo med uporabniki o želenih statističnih izračunih iz korpusa. V zaključku predstavimo načrte za prihodnje delo.

2. Spremljevalni korpusi

V mednarodnem prostoru so spremljevalni korpusi prisotni že od 20. stoletja. Eden prvih je bil the Bank of English, ki je bil prvič objavljen leta 1991. Vsebuje več kot 650 milijonov besed[2] in je danes vključen v 4,5-milijardni korpus COBUILD založbe Collins. Korpus ni prosto dostopen, poleg zaposlenih na založbi Collins ga lahko uporabljajo tudi zaposleni in študentje na Univerzi v Birminghamu.

Za angleščino je danes pomemben predvsem korpus NOW (News on the Web; Davies, 2016-), ki vsebuje več kot 15 milijard besed iz spletnih časopisov in revij. Korpus zajema besedila od 2010 naprej. Kot je omenjeno na spletni strani,[3] korpus vsak mesec naraste za 180-200 milijonov besed.

Obsežna zbirka korpusov za spremljanje sprememb v jeziku, ki poleg angleščine pokriva še več kot 35 drugih jezikov, so korpusi Timestamped JSI. Korpusi vsebujejo novice, ki jih zbira JSI Newsfeed na Institutu "Jožef Stefan" (Trampuš in Novak, 2012). Korpusi za 18 jezikov so na voljo v orodju Sketch Engine (Kilgarriff et al., 2004),[4] v katerem imajo poleg ostalih funkcij orodja uporabniki na voljo tudi t. i. Trende (Herman, 2013), funkcijo, ki pomaga prepoznavati trende v rabi besed. Korpusi v Sketch Enginu

[1] https://sled.ijs.si/

[2] Žal nismo našli podatka, kdaj je bil korpus nazadnje posodobljen.

[3] https://www.english-corpora.org/now/

[4] https://www.sketchengine.eu/

vsebujejo besedila od 2014 do aprila 2021 (čas zadnje posodobitve) in so različnih velikosti; korpus angleščine na primer vsebuje približno 60 milijard besed.

Obstaja še precej drugih spremljevalnih korpusov, ki pa so pogosto na voljo zgolj za interno rabo. Primer takšnega korpusa je ONLINE, dinamični spremljevalni korpus češkega jezika, ki ga izdeluje Inštitut za češki nacionalni korpus.[5] Velik je približno 6,3 milijarde besed in vsebuje spletne novice, komentarje (pod spletnimi novicami), besedila s forumov in družabnih omrežij (Facebook, Twitter, Instagram). Korpus ONLINE je razdeljen na dva komplementarna korpusa: ONLINE_NOW in ONLINE_ARCHIVE. Prvi je posodobljen vsak dan in pokriva obdobje zadnjega meseca in preteklih šestih mesecev. ONLINE_ARCHIVE pokriva obdobje od februarja 2017 do prvega meseca, ki ga vsebuje ONLINE_NOW. Tako se vsebina zadnjega meseca po starosti v korpusu ONLINE_NOW na začetku vsakega meseca preseli v ONLINE_ARCHIVE.

Obstajajo tudi manjši in bolj specializirani spremljevalni korpusi, kakršen je npr. korpus Coronavirus (Davies, 2019-), ki zajema obdobje od januarja 2020 do danes in vsebuje več kot 1,4 milijarde besed. V njem so spletne novice v angleščini, vsak dan pa naraste za 3 do 4 milijone besed.

Do določene mere vlogo spremljevalnega korpusa opravljajo tudi diahroni korpusi, seveda pod pogojem, da vsebujejo čim novejša besedila. Kot primer lahko navedemo korpus sodobne ameriške angleščine (Corpus of Contemporary American English; Davies, 2008-), ki vsebuje besedila od leta 1990 do marca 2020 (zadnja posodobitev) in obsega več kot milijardo besed. Prednost korpusa je, da je žanrsko uravnotežen, saj vsebuje besedila iz osmih različnih žanrov (govorjeni jezik, leposlovje, revije, časopise, znanstvena besedila, televizijske in filmske podnapise, bloge in ostale spletne strani). Slovenski ekvivalent bi bil korpus Gigafida 2.0 (Krek et al., 2019),[6] ki obsega 1,13 milijarde besed, vendar pa je v primerjavi s korpusom sodobne ameriške angleščine manj ažuren (vsebuje samo besedila do leta 2018).

Za slovenščino do danes še ni obstajal pravi spremljevalni korpus. Obstajajo sicer viri, kot je Jezikovni sledilnik (Kosem et al., 2021),[7] ki že izkorišča najsodobnejše podatke o jezikovni rabi, v konkretnem primeru od JSI Newsfeeda, za izdelavo neke vrste začasnih korpusov, na katerih se potem izvajajo statistični izračuni. Taka ciljna raba je seveda tudi potrebna, vendar pa je namenjena nestrokovni javnosti; po drugi strani strokovna javnost, kot so leksikografi_ke, jezikoslovci_ke, drugi raziskovalci_ke potrebujejo dostop do izvirnih besedil, če želijo opravljati še druge analize.

3. Korpus Trendi

Izdelave prvega spremljevalnega korpusa za slovenščino, ki smo ga poimenovali Trendi, smo se lotili v okviru projekta SLED. Poleg izdelave in rednega posodabljanja korpusa Trendi ima projekt še dva cilja: pripravo na korpusnih podatkih temelječe statistike o različnih vidikih rabe besed in izdelavo orodja, ki bo besedila avtomatsko opremilo s podatkom o tematski kategoriji.

3.1. Metodologija in vsebina korpusa

Z metodološkega vidika smo pri snovanju korpusa Trendi morali sprejeti dve odločitvi: obdobje, ki ga bo korpus pokrival, in kako pogosto bo korpus posodobljen. Pri odločitvi o obdobju smo izhajali iz želje, da bi korpus Trendi vedno pokrival manko najnovejše različice referenčnega (pisnega) korpusa Gigafida, trenutno je to 2.0. V tem trenutku to pomeni, da bo Trendi vseboval besedila od 2019 naprej. To pomeni, da se ob objavi nove različice korpusa Gigafida (npr. korpus Gigafida 3.0 bo objavljen v sklopu projekta *Razvoj slovenščine v digitalnem okolju* - RSDO),[8] obdobje korpusa Trendi ustrezno prilagodi.

Tesna povezanost s korpusom Gigafida tudi pomeni, da bo korpus Trendi predstavljal standardno pisno slovenščino. Odločitev se nam zdi smiselna tudi zato, ker sta nestandardna oz. govorjena slovenščina pokrita s korpusi, kot sta JANES[9] in Gos,[10] in je torej njun razvoj predmet ločenih projektov. Navsezadnje pa ne gre pozabiti na nastajajoči korpus metaFida,[11] ki bo združil vse slovenske korpuse.

Pri pripravi seznama virov za vključitev v korpus Trendi smo izhajali iz seznama slovenskih spletnih virov, ki jih najdemo v servisu JSI Newsfeed. Izdelali smo seznam vseh virov od leta 2019 do konca 2021, pridobili smo tudi podatek o skupnem številu besedil na vir. Nato smo pri pripravi seznama za korpus Trendi podrobno analizirali vsakega od 243 virov. 90 virov smo izključili, ker je šlo za tuje ali slovenske spletne strani z vsebino v tujem jeziku. Nato smo s seznama odstranili še 34 virov, nekatere zato, ker niso vsebovali medijskih novic (blogi, spletne strani vladnih uradov in podjetij), druge zato, ker je njihova vsebina preveč specializirana (npr. repozitoriji akademskih publikacij so primernejši za korpuse, kot je Korpus akademske slovenščine). Ena od strani (preberi.si) je bila s seznama odstranjena zato, ker je agregator novic iz drugih virov. Končni seznam korpusa Trendi tako vsebuje 110 virov, med tistimi, ki so v obdobju 2019-2021 prispevali največ novic, so sta.si (260.080 besedil), rtvslo.si (97.924), siol.net (69.471), delo.si (65.415), 24ur.com (61.623), dnevnik.si (47.749) in vecer.com (45.548).

Seznam virov se bo redno posodabljal, saj lahko pričakujemo pojav novih spletnih strani, pa tudi ukinitev obstoječih. Kot primer lahko navedemo spletno stran necenzurirano.si, ki se je pojavila šele leta 2020 in je že 28. po številu novic (8.494). Dodajanje novih virov v korpus pomeni tudi večje število besed na mesečni ravni in posledično večji korpus Trendi. Trenutni okvirni izračuni kažejo, da se bo Trendi vsak mesec povečal za 10-15 milijonov pojavnic, pri čemer je bil povprečen mesečni obseg leta 2019 12,5 milijona pojavnic, leta 2021 pa že 21 milijonov pojavnic.

Zaradi narave korpusa Trendi bodo potrebne redne posodobitve, ki so zaenkrat predvidene na mesečni ravni, kot je praksa pri podobnih tujih korpusih. To se zdi trenutno realno, upoštevajoč časovno zahtevnost pridobivanja in

[5] https://korpus.cz/

[6] https://viri.cjvt.si/gigafida/

[7] https://viri.cjvt.si/sledilnik/slv/

[8] https://slovenscina.eu/

[9] https://www.clarin.si/kontext/query?corpname=janes

[10] http://www.korpus-gos.net/

[11] https://www.clarin.si/kontext/query?corpname=mfida01

označevanja besedil, pretvorb v potreben format in vključevanje korpusa v konkordančnike.

3.2. Priprava besedil

Za pripravo besedil smo pripravili cevovod, ki vključuje pridobivanje besedil, označevanje na različnih ravneh, združevanje po virih in obdobjih ter pretvorbo v različne formate. Pridobivanje besedil je zaenkrat vezano na servis JSI Newsfeed, ki uporablja protokol RSS novic, vendar pa smo sredi priprave lastnega postopka luščenja. Za to smo se odločili predvsem zato, ker smo odkrili, da so pri mnogih virih potrebne izboljšave pri pridobivanju besedil, npr. poleg besedila so izluščeni še drugi deli strani, besedilo ni pridobljeno v celoti ipd. Poleg tega strani včasih vsebujejo pomembne metapodatke o besedilu, ki trenutno niso del zajema. V novem postopku bomo ročno preverili rezultate pridobivanja besedil z vsakega vira in prilagodili algoritem za vsak vir, kjer se bo izkazala potreba po prilagoditvi.

Nekateri viri, kot so sta.si, delo.si itd. imajo vsebine zaklenjene oziroma so dostopne samo naročnikom. Pri pridobivanju prek protokola RSS so tako prosto dostopni samo povzetki ali prvih nekaj odstavkov, včasih celo samo naslov in podnaslov. Pri reševanju problema smo združili moči z ekipo, ki v okviru projekta RSDO oz. priprave korpusa Gigafida 3.0 sklepa pogodbe z besedilodajalci. Dogovor z besedilodajalci vključuje redno dostavljanje celotnih besedil. Posledično bo končna oblika cevovoda za korpus Trendi kombinacija priprave besedil, pridobljenih s spleta, in besedil, ki jih bodo v digitalni obliki poslali besedilodajalci.

Del postopka pridobivanja besedil je tudi deduplikacija, ki je trenutno omejena zgolj na raven vira besedila; del cevovoda je namreč preverjanje, da se besedilo z istim URL-jem ne ponovi. Zavedamo se, da zaradi pokrivanja istih dogodkov obstaja velika prekrivnost med viri. Še več, mnogi viri osnujejo številne novice na podlagi vsebin sta.si, kar pripelje do podvajanja besedila na ravni stavkov, odstavkov ali tudi celotne vsebine. Kljub temu za namene korpusa Trendi deduplikacija na ravni vsebine ni predvidena, saj želimo uporabnikom omogočiti analizo vsebin posameznih virov ter primerjalne analize med viri. Deduplikacija pa bo najbrž opravljena pri pripravi besedil za novo različico korpusa Gigafida, kot je bila praksa v preteklih različicah (Krek et al., 2019).

Sledi postopek strojnega označevanja besedil, za kar uporabljamo označevalni cevovod CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019),[12] ki se kot referenčno orodje za slovnično označevanje besedil v slovenščini aktivno razvija v okviru projekta RSDO. Orodje je nadgradnja odprtokodnega orodja Stanza (Qi et al., 2020), ki v primerjavi z izvorno programsko opremo podrobneje naslavlja specifike slovenščine, zlasti na ravni stavčne segmentacije, tokenizacije, oblikoskladenjskega označevanja in lematizacije po sistemu JOS (Erjavec et al., 2010). Poleg navedenih ravni orodje besedila tudi skladenjsko razčleni po sistemu Universal Dependencies (Dobrovoljc et al., 2017) in v njih označi imenske entitete (Zupan et al., 2017), kot so imena oseb, krajev, organizacij ipd.

Po končanem označevanju se v cevovodu opravi še pretvorba besedil iz privzetega formata označevalnega orodja (CONNL-U) v TEI XML, ki ga med drugim potrebujemo za statistične izračune s programom LIST (Krsnik et al., 2019). V ta proces sta vključena še dva povezana postopka združevanja besedil: združevanje besedil po viru na dan (vsakodneven postopek) in združevanje besedil istega vira za cel mesec (enkrat na mesec, na začetku novega meseca za nazaj). V zadnjem koraku, ki ga izvajamo enkrat mesečno in ga moramo pognati ločeno zaradi kombinacije XSLT in skripte Perl, je opravljena še pretvorba mesečnih datotek (razdeljenih po viru) v format VERT, ki ga uporabljata konkordančnika KonText (Machálek, 2020) in NoSketch Engine (Rychlý, 2007).

3.3. Prva različica korpusa Trendi

Prva različica korpusa Trendi, imenovana Trendi 2022-05, je bila objavljena junija 2022 in vsebuje 565.308.991 pojavnic oz. malo več kot 473 milijonov besed. V korpusu je 1.436.548 besedil od 48 izdajateljev, pri čemer imajo največje deleže Slovenska tiskovna agencija (337.484; 23,5 %), Delo d.o.o. (128.164; 8,9 %), Radiotelevizija Slovenija (124.861; 8,7 %), Media24 d.o.o. (100.587; 7 %), PRO PLUS d.o.o. (86.578; 6 %) in TSMedia d.o.o. (83.342; 5,8 %).

3.4. Dostopnost korpusa Trendi

Korpus Trendi je za brskanje prosto dostopen v treh konkordančnikih CLARIN.SI – konkordančniku KonText[13] in dveh različicah konkordančnika NoSketchEngine,[14] tako KonText kot NoSketch Engine imata več enakih funkcionalnosti (enostavno in napredno iskanje ipd.), vendar pa KonText ponuja možnost registracije in shranjevanje iskanj in priljubljenih korpusov, NoSketchEngine pa dodatne funkcionalnosti, kot je luščenje ključnih besed (angl. *keywords*) iz korpusov, za uporabo katerih ni potrebna registracija. Konkordančnik NoSketch Engine je na CLARIN.SI poleg starejše različice (Bonito) po novem na voljo tudi v novejši različici uporabniškega vmesnika (Crystal),[15] ki zagotavlja izboljšano uporabniško izkušnjo in dolgoročnejše vzdrževanje.

Odprto dostopna različica korpusa Trendi bo zaradi omejitev avtorskih pravic izdelana po isti metodi kot ccGigafida 1.0 (Logar et al., 2013), tj. vzorčeni bodo naključni odstavki posameznih besedil, in bo na voljo v repozitoriju CLARIN.SI.

Korpus bo v repozitoriju CLARIN.SI na voljo tako v formatu TEI kot v formatu CONNL-U, saj je slednji preferenčni format pri nalogah, ki vključujejo nadaljnje procesiranje podatkov, npr. strojno učenje, luščenje podatkov ipd.

3.5. Tematska kategorizacija besedil

Ena od aktivnosti projekta SLED je tudi izdelava orodja za avtomatsko kategorizacijo besedil glede na tematiko. Za izdelavo takšnega orodja oz. modela zanj potrebujemo dvoje: klasifikacijo kategorij in učno množico.

[12] https://pypi.org/project/classla/
[13] https://www.clarin.si/kontext/
[14] https://www.clarin.si/noske/
[15] https://www.clarin.si/ske/

Pri izdelavi nabora kategorij smo se opirali na podatke iz treh skupin virov:

- slovenskih novičarskih portalov, izbrali smo jih šest, tj. rtvslo.si, delo.si, sta.si, dnevnik.si, 24ur.com in vecer.com.
- nabora tematskih kod oz. kategorij Mednarodnega tiskovnega telekomunikacijskega sveta (IPTC).[16] S tem smo tudi želeli zagotoviti čim boljšo usklajenost naših kategorij z mednarodnim standardom.
- kategorij v sodobnih sinhronih in spremljevalnih korpusih, pri čemer sta bila relevantna predvsem češki korpus SYN_2015 (Křen et al., 2016) in estonski nacionalni korpus (Koppel in Kallas, v tisku).

Glavno vodilo pri pripravi klasifikacije je bilo pripraviti relativno majhen nabor kategorij, v katere lahko uvrstimo vse novice na različnih portalih. S tem bi zagotovili tudi boljše delovanje modela. Posledično smo pri analizi uporabljenih virov več pozornosti posvečali krovnim kategorijam, kar je bilo sploh potrebno pri naboru IPTC, ki ima približno 1.400 kategorij, razdeljenih v tri nivoje (s tem da krovni nivo sestavlja le 17 kategorij). Za ponazoritev smiselnosti uporabe zgolj krovnih kategorij lahko vzamemo kategorijo šport, ki ima na večini novičarskih portalov nadaljnje kategorije, od katerih se vedno pojavita samo *nogomet* in *košarka*, ostale pa le na nekaterih portalih, npr. dnevnik.si nima *zimskih športov*, ima pa ločeno podstran za novice o *Luki Dončiću*; rtvslo.si je edini, ki ima podstran za novice o *Formuli 1*, 24ur.si ima ločene podstrani za *Ligo prvakov* in *Ligo Evropa* (nogomet) ter *borilne športe*.

Naša končna klasifikacija vsebuje 12 kategorij:

- **umetnost in kultura**. Vključuje besedila o kulturi, umetnosti, filmih, knjigah, gledališču, pa tudi recenzije ipd.
- **črna kronika**. Naravne in ostale nesreče, človeški delikti, kriminal.
- **gospodarstvo**. Vključuje besedila s področja ekonomije, trgov, financ, zaposlitev ipd.
- **okolje**. Zajema okoljevarstvo, planet, energente, tudi kmetijske teme.
- **zdravje**. Fizično in mentalno zdravje ljudi, medicina, farmacija, zdravstvena infrastruktura.
- **prosti čas**. Hobiji, rekreacija, potovanja, turizem, ljubljenčki, dom in družina, bivanje.
- **politika in pravo**. Mednarodne in nacionalne novice s področna državne uprave, pravnih postopkov in družbenih razmerij, konfliktov, vojn.
- **znanost in tehnologija**. Znanstvena odkritja, zanimivosti, tehnološke inovacije, informacijska tehnologija, računalništvo.
- **družba**. Družbena vprašanja in razmerja, enakost, diskriminacija, religija, etika ipd.
- **šport**. Športni rezultati in zanimivosti z različnih športnih področij.
- **vreme**. Meteorološke napovedi, opisi vremenskih posebnosti, stanj, procesov.
- **zabava**. Estrada, moda, slog.
- **izobraževanje**. Procesi posredovanja in pridobivanja znanja ter veščin. Vse stopnje izobraževanja, od vrtca do univerzitetnega izobraževanja, pa tudi vseživljenjsko učenje.

Kot prikazuje primerjalna Tabela 1, obstaja precejšnja prekrivnost tako s kategorijami novičarskih portalov kot s kategorijami IPTC in tujih korpusov. V nekaterih primerih, npr. *gospodarstvo, prosti čas, politika* in *družba*, naša kategorija zajema več kategorij ostalih virov. Tako ima za prosti čas estonski korpus kar sedem ločenih kategorij. Edini primer, ko se eno od kategorij tujih virov lahko uvrsti v dve naši, sta *umetnost in kultura* ter *zabava*. Kategoriji smo namreč ločili po eni strani zato, ker ima veliko slovenskih novičarskih portalov ločene podstrani zanju, po drugi strani pa zaradi samega jezika - kulturno-umetniške vsebine so za razliko od zabavnih pogosto precej bolj strokovne.

Medtem ko v naše kategorije lahko umestimo vseh 17 kategorij IPTC, pa češki oz. estonski korpus določenih kategorij nimata, npr. estonski nima *črne kronike*, češki pa ne *okolja, zdravja, znanosti in tehnologije* ter *zabave*. Oba tudi nimata ločene kategorije za *vreme*, ki pa jo ima IPTC in smo jo dodali zato, ker jo ima večina slovenskih novičarskih portalov.

Če pogledamo še prekrivnost kategorij s stranmi oz. podstranmi šestih slovenskih novičarskih portalov, vidimo, da so problematične kategorije predvsem *politika*, *družba* in *izobraževanje*. Gre za sicer legitmne kategorije, ki pa na novičarskih portalih nimajo svojih podstrani, temveč so novice razpršene po drugih podstraneh, ki so večinoma opredeljene glede na geografski izvor novice, npr. Slovenija, Svet, Lokalno. Medtem ko so se avtorji češkega korpusa odločili slediti takšni delitvi tudi pri kategorijah (*current events, foreign news, domestic news, regional news*), smo se mi raje držali tematike. To za izdelavo učnih množic pomeni nekoliko več ročnega dela oz. iskanje drugih kazalcev, s katerimi lahko odkrijemo tematiko prispevka na posameznem portalu. Izjema je portal sta.si, ki že ima ustrezne kategorije, in sicer *Šolstvo* in *Družba*, za politiko pa *Državni zbor*, *Evropska unija*, *Mednarodna politika, Slovenska notranja politika* in *Slovenska zunanja politika*.

Učne množice smo izdelali z mapiranjem kategorij različnih virov novic na našo interno kategorizacijo. Tako lahko besedila iz določenih kategorij konkretnih virov uporabimo za učenje modela. Pri pripravi učnih množic bomo vzorčili tako količino podatkov iz posameznega vira kot količino podatkov v kategoriji in s tem zagotovili raznolikosti učnih množic, pa tudi robustnost končnega modela.

Za modeliranje bomo uporabili orodje fasttext (Joulin et al, 2016) z vložitvami CLARIN.SI (Ljubešić in Erjavec, 2018) in model SloBERTa (Ulčar in Robnik-Šikonja, 2021). Glede na razliko v rezultatih (pričakujemo, da se bo model SloBERTa odrezal boljše, a morda razlika v rezultatih ne bo tako opazna) in kompleksnosti klasifikatorja (fasttext je precej hitrejši in zahteva bistveno manj spominskih kapacitet), bomo izbrali klasifikator, ki ga bomo uporabili na novih besedilih.

[16] https://cv.iptc.org/newscodes/subjectcode

kategorija	slovenski portali (6)	češki korpus	estonski korpus	IPTC
umetnost in kultura	5	culture	culture & entertainment	arts, culture and entertainment
črna kronika	6	crime	/	disaster and accident
gospodarstvo	6	economy	economy, finance & business; agriculture; construction & real estate	economy, business and finance; labour
okolje	2	/	nature & environment	environmental issue
zdravje	3	/	health	health
prosti čas	4	leisure	beauty; cars; food & drinks; gambling & casinos; home, family & children; pets and animals; travel & tourism; video games	lifestyle and leisure
politika in pravo	1	politics	politics & government	politics; crime, law and justice; unrest, conflicts and war
znanost in tehnologija	5	/	science, technology & IT	science and technology
družba	1	social life	society; religion; sex; women	social issue; religion and belief; human interest
šport	6	sports	sports	sport
vreme	4	/	/	weather
zabava	4	/	culture & entertainment*	arts, culture and entertainment*
izobraževanje	1	/	education	education

Tabela 1: Primerjava tematskih kategorij projekta SLED z domačimi novičarskimi portali in tujimi viri.

3.6. Rezultati uporabniške ankete

Ker je Trendi prvi korpus svoje vrste v slovenskem okolju, smo ga želeli zasnovati karseda skladno z uporabniškimi pričakovanji. Ta smo v decembru 2021 preverili s pomočjo uporabniške ankete, s katero smo ugotovili, katerih podatkov o aktualni rabi jezika si raziskovalna skupnost želi in v kakšni obliki (npr. različni seznami, kot so kandidati za neologizme, besede in besedne zveze z najbolj izstopajočo rabo v določenem obdobju (dnevu, tednu, mesecu), izstopajoče besede in besedne zveze glede na vir ipd.).

Anketa[17] je bila izdelana na platformi 1KA, sestavljena pa je bila iz 9 vprašanj: med temi je bilo 5 vsebinskih, 4 pa so zbirala demografske podatke (spol, starost, področje delovanja). Diseminirana je bila po e-poštnih seznamih slovenskih jezikoslovnih raziskovalnih skupnosti (npr. SlovLit ter e-poštni seznam Slovenskega društva za jezikovne tehnologije) ter po družbenem omrežju Facebook (na uradni strani Centra za jezikovne vire in tehnologije Univerze v Ljubljani ter v neformalnih jezikoslovnih uporabniških skupinah, kot je *Prevajalci, na pomoč!*).

V celoti izpolnjenih vprašalnikov je bilo 100. Vzorec, ki ga je zajela anketa, zajema predvsem osebe ženskega spola (82 %), manjši delež pa je moških (18 %). Po starosti vzorec zajema predvsem generacije med 26. in 55. letom starosti (80 % vseh udeleženk_cev), največ med 26. in 35. letom (33 %) in med 46. in 55. letom (32 %). Večina

[17] Podrobnejše poročilo o izvedeni anketi je na voljo na spletni strani projekta: https://sled.ijs.si/wp-content/uploads/2022/02/SLED_anketa_porocilo_2022-2-03_final.pdf

udeleženk_cev je zaposlenih bodisi v javnem sektorju (61 %) bodisi je samozaposlena (20 %), le manjši delež ima še študentski status (3 %) ali pa so zaposleni v podjetjih (6 %), upokojeni (4 %) ali v iskanju zaposlitve (5 %). Po področju delovanja, pri katerem so udeleženci_ke lahko izbrali_e več možnosti, prednjačita lektoriranje (60 %) in prevajanje (46 %), visok delež pa imajo tudi ljubiteljsko raziskovanje jezika (38 %), strokovno in znanstveno pisanje (34 %), jezikoslovne raziskave (32 %) ter kreativno pisanje in blogerstvo (22 %). Skupno 40 % zajemajo tudi različne kategorije poučevanja jezika (slovenščina kot 1. jezik na osnovni ali srednji šoli, slovenščina kot 2. ali tuji jezik, jezikoslovni predmeti na višji/univerzitetni ravni). Vzorec nakazuje, da je anketa zajela različna področja jezikoslovno-raziskovalnega udejstvovanja.

V nadaljevanju predstavljamo podrobnejšo analizo odgovorov na vsebinska vprašanja.

3.6.1. Scenariji uporabe in uporabniško zanimanje

Anketiranci_ke so navedli_e, kateri podatki v orodju, ki bi spremljalo aktualno jezikovno rabo, bi jih najbolj zanimali, in pri vsakem od 6 predlaganih scenarijev uporabe (s konkretnimi primeri za lažjo predstavo) ocenili_e svojo stopnjo zanimanja (1 - sploh me ne zanima, 5 - zelo me zanima). Med scenariji so npr. *katere besede/besedne zveze so najznačilnejše za določeno obdobje v primerjavi z drugim obdobjem*? (npr. katere besede so se mnogo pogosteje uporabljale v februarju 2020 kot pa v februarju 2021); *v katerem obdobju je določena beseda/besedna zveza najpogostejša*? (npr. ali je bila beseda "tajkun" res najpogostejša v obdobju 2008-2009?); *ali raba besede/besedne zveze v zadnjem obdobju glede na trende narašča ali pada?* (npr. ali se "epidemija" uporablja vse pogosteje ali vse redkeje?).

Rezultati kažejo, da se anketirancem_kam vsi predlagani scenariji zdijo zanimivi: kategoriji "Zanima me" (4) in "Zelo me zanima." (5) namreč pri vsakem scenariju skupaj zajemata med 74 in 88 %. Po stopnji zanimanja najbolj izstopa scenarij, v katerem je mogoče primerjati trend rabe dveh ali več besed/besednih zvez (npr. *anticepilec* vs. *proticepilec*), enako pa anketiranke_ce zanima tudi, ali raba določene besede/besedne zveze v zadnjem obdobju glede na trende narašča ali pada.

Dobre tri četrtine vprašanih (76 %) je odgovorilo, da bi jim podatki o aktualni jezikovni rabi koristili pri delu, le 9 % tovrstni podatki ne bi koristili (15 % je neodločenih). Rezultati ankete torej potrjujejo, da jezikoslovno skupnost podatki o trendih jezikovne rabe zanimajo in da obstaja realna potreba po jezikovnem viru, ki tovrstne podatke prinaša sprotno in ažurno.

3.6.2. Načini prikaza podatkov

Na lestvici od 1 (sploh ni pomembno) do 5 (zelo pomembno) so anketiranci_ke ocenili_e tudi, kateri od predlaganih načinov prikaza podatkov (grafi s trendi jezikovne rabe. tabele s številskimi podatki, seznami besed oz. besednih zvez z naraščajočo/padajočo rabo, drugo) se jim zdijo pomembni. Če združimo deleže kategorij "pomembno" (4) in "zelo pomembno" (5), dobimo deleže 79 % za grafe, 64 % za tabele s številskimi podatki in 87 % za sezname besed s padajočo/naraščajočo rabo. Anketiranke_ce torej najbolj zanimajo preprosti seznami, najmanj pa napredne tabele s številskimi podatki.

3.6.3. Uporabniški predlogi

V odprtem vprašanju so imeli anketiranci_ke možnost izraziti predloge oz. dodatne scenarije, ki bi jih zanimali o aktualni jezikovni rabi. Dodatnih predlogov je bilo 15. Nanašajo se npr. na povezljivost orodja z drugimi jezikovnimi viri (npr. integracija v Slovenski oblikoslovni leksikon Sloleks in v korpus pisne standardne slovenščine Gigafida) in dostop do podatkov (npr. možnost dostopa do podatkov preko javnega API-ja), primerjavo sopomenskih različic besed oz. besednih zvez (npr. *oče* vs. *ata*), vključitev zgledov rabe in spremljanje jezikovne rabe daljših enot (npr. frazemov). Večina dodatnih predlogov sicer presega obseg projekta SLED, a predstavljajo pomembno povratno informacijo za razmislek o prihodnjem razvoju in integraciji spremljevalnega korpusa in iz njega izluščenih podatkov v ostale jezikovne vire.

4. Sklep in nadaljnje delo

V prispevku smo predstavili različne aktivnosti projekta SLED, s poudarkom na korpusu Trendi, nastajajočem spremljevalnem korpusu slovenskega jezika. Opisali smo metodologijo njegove izdelave, vsebino in oblike, v katerih je na voljo uporabnikom_cam. Predstavili smo tudi klasifikacijo tematskih kategorij, ki je bila oblikovana za namene izdelave modela za avtomatsko tematsko kategorizacijo besedil. Zadnji del je bil namenjen predstavitvi rezultatov ankete o uporabniških pričakovanjih o podatkih o aktualni rabi jezika, ki jih želi imeti zainteresirana skupnost.

V prihajajočih mesecih bomo nadaljevali z objavami mesečnih različic korpusa, pripravili prve statistične izračune in dokončali ter evalvirali algoritem za avtomatsko kategorizacijo besedil. Pomembno je, da smo veliko časa posvetili vzpostavitvi avtomatskih postopkov priprave besedil in izračunov, saj bo to pospešilo posodabljanje podatkov v konkordančnikih in na repozitorju CLARIN.SI.

Prav tako je ključna aktivnost izboljšava postopka pridobivanja besedil, ki bo poskrbela, da bodo odpravljene določene pomanjkljivosti trenutne metode. Ker bo vzpostavljena tesna povezanost med korpusom Trendi in referenčnim korpusom Gigafida, bo vsaka izboljšava postopkov koristila obema korpusoma.

S korpusom Trendi je slovenska jezikovna infrastruktura bogatejša za pomemben vir, ki bo relevanten tako za raziskovalno skupnost kot širšo javnost.

5. Zahvala

Projekt SLED (*Spremljevalni korpus in spremljajoči podatkovni viri*) financira Ministrstvo za kulturo Republike Slovenije kot del *Javnega razpisa za (so)financiranje projektov, namenjenih gradnji in posodabljanju infrastrukture za slovenski jezik v digitalnem okolju 2021–2022*. Raziskovalna programa št. P6-0411 (*Jezikovni viri in tehnologije za slovenski jezik*) in št. P6-0215 (*Slovenski jezik - bazične, kontrastivne in aplikativne raziskave*) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

6. Literatura

Mark Davies. 2008–. The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/.

Mark Davies. 2016–. Corpus of News on the Web (NOW). Available online at https://www.english-corpora.org/now/.

Mark Davies. 2019–. The Coronavirus Corpus. Available online at https://www.english-corpora.org/corona/.

Kaja Dobrovoljc, Tomaž Erjavec in Simon Krek. 2017. The Universal Dependencies Treebank for Slovenian. V: *Proceedings of the 6thWorkshop on Balto-Slavic Natural Language Processing*, BSNLP@EACL 2017, str. 33–38.

Tomaž Erjavec, Darja Fišer, Simon Krek in Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. V: P*roceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).*

Ondrej Herman. 2013. *Automatic methods for detection of word usage in time.* Diplomska naloga. Masaryk University, Faculty of Informatics.

Armand Joulin, Edouard Grave, Piotr Bojanowski in Tomas Mikolov. 2016. *Bag of Tricks for Efficient Text Classification*. arXiv. https://arxiv.org/abs/1607.01759.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz in David Tugwell. 2004. The Sketch Engine. V: G. Williams in S. Vessier, ur., *Proceedings of the Eleventh EURALEX International Congress, Lorient, France*, str. 105-116. Lorient: Université de Bretagne Sud.

Kristina Koppel in Jelena Kallas. (v tisku). *Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu.* Eesti Rakenduslingvistika Ühingu aastaraamat.

Iztok Kosem, Simon Krek, Polona Gantar, Špela Arhar Holdt in Jaka Čibej. 2021. Language monitor: tracking the use of words in contemporary Slovene. V: I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek in C. Tiberius , ur., *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference. 5–7 July 2021, virtual*, str. 514–527. Brno: Lexical Computing CZ, s.r.o., https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_33_pp514-528.pdf.

Simon Krek et al. 2019. *Corpus of Written Standard Slovene Gigafida 2.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1320.

Luka Krsnik et al. 2019. *Corpus extraction tool LIST 1.2*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1276.

Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička in Adrian Jan Zasina. 2016. SYN2015: Representative Corpus of Contemporary Written Czech. V: P*roceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, str. 2522–2528, Portorož, Slovenia. European Language Resources Association (ELRA).

Nikola Ljubešić in Kaja Dobrovoljc. 2019. 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.

Nikola Ljubešić in Tomaž Erjavec. 2018. *Word embeddings CLARIN.SI-embed.sl 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1204.

Nataša Logar, Tomaž Erjavec, Simon Krek, Miha Grčar, in Peter Holozan. 2013. *Written corpus ccGigafida 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1035.

Tomáš Machálek. 2020. KonText: Advanced and Flexible Corpus Query Interface. V: Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, str. 7003–7008. https://www.aclweb.org/anthology/2020.lrec-1.865

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton in Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Pavel Rychlý. 2007. Manatee/Bonito-A Modular Corpus Manager. V: RASLAN, str. 65–70.

Mitja Trampuš in Blaž Novak. 2012. The Internals Of An Aggregated Web News Feed. V: *Proceedings of 15th Multiconference on Information Society 2012* (IS-2012). http://ailab.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf.

Matej Ulčar in Marko Robnik-Šikonja. 2021. SloBERTa: Slovene monolingual large pretrained masked language model. V: *Proceedings of the 24th International Multiconference – IS2021* (SiKDD). https://ailab.ijs.si/dunja/SiKDD2021/Papers/Ulcar+Robnik.pdf.

Katja Zupan, Nikola Ljubešić in Tomaž Erjavec. Smernice za označevanje imenskih entitet v slovenskem jeziku. https://www.clarin.si/repository/xmlui/bitstream/handle/11356/1238/SlovenianNER-slv-v1.0.pdf?sequence=7&isAllowed=y.

Automatic Text Analysis in Language Assessment: Developing a MultiDis Web Application

Sara Košutar*, Dario Karl‡, Matea Kramarić*, Gordana Hržica*

* Faculty of Education and Rehabilitation Sciences, University of Zagreb
University Campus Borongaj, Borongajska cesta 83 f, 10 000 Zagreb
sara.kosutar@erf.unizg.hr
matea.kramaric@erf.unizg.hr
gordana.hrzica@erf.unizg.hr
‡Department of Data Science, InSky Solutions
Medačka ulica 18, 10 000 Zagreb
dario.karl.sl@gmail.com

Abstract

Language sample analysis provides rich information about the language abilities in the written or spoken text produced by a speaker in response to a language task. Language sample analysis is generally used to assess the abilities of children during language acquisition, but also the abilities of adult speakers across the lifespan. Its wide range of uses also allows for the assessment of language abilities in educational contexts such as second language acquisition or fluency, the abilities of bilingual speakers in general, and it is also used for diagnosis in speech and language pathology. Various computer programs have been developed to assist in the language sample analysis. However, these programs have been developed mainly for English and are often not fully open-access or do not provide data on population metrics, history of data uploaded by a user, and/or improvements in basic language measures. The time needed for transcription and the linguistic knowledge required for manual analysis are considered to be the main obstacles to its implementation The goal of this paper is to present a web-based application MultiDis intended for the analysis of language samples at the microstructural text level in Croatian. The application is still under development, but the current version fulfils its main purpose – it enables the (semi-) automatic calculation of measures reflecting language productivity, lexical diversity, syntactic complexity, and discourse cohesion in spoken language, and provides users with socio-demographic and linguistic metadata as well as the history of uploaded transcripts. We will present the challenges we have faced in developing the application (e.g., annotation system, text standardisation), future improvements we plan to make to the application (e.g., syntactic parsing, speech-to-text, multilingual analysis), and the possibilities of its use in the wider scientific and professional community.

1. Language sample analysis

Language sample analysis provides rich information about the language abilities in the written or spoken text produced by a speaker in response to a language task, e.g. storytelling, written essay, description of a picture, answering questions, etc. It is an ecologically valid means of language assessment that can be used along with standardised language tests because it provides data that tests cannot. Compared to standardised tests, language sample analysis has greater ecological validity because it reflects the natural everyday situation of language production. Consequently, it allows for a more in-depth analysis of specific morphosyntactic, semantic, and pragmatic features. Due to its lower bias, it proved to be more suitable for studying regional variations and dialects compared to standard questionnaires (e.g., Samardžić and Ljubešić, 2021). Language sample analysis is generally used to assess children's abilities during language acquisition, but also adult speakers' abilities across the lifespan (e.g., Westerveld et al., 2004). Its wide range of uses allows for the assessment of language abilities in educational contexts such as second language acquisition or fluency (e.g., Clercq and Housen, 2017), the abilities of bilingual speakers in general (e.g., Gagarina et al., 2016), and it is also used for diagnosis in speech and language pathology (e.g., Justice et al., 2006). This type of analysis is widely used in some countries, but in many countries, scientists and professionals are unaware of its benefits or find it too complex and time-consuming (see Heilmann, 2010; Klatte et al., 2022).

The process of collecting language samples involves several steps. First, a speaker is given a language task, for example telling a story based on a picture, and is recorded while performing this task. The recordings are then transcribed using special codes and are divided into smaller units of analysis, e.g., communication units (C-units; see Labov and Waletzky, 1967). Special codes mark different features of the spoken language or deviations (e.g., repetitions, omissions of vowels, use of regionally marked words, morphosyntactic errors, etc.). When written language samples are collected, the speaker responds to the task in writing, but all further steps are the same. Once the transcripts are produced, they can be analysed in various computer programs that enable (semi-)automatic calculation of different language measures.

Language sample analysis provides information about language abilities at two levels of text structure (Gagarina et al., 2012). First is the microstructural level, which refers to the internal linguistic organisation and includes text length, vocabulary use, morphosyntax, cohesive devices, etc. At the microstructural level, one can observe, for example, which language structures have emerged during language acquisition or how complex they are in terms of their internal features. The macrostructural analysis allows for assessing the ability of the hierarchical organisation of the text (e.g., in storytelling, whether the speaker has expressed a goal, an attempt, an outcome, etc.). At the macrostructural level, one can examine how successfully a speaker connects sentences according to a language task. By examining these elements, one gains insight into the quality of an individual's language when performing a particular language task, but also indirectly information on her or his language skills in general.

1.1. Language measures

Different aspects of microstructure correspond to several dimensions, such as productivity, lexical diversity, and syntactic complexity. A set of (semi-)automatic measures has been proposed to assess language abilities at the microstructural level. Productivity refers to the amount of language (words or utterances) produced (Leadholm and Miller, 1992). Measures of productivity include the total number of C-units or the total number of words (TNW). C-units are often used instead of utterances in spoken language analysis (see MacWhinney, 2000). The basic criteria for dividing a sequence of spoken words into utterances are intonation and pauses. However, transcribers may rate the utterances differently against these criteria, which results in lower inter-rater reliability (Stockman, 2010). C-units consist of one or more clauses. A clause is any syntactic unit consisting of at least one predicate. A complex sentence with one or more dependent clauses constitutes one C-unit, while a compound sentence is divided into two or more C-units, depending on the number of independent clauses. Studies have shown that measures of productivity can distinguish children with typical language status from children with developmental language disorders (DLD; Wetherell et al., 2007), bilingual from monolingual children (Hržica and Roch, 2021), and adult speakers according to their language skills (Nippold et al., 2017).

Measures of lexical diversity are used to assess vocabulary abilities. The more diverse the vocabulary produced, the greater the lexical diversity. Measuring lexical diversity is more complex and therefore methodologically challenging. Traditional measures include the number of different words (NDW; Miller, 1981) and the type-token ratio (TTR; Templin, 1957). Types and tokens can be easily calculated automatically, whereas lemmas are more difficult to calculate automatically, and require specialized natural language processing tasks. In particular, this requires morphological analyses such as lemmatisation, part-of-speech (POS) tagging, or morphological segmentation. In languages with rich morphology, the lemma-token ratio would be more appropriate, but due to the time-consuming nature of the task, this has rarely been done (see Balčiūnienė and Kornev, 2019). Another problem with measures of lexical diversity and measures of productivity is that they are affected by the length of a language sample (Malvern et al., 2004; McCarthy, 2005).

To overcome these limitations, alternative measures have been developed, such as D (Malvern and Richards, 1997) and moving average type-token ratio (MATTR; Covington and McFall, 2010). The measure D is based on modelling the decrease in TTR with the increasing size of the language sample using mathematical algorithms. MATTR calculates TTR for text windows of a fixed size, e.g., 500 words. The window moves through the text and calculates TTR for words 1-501, 2-502, etc. At the end of the text, all TTRs are averaged to determine the final score. However, it is not yet clear which of these measures provides more reliable results, as the results of validation studies vary (see deBoer, 2014; Fergadiotis et al., 2015). Regardless of methodological limitations, these measures can distinguish the abilities of children and adults with typical language status from children or adults with DLD (e.g., Hržica et al., 2019; Kapantzoglou et al., 2019). Measures of lexical diversity have also been found to correlate with standardised vocabulary tests in bilingual children (e.g., Hržica and Roch, 2021).

Syntactic complexity refers to the range of syntactic structures and the degree of sophistication of these structures in language production (Ortega, 2003). It is usually measured by calculating the average length of the C-unit. The length of the C-unit increases when there is a dependent clause or when the syntax within the clause is more complex, for example when the clause is extended by adding attributes, appositions, or adjectives. Measures of syntactic complexity have been shown to distinguish between different groups of speakers, including children with DLD and adults of different ages (e.g., Rice et al., 2010; Nippold et al., 2017). In addition to the average length of syntactic units, other measures of syntactic complexity include clausal density (i.e., the total number of main and subordinate clauses divided by the total number of C-units) and mean length of clause (main or subordinate), and they are also commonly used (e.g., Scott and Stokes, 1995; Norris and Ortega, 2009). Because of the variety of measures and the different methods of calculation, little is known about which measures are appropriate concerning typological differences between languages, and some of these measures are not always automatic.

In the last decades of the 20th century, various computer programs have been developed to support language sample analysis (overview: Pezold et al., 2020), but they are often not user-friendly. More recently, web-based programs have been introduced that allow for the analysis of language use at different linguistic levels (e.g., Coh-Metrix; McNamara et al., 2014). The measures are based on basic calculations (e.g., TTR, MLU), but there are also advanced measures based on language technologies such as the annotation of morphological, syntactic, and semantic features. Such applications are mainly developed for English or other widely spoken languages and are often not fully open-access. There is an increasing awareness of the importance of language sample analysis as a complementary method in language assessment. The time needed for transcription and the linguistic knowledge required for manual analysis are considered to be the main obstacles to its implementation (Pezold et al., 2020). Therefore, the development of a tool for the automatic calculation of language measures could make naturalistic language assessment more feasible.

2. Goal of the paper

The goal of this paper is to present a web-based application MultiDis, intended for the analysis of language samples at the microstructural level in Croatian, which enables the (semi-)automatic calculation of measures reflecting language productivity, lexical diversity, syntactic complexity, and discourse cohesion in spoken and written language. We will present the challenges we have faced in developing the application, future improvements we plan to make to the application, and the possibilities of its use in the wider scientific and professional community.

3. Development of the MultiDis web application

Existing computer-based resources used to analyse children's or adults' language abilities are either developed for English only or do not provide data on population

metrics, history of data uploaded by a user, and/or improvements in basic language measures such as NDW or TTR. The Computerized Language Analysis (CLAN; MacWhinney, 2000), for example, is a freely available desktop application whose users are expected to have a high level of language and transcription expertise. Text Inspector (2018), on the other hand, is a web-based application, but it is only designed for the text analysis of the English language and the target users are mainly first or second language acquisition teachers. We aim to develop a web-based application that fosters the analysis of language samples in Croatian. Our target users work at least partly with spoken language (e.g., language diagnostics performed by speech and language pathologists), so the application should support both written and spoken language analysis. The application is currently being developed, and we will present the coding system, language resources, data collection and language measures that have been implemented so far.

3.1. Annotation codes

Considering that our target users mostly work with spoken language, there are several codes which can be used to annotate the data. Computer programs for language analysis such as CLAN (MacWhinney, 2000) have an entire system of very specific annotation codes. In the MultiDis web application, a new and simpler system of annotation codes was developed to provide a faster and more organised annotation process. The system of the codes was designed to include several categories with individual codes and subsets of codes. The main idea is to have a system of annotation codes that can be changed over time according to the following criteria:

hierarchical (with categories and subcategories of codes)
extensible (adding new categories and codes)
easily customizable system (each category has a recognizable first character).

To date, the following categories have been established: *phonotactic codes* include conversation markers and elements of communication; *citation codes* indicate references to another utterance within the language sample; *phonetic codes* indicate pronunciation and other elements specific to spoken language; *sociolinguistic codes* indicate dialectisms, neologisms, foreign words, etc.; *correction codes* indicate errors made at a particular level of linguistic structure – phonological, morphosyntactic and/or lexical. There is also an additional code for corrections – a marker that can be used to exclude a particular segment from the transcript and provide a correct or standardised form that the application will use to standardise any text before moving on to a later stage of language analysis. A full description of codes is available on the web page of the application: http://www.multidis.com.hr/statistics/.

An example of multiple annotation codes would be a sentence in (1), that would look like (2) in the following uploaded transcript. Angle brackets point to a segment that needs to be excluded and round brackets point to a 'standardised' form of that segment. In addition, the @d code preceding the token *ćuko* 'dog' refers to a dialectism. The application will convert the sentence in (2) into the standardised form or the sentence as in (3), mapping the dialectism and providing this information in the final analysis report.

(1) *Dečko i ćuko su ulovili žabicu*. 'The boy and the dog caught the frog'
(2) *Dečko i <ćuko> (@d pas) su ulovili žabicu*. 'The boy and the dog caught the frog'
(3) *Dečko i pas su ulovili žabicu*. 'The boy and the dog caught the frog'

The annotation system and parsing rules for the transcripts were implemented using common Regular expressions (regex) in Python (Van Rossum, 2020). Regular expressions allow the system to recognise specific codes, save the data and convert the language into a standard form, so that existing language resources, such as tokenizers and lemmatizers, achieve a higher hit rate and precision. After annotation and parsing, the application will provide a standardised language text on which further language sample analysis is performed.

3.2. Language resources

The next step in the development of the application was the integration of an open-source Python library. We started with Stanza (Qi et al., 2020) to solve the following tasks common in natural language processing:

lemmatisation
POS tagging
syntactic parsing (sentence and clause segmentation).

In the early stages of developing the MultiDis web application, one of the main linguistic resources used was Stanza, a Python natural language processing toolkit for human language developed at Stanford University (Qi et al., 2020). Stanza enables quick out-of-the-box processing of multilingual texts. Since we plan to test our use case – based on the analysis of children's spoken language – on multiple languages, Stanza has an advantage over several other natural language processing models, frameworks and neural pipelines, such as Podium (Tutek et al., 2021), CLASSLA (Ljubešić and Dobrovoljc, 2019) or BERTić (Ljubešić and Lauc, 2021). Lemmatisation and POS tagging are fairly accurate (> 85 % of the cases), as they do not interfere with the computation of currently implemented language measures, though the process of delimiting the boundaries of C-units has been an obstacle that is currently being resolved. We are also exploring other options and planning further analysis and accuracy testing for this task. Since the language samples that the application will analyse are non-literary texts, we also plan to explicitly compare the aforementioned tools in the tasks of lemmatisation, POS tagging and morphosyntactic description (MSD) using our datasets to improve the application's baseline accuracy in these tasks. The standard for POS tagging is MulTextEast language resources (Erjavec, 2010), version 4 for the Croatian language. In this way, a token *ćuko* 'dog' is annotated as a dialectism using the annotation codes for the transcript parsing, and the standardised form *pas* 'dog' receives a morphosyntactic tag *Ncmsn* (nominative case, common noun, masculine, singular).

3.3. Data collection – manual annotation of transcripts with the new coding system

In the next step of developing the MultiDis web application, it was important to test the annotation system and the parsing of the language samples, as the aim was to obtain a standardised text with the data on the participants'

socio-demographic and language characteristics, parsed with the appropriate annotation codes and available to the user along with the morphosyntactic data. Before running the analysis, the texts were manually transcribed by students and volunteers within the courses *Computer Analysis of Child Language* and *Volunteering* at the Department of Speech and Language Pathology at the Faculty of Education and Rehabilitation Sciences, University of Zagreb. The test transcripts are the result of a storytelling task, mostly *Frog where are you?* (Mayer, 1969) and *Multilingual Assessment Instrument for Narratives* (MAIN; Gagarina et al., 2012; Gagarina et al., 2019; Hržica and Kuvač Kraljević, 2012, 2020). After the implementation of annotation codes, these transcripts have been successfully standardised and prepared for the final analysis. Any other transcript can be uploaded to the application and the user can only receive data about their uploaded transcripts and not about the transcripts of other users.

3.4. Automation of language measures

Using the standardised text and the provided language data from the previous step in the analysis, the next task of the MultiDis web application is to provide users with a detailed analysis of language measures. It is important to note that the measures are currently calculated intertextually, but we plan to compare the individual results with the population results, as well as with the baseline data. The application incorporates diverse measures that can be used in the language assessment such as productivity, lexical diversity, syntactic complexity and discourse cohesion. The list of language measures included in the MultiDis web application is available in Table 1.

Category	Measure	Description
Language productivity	Number of communication units (NCU)	The total number of communication units
	Total number of words (TNW)	The total number of tokens (repeated tokens are excluded)
	Number of different words (NDW)	The total number of word forms – types
Lexical diversity	Type-token ratio (TTR)	The total number of tokens divided by the total number of types
	Index of lexical diversity D*	Based on the VOCD algorithm calculates the probability of the next token in a sequence based on an arbitrarily chosen *n*-token sample from the text
	Moving average type-token ratio (MATTR)	Based on a window length pre-defined by a user, the text is divided into segments and for each window length, the TTR is calculated – the average TTR ratio of each segment is the measure of MATTR
Syntactic complexity	Mean length of the communication unit	The total number of words is divided by the total number of communication units
	Clausal density	The total number of main and subordinate clauses is divided by the total number of communication units
	Mean length of clause	The total number of tokens is divided by the total number of clauses
Discourse cohesion	Ratio of connectives	The total number of connectives is divided by the total number of C-units.

	Ratio of different connectives**	The total number of one type of connective is divided by the total number of all other types of connectives in the text

Table 1: List of language measures implemented in the MultiDis web application (*being tested; **in the process of implementation).

The process of automatic analysis of language measures is based on precise segmentation of C-units and clauses, as well as on the results of tokenisation and lemmatisation. Each simple sentence (e.g., *The dog is playing with the frogs*), each complex sentence containing a subordinate clause or a parenthetical phrase (e.g., *When the dog chased the cat away, the birds were happy*), and each clause of a compound sentence was considered as one C-unit (e.g., *One goat is in the water and the other is grazing grass*). Given the fact that we need 100% accuracy on this task, at this stage, we are still in the process of developing an automatic way of detecting connectives in the text as well as clause delimiters. Thus, a user still has to manually divide the text into C-units following the above-mentioned criteria before uploading a language sample to the application. This also means that the user can change any automatically parsed C-unit. Collecting a larger amount of data will make it possible to train and apply an appropriate machine learning model to enable automatic segmentation of C-units and clauses.

At the current stage of developing the application, a user can obtain the results of all available language measures based on C-unit segmentation, as well as the morphosyntactic data and the data provided by the annotation codes. It is important to note that the MATTR measure does not have a fixed window length; instead, there is a default window size that contains 10% of the total number of tokens, and the user can manually adjust the window size. In this way, we have avoided the possibility for the results on MATTR to be the same as the results on TTR for language samples with less than 500 tokens, and we have allowed the user to define the best window size for this measure. Measure D and the number of different connectives are currently being implemented and tested before these results are made available to users. The remaining measures listed in Table 1 have been successfully implemented.

4. Technical specifications of the MultiDis web application

The MultiDis web application is deployed on the Croatian Academic and Research Network (CARNET) server as a monolithic Docker service. All requests are first forwarded to a Nginx service for the static files and only then to the application itself via a Gunicorn service (Python Web Server Interface Gateway HTTP Server). The application and the entire backend logic are written in the Python programming language (Van Rossum, 2020) within the Django web framework. All data is stored in a MySQL database instance on the server. As mentioned earlier, a Stanza PyTorch model (Qi et al., 2020) is run with the application to infer the language data and provide morphosyntactic information. Other open-source libraries and packages used are python-docx, NumPy and Pandas.

The application is designed so that each segment can be improved, without compromising our main goals or the user's experience. In this sense, we can also include written language samples and provide new annotation codes and categories for written language or implement measures that are only used in the analysis of adult language. Lemmatisation and POS tagging can be improved by replacing the existing model with a new, customized and open-source model that can be extended to languages other than Croatian.

5. Future extensions

The MultiDis web application is still under development, but the current version fulfils its main purpose – it allows for (semi-)automatic analysis of spoken language, and provides users with socio-demographic and linguistic metadata as well as the history of uploaded transcripts. In addition to the implementation of a service for the automatic determination of C-units and clause boundaries, additional data will be made available to users, such as the analysis of Croatian dialects and reference data for language measures, at least for some populations and some text types. Several other options are also being considered, such as fully automatic parsing of the original language sample without the manual annotation codes and an experimental speech-to-text service. As the tools and resources to develop this application are also available for other languages, the application could be scaled for multilingual analysis, preferably in collaboration with other researchers.

6. Conclusion

The MultiDis web application is freely available at http://www.multidis.com.hr/ and can be used by linguists, speech and language pathologists, teachers etc., to assess the language abilities of both children and adult speakers of Croatian. It can help clinicians and educators in language sample analysis by resolving some of the main obstacles to its use. A simpler coding system fosters transcription and future development of speech-to-text could ease this process even further. Automatic lemmatisation and morphological tagging save time and enable more precise calculation of language measures. The language measures included in the application were selected based on previous research and adequately reflect the different aspects of the participants' language abilities. Therefore, the MultiDis web application supports its users by reducing both the transcription time and the linguistic knowledge required to technically perform the analysis.

7. Acknowledgements

This work was supported by the Croatian Science Foundation under the project entitled *Multilevel approach to spoken discourse in language development*

(UIP-2017-05-6603), by the Arts and Humanities Research Council under the project entitled *Feast and Famine Project: Confronting Overabundance and Defectivity in Language* (AH/T002859/1) and by the COST Action under the project *NexusLinguarum – European network for Web-centred linguistic data science* (CA18209). Sara Košutar was supported by the project *Young Researchers' Career Development project – Training of New Doctoral Students*. Any opinions, findings, conclusions, or recommendations presented in this manuscript are those of the author(s) and do not necessarily reflect the views of the Croatian Science Foundation.

8. References

Ingrida Balčiūnienė and Aleksandr N. Kornev. 2019. Evaluation of narrative skills in language-impaired children. Advantages of a dynamic approach. In: E. Aguilar-Mediavilla, L. Buil-Legaz, R. López-Penadés, V. A. Sanchez-Azanza and D. Adrover-Roig, eds., *Atypical Language Development in Romance Languages*, pages 127–414. John Benjamins Publishing Company, Amsterdam and Philadelphia.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Bastien de Clercq and Alex Housen. 2017. A Cross-Linguistic Perspective on Syntactic Complexity in L2 Development: Syntactic Elaboration and Diversity. *The Modern Language Journal*, 101(2):315–334.

Fredrik deBoer. 2014. Evaluating the comparability of two measures of lexical diversity. *System*, 47:139–145.

Tomaž Erjavec. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2544–2547, Valletta, Malta.

Gerasimos Fergadiotis, Heather Harris Wright and Samuel B. Greenc. 2015. Psychometric Evaluation of Lexical Diversity Indices: Assessing Length Effects. *Journal of Speech, Language, and Hearing Research*, 58(3):840–852.

Natalia Gagarina, Daleen Klop, Sari Kunnari, Koula Tantele, Taina Välimaa, Ingrida Balčiūnienė, Ute Bohnacker, and Joe Walters. 2012. MAIN: Multilingual assessment instrument for narratives. *ZAS Papers in Linguistics*, 56:1–155.

Natalia Gagarina, Daleen Klop, Sari Kunnari, Koula Tantele, TainaVälimaa, Ute Bohnacker, and Joel Walters. 2019. MAIN: Multilingual Assessment Instrument for Narratives – Revised. *ZAS Papers in Linguistics*, 63:1–21.

Natalia Gagarina, Daleen Klop, Ianthi M. Tsimpli, and Joel Walters. 2016. Narrative abilities in bilingual children. *Applied Psycholinguistics*, 37(1):11–17.

John J. Heilmann. 2010. Myths and Realities of Language Sample Analysis. *Perspectives on Language Learning and Education*, 17(1): 4–8.

Gordana Hržica, Sara Košutar, and Matea Kramarić. 2019. Rječnička raznolikost pisanih tekstova osoba s razvojnim jezičnim poremećajem [Lexical diversity in written texts of persons with developmental language disorder]. *Hrvatska Revija za Rehabilitacijska Istraživanja*, 55(2):14–30.

Gordana Hržica and Jelena Kuvač Kraljević. 2012. MAIN – hrvatska inačica: Višejezični instrument za ispitivanje pripovijedanja [MAIN – Croatian version: Multilingual Assessment Instrument for Narratives]. *ZAS papers in linguistics*, 56:201–218.

Gordana Hržica and Jelena Kuvač Kraljević. 2020. The Croatian adaptation of the Multilingual Assessment Instrument for Narratives. *ZAS Papers in Linguistics*, 64:37–44.

Gordana Hržica and Maja Roch. 2021. Lexical diversity in bilingual speakers of Croatian and Italian. In: S. Armon-Lotem and K. K. Grohmann, eds., *LITMUS in Action: Cross comparison studies across Europe*, pages 100–129. John Benjamins Publishing Company Trends in Language Acquisition Research (TILAR), Amsterdam.

Laura M. Justice, Ryan P. Bowles, Joan N. Kaderavek, Teresa A. Ukrainetz, Sarita L. Eisenberg, and Ronald B. Gillam. 2006. The Index of Narrative Microstructure: A Clinical Tool for Analyzing School-Age Children's Narrative Performances. *American Journal of Speech-Language Pathology*, 15(2):177–191.

Maria Kapantzoglou, Gerasimos Fergadiotis, and Alejandra Auza Buenavides. 2019. Psychometric evaluation of lexical diversity indices in Spanish narrative samples from children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research*, 62(1):70–83.

Inge S. Klatte, Vera van Heugten, Rob Zwitserlood, and Ellen Gerrits. 2022. Language Sample Analysis in Clinical Practice: Speech-Language Pathologists' Barriers, Facilitators, and Needs. *Language, speech, and hearing services in schools*, 53(1):1–16.

William Labov and Joshua Waletzky. 1967. Narrative analysis: Oral versions of personal experience. In: J. Helm, ed., *Essays on the verbal and visual arts*, pages 3–38. University of Washington Press, Seattle and London.

Barbara J. Leadholm and Jon F. Miller. 1992. *Language sample analysis: The Wisconsin guide*. Wisconsin State Department of Public Instruction, Madison.

Nikola Ljubešić and Kaja Dobrovoljc. 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*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Nikola Ljubešić and Davor Lauc. 2021. BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In: *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk: Transcription format and programs (3rd ed.)*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

David Malvern and Brian Richards. 1997. A new measure of lexical diversity. In: A. Ryan and A. Wray, eds., *Evolving models of language*, pages 58–71. Multilingual Matters, Clevedon.

David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical Diversity and Language*

Development. Quantification and Assessment. Palgrave Macmillan, London.

Mercer Mayer (1969). *Frog, where are you?* Dial Press, New York.

Phillip M. McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. PhD thesis, University of Memphis.

Danielle S. McNamara, Arthur C. Graesser, Phillip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, New York.

Jon M. Miller. 1981. *Assessing language production in children: experimental procedures*. University Park Press, Baltimore.

Marilyn A. Nippold, Laura M. Vigeland, Megan W. Frantz-Kaspar, and Jeannene M. Ward-Lonergan. 2017. Language Sampling With Adolescents: Building a Normative Database With Fables. *American Journal of Speech-Language Pathology*, 26(3):908–920.

John M. Norris and Lourdes Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4):555–578.

Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. *Applied Linguistics*, 24(4): 492–518.

Mollee J. Pezold, Caitlin M. Imgrund, and Holly L. Storkel. 2020. Using Computer Programs for Language Sample Analysis. *Language, Speech, and Hearing Services in Schools*, 51(1):103–114.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 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*, pages 101–108, Stroudsburg, PA. Association for Computational Linguistics.

Mabel L. Rice, Filip Smolik, Denise Perpich, Travis Thompson, Nathan Rytting, and Megan Blossom. 2010. Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, 53(2):333–349.

Tanja Samardžić and Nikola Ljubešić. 2021. Data Collection and Representation for Similar Languages, Varieties and Dialects. In: M. Zampieri and P. Nakov, eds., *Similar Languages, Varieties, and Dialects: A Computational Perspective, Studies in Natural Language Processing*, pages 121–137, Cambridge University Press, Cambridge.

Cheryl M. Scott and Sharon L. Stokes. 1995. Measures of syntax in school-age children and adolescents. *Language, Speech, and Hearing Services in Schools*, 26(4):309–319.

Ida J. Stockman. 2010. Listener reliability in assigning utterance boundaries in children's spontaneous speech. *Applied Psycholinguistics*, 31(3):363–395.

Mildred C. Templin. 1957. *Certain language skills in children; their development and interrelationships*. University of Minnesota Press, Minneapolis.

Text Inspector. 2018. *Online lexis analysis tool at textinspector.com*

Martin Tutek, Filip Boltužić, Ivan Smoković, Mario Šaško, Silvije Škudar, Domagoj Pluščec, Marin Kačan, Dunja Vesinger, Mate Mijolović, and Jan Šnajder. 2021. *Podium: a framework-agnostic NLP preprocessing toolkit. GitHub repository*. https://github.com/TakeLab/podium

Guido Van Rossum. 2020. The Python Library Reference, release 3.8.2. Python Software Foundation. https://py.mit.edu/_static/spring21/library.pdf

Marleen F. Westerveld, Gail Gillon, and Jon F. Miller. 2004. Spoken language samples of New Zealand children in conversation and narration. *Advances in Speech Language Pathology*, 6(4):195–208.

Danielle Wetherell, Nicola Botting, and Gina Conti-Ramsden. 2007. Narrative in adolescent specific language impairment (SLI): a comparison with peers across two different narrative genres. *International journal of language & communication disorders*, 42(5):583–605.

Assessing Comparability of Genre Datasets via Cross-Lingual and Cross-Dataset Experiments

Taja Kuzman[†] [*], **Nikola Ljubešić**[†], **Senja Pollak**[†]

[†]Department of Knowledge Technologies, Jožef Stefan Institute
taja.kuzman@ijs.si, nikola.ljubesic@ijs.si, senja.pollak@ijs.si

[*]Jožef Stefan International Postgraduate School

Abstract

This article explores comparability of an English and a Slovene genre-annotated dataset via monolingual and cross-lingual experiments, performed with two Transformer models. In addition, we analyze whether translating the Slovene dataset into English with a machine translation system improves monolingual and cross-lingual performance. Results show that cross-lingual transfer is possible despite the differences between the datasets in terms of genre schemata and corpora construction methods. Furthermore, the XLM-RoBERTa model was shown to provide good results in both settings already when learning on less than 1,000 instances. In contrast, the trilingual CroSloEngual BERT model was revealed to be less suitable for this text classification task. Moreover, the results reveal that although the English dataset is 40 times larger than the Slovene dataset, it provides similar or worse classification results.

1. Introduction

Texts in datasets can be grouped by genres based on their common function, form and the author's purpose (Orlikowski and Yates, 1994). Labeling texts with genres allows for a deeper insight into the composition and quality of a web corpus that was collected with automatic means, more efficient queries in information retrieval tools (Vidulin et al., 2007), as well as improvements of various language technologies tasks, such as part-of-speech tagging (Giesbrecht and Evert, 2009) and machine translation (Van der Wees et al., 2018). That is why automatic genre identification (AGI) has been a subject of numerous studies in the computational linguistics and information retrieval fields (e.g., see Egbert et al. (2015), Sharoff (2018)).

As in other text classification tasks, a large manually annotated dataset is required in AGI in order to train and test a classifier. While there exist some large English genre-annotated datasets, such as the Corpus of Online Registers of English (CORE) (Egbert et al., 2015) with 53,000 texts and the Leeds Web Genre Corpus (Asheghi et al., 2016) with 5,000 texts, for other languages there is either no dataset or mostly a small one, consisting of 1,000 to 2,000 texts, such as genre-annotated corpora for Russian (Sharoff, 2018), Finnish (Laippala et al., 2019), Swedish and French (Repo et al., 2021). This means that for obtaining a large dataset needed for genre identification of other languages, costly and time-consuming annotation campaigns are still needed, leaving most languages under-resourced in regard to the technologies based on the AGI.

However, it might be possible to overcome this obstacle by leveraging the cross-lingual transfer, applying models trained on high-resource languages to the low-resource languages. Recently, Repo et al. (2021) showed that it is possible to achieve good levels of cross-lingual transfer in AGI experiments. They performed experiments in zero-shot cross-lingual automatic genre identification by training multilingual Transformer-based models on the English CORE corpus (Egbert et al., 2015) and testing them on smaller Finnish, Swedish, and French datasets. Rönnqvist et al. (2021) extended this research, training the models on a multilingual dataset, created from the four corpora, which further improved the results.

These promising results stimulated creation of genre-annotated datasets for other languages, and for Slovene, a web genre identification corpus GINCO 1.0 (Kuzman et al., 2021) was created. Its genre schema was based on the CORE schema with the possibility of cross-lingual experiments in mind (see Kuzman et al. (2022)). However, a linguistic analysis of the categories (Biber and Egbert, 2018) and a low inter-annotator agreement, reported by Egbert et al. (2015) and Sharoff (2018), revealed some shortcomings of the CORE schema that could impact the reliability of the dataset. Thus, Kuzman et al. (2022) diverged from the original schema when annotating GINCO, striving towards a more reliably annotated dataset. In addition to this, the CORE and GINCO datasets were created following different corpora collection and annotation approaches (see Section 3.1.). Due to these differences, it remained unclear whether the datasets are comparable enough to allow cross-lingual transfer which would eliminate the need for extensive annotation campaigns of Slovene and other under-resourced languages of interest. This article provides first insight into this, exploring the comparability of the two datasets through cross-dataset and cross-lingual experiments.

2. Goal of the Paper

This paper analyzes comparability of two genre-annotated datasets, the Corpus of Online Registers of English (CORE) (Egbert et al., 2015) and the Slovene Web genre identification corpus GINCO 1.0 (Kuzman et al., 2021). We perform cross-dataset and cross-lingual automatic genre identification experiments to address the main research question (Q1): Is the CORE dataset comparable to the GINCO dataset enough to provide good cross-lingual transfer, as it was achieved by Repo et al. (2021) who

used comparably encoded Finnish, Swedish and French datasets?

To compare the corpora and to analyze their usefulness for monolingual as well as for cross-lingual automatic genre identification, first, labels from both corpora were mapped to a joint schema, the GINCORE schema. Then, multilingual pre-trained Transformer-based models were trained on the English CORE dataset with GINCORE labels (EN-GINCORE), the Slovene GINCO dataset with GINCORE labels (SL-GINCORE) and the SL-GINCORE dataset that was machine translated into English (MT-GINCORE). We conduct 1) monolingual in-dataset AGI experiments, training and testing on the same dataset, 2) cross-lingual and cross-dataset AGI experiments, training on one dataset and testing on the other. The machine-translated dataset is added to the comparison to explore two additional research questions: Q2) In monolingual in-dataset experiments, do multilingual models, which were pre-trained on more English than Slovene data, perform differently on Slovene dataset (SL-GINCORE) than on a Slovene dataset, machine-translated to English (MT-GINCORE)? and Q3) In cross-lingual cross-dataset experiments, does translating the training data (MT-GINCORE) into the language of test data (EN-GINCORE) provide better results than using training and testing data in different languages (SL-GINCORE and EN-GINCORE)?

The experiments were performed with two multilingual Transformer-based pre-trained language models, massively multilingual XLM-RoBERTa model (Conneau et al., 2020), and the trilingual Croatian-Slovene-English CroSloEngual BERT model (Ulčar and Robnik-Šikonja, 2020). This provides an answer to the fourth research question (Q4): Does CroSloEngual BERT, pre-trained on a smaller number of languages, perform better in the cross-lingual AGI experiments than a massively multilingual XLM-RoBERTa model?

3. Data Preparation

3.1. Original Datasets

In this research, three datasets were used: the Corpus of Online Registers of English (CORE) (Egbert et al., 2015), the Slovene Web genre identification corpus GINCO 1.0 (Kuzman et al., 2021) and the GINCO 1.0 corpus, machine translated to English.

The CORE corpus consists of web texts that were extracted from the "General" part of the Corpus of Global Web-based English (GloWbE) (Davies and Fuchs, 2015). The GloWbE corpus was collected via Google searches with high frequency English 3-grams as the queries (Davies and Fuchs, 2015). After obtaining the texts, further cleaning was performed, more specifically, the boilerplate was removed with the Justext tool (Pomikálek, 2011).

The CORE corpus was annotated based on a hierarchical schema which consists of 8 main genre categories, such as *Narrative*, *Opinion*, *Spoken*, and 54 subcategories, e.g., *News Report/Blog*, *Instruction*, *Travel Blog*, *Magazine Article*. The annotation was single-label, i.e., each annotator, recruited through a crowd-sourcing platform, could assign one main category and one subcategory to a text. However, as each text was annotated by four annotators, that means that it can have up to four labels. The corpus that we obtained from the authors and used in this research consists of 48,415 texts, labeled with 8 main categories and 47 subcategories. The corpus was further cleaned by removing duplicated texts and texts with more than one assigned label, resulting in 41,502 texts.

The GINCO corpus (Kuzman et al., 2022) consists of a random sample of web texts from two Slovene web corpora, slWaC 2.0 corpus (Erjavec and Ljubešić, 2014) from 2014 and MaCoCu-sl 1.0 corpus (Bañón et al., 2022) from 2021. Both web corpora were created by crawling the Slovene top-level domain and some generic domains that are interlinked with the national domain. As in GloWbE, the boilerplate was removed with the Justext tool (Pomikálek, 2011). The GINCO corpus consists of two parts, the "suitable" part, annotated with genres, and "not suitable" part, consisting of texts not suitable for genre annotation, such as texts in other languages, machine-translated texts etc. In this research, only the suitable part, consisting of 1,002 texts, was used.

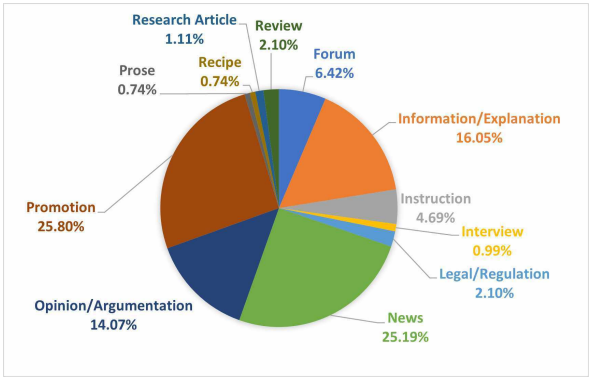
For the annotation, a GINCO schema was used, consisting of 24 labels, e.g., *News/Reporting*, *Opinion/Argumentation*, *Promotion of a Product*. The schema is based on the subcategory level of the CORE schema and on other schemata from previous genre studies. The texts were annotated by two annotators with the background in linguistics. In case of disagreement, final labels were determined at frequent meetings. Multi-label annotation was allowed, i.e., each text could be annotated with up to three classes which were ordered according to their prevalence in the text as a primary, secondary and tertiary label. However, in these experiments, only the primary labels are used. Each paragraph in the texts is accompanied with metadata (attribute `keep`) with information on whether it was manually identified to be a part of the main text and thus useful for the annotation. In this research, paragraphs not deemed to be useful were discarded.

The machine-translated GINCO corpus (MT-GINCO) was created by translating the Slovene GINCO 1.0 to English with the DeepL[1] machine translation system. The system is stated by its developers to be "3x more accurate" than its closest competitors, i.e., Google Translate, Amazon Translate and Microsoft Translator, based on internal blind tests (DeepL, nd). DeepL was confirmed to outperform Google Translate also in an independent study of Yulianto and Supriatnaningsih (2021). The GINCO corpus was translated into British English, as this variety seems to be more frequent than American English in the general part of the GloWbE corpus on which the CORE corpus is based (GloWbe, nd). The prevalence of the British variety in the CORE corpus was also confirmed with a lexicon-based American-British-variety Classifier (Rupnik et al., 2022) which identified 40% of texts to be British, 25% to be American, while the rest contain a mixture of both varieties or no signal for any of them.

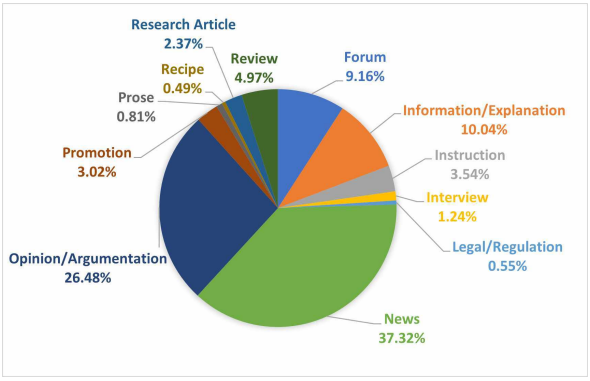
3.2. GINCORE Schema

To be able to perform cross-dataset experiments, the CORE and GINCO schemata were mapped to a joint

[1] `https://www.deepl.com/translator`



(a) SL-GINCORE and MT-GINCORE.



(b) EN-GINCORE.

Figure 1: The differences between the distributions of GINCORE labels in the GINCO corpora MT-GINCORE and SL-GINCORE, and in the EN-GINCORE (CORE corpus).

schema – the GINCORE schema. The schemata were mapped based on descriptions of categories in previous research, in the annotation guidelines for GINCO[2] and the guidelines for CORE, created for the needs of annotation of Finnish, French and Swedish corpora using the CORE schema[3] in further research (Laippala et al., 2019; Laippala et al., 2020). Furthermore, manual inspection of instances from the GINCO and CORE corpora was performed to analyze to which extent the annotations in the corpora match the guidelines. The basis of the GINCORE schema was the GINCO schema as it was shown to provide a more reliable annotation than CORE (see Kuzman et al. (2022)). Moreover, it is easier to map 54 CORE subcategories with a very high granularity to 24 broader GINCO categories than vice versa. The CORE schema consists of broad main categories and more specific subcategories. As the GINCO schema was based on the subcategories of the CORE schema, the subcategories level was used for the mapping from CORE to GINCORE.

Some of genre categories in both schemata are identical and can be directly mapped, namely *Recipe*, *Review*, *Interview* and *Legal/Regulation*. As the GINCO and CORE schemata differ in granularity, broader GINCORE labels were created which efficiently cover categories from both schemata. Some CORE categories were not included in the mapping, because a) these labels revealed to be very infrequent and there is no sufficient information about them available, b) the labels were too broad or problematic for annotators and as a result include instances that are too heterogeneous and cannot be mapped to just one GINCORE label. The resulting GINCORE schema[4] covers 43 CORE subcategories and all 24 GINCO categories by using 20 labels: 15 labels that are present in both corpora, and 5 labels, newly introduced by the GINCO schema and thus present only in the GINCO dataset.

3.3. GINCORE Datasets

For the purpose of performing cross-dataset experiments, only the GINCORE classes that have more than 5 instances in each of the datasets were used, resulting in a smaller set of 12 GINCORE labels: *News*, *Forum*, *Opinion/Argumentation*, *Review*, *Research Article*, *Information/Explanation*, *Promotion*, *Instruction*, *Prose*, *Interview*, *Legal/Regulation*, and *Recipe*. The texts annotated with other GINCORE labels were not included in the experiments. Thus, the final datasets are slightly smaller:

- the English CORE dataset with 12 GINCORE labels, henceforth referred to as the English GINCORE dataset (EN-GINCORE), consists of 33,918 texts;
- the Slovene GINCO dataset with 12 GINCORE labels, henceforth referred to as the Slovene GINCORE dataset (SL-GINCORE), consists of 810 texts;
- the machine-translated English GINCO dataset with 12 GINCORE labels, henceforth referred to as the Machine-Translated GINCORE dataset (MT-GINCORE), consists of 810 texts.

The text instances were not pre-processed, i.e. each instance is a running text as it was extracted from the original web page from which the boilerplate and HTML tags were removed. In GINCO datasets (SL-GINCORE and MT-GINCORE), the texts consist of paragraphs, which is indicated by the <p> tag, while in the CORE dataset (EN-GINCORE), the partitioning into paragraphs is not preserved. In addition to this, the datasets differ significantly in terms of length of the texts. In the CORE dataset, the median length is 649 words, while the minimum and maximum text length is 52 words and 118,278 words respectively. In the GINCO datasets, most texts are significantly shorter, with the median length of 198 words, minimum length of 12 words and maximum length of 4,134 words. As the Transformer models, used in the experiments, can

[2]The guidelines for GINCO are available here: https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/.

[3]The guidelines for the annotation campaigns using the CORE schema are available here: https://turkunlp.org/register-annotation-docs/.

[4]The final table with all the GINCORE mappings is available here: https://tajakuzman.github.io/GINCO-Genre-Annotation-Guidelines/genre_pages/GINCORE_mapping.html.

process maximum instance length of 512 tokens, this means that while the models will in most cases be trained on complete texts from the GINCO datasets, more than half of the texts from the CORE dataset will not be used in their entirety and the models will be trained only on the first part of these instances.

Here, it should be also noted that the CORE dataset and the GINCO datasets are characterized by a different distribution of GINCORE classes. Frequency of some classes, such as *Promotion*, is significantly different, as can be seen in Figure 1.

4. Machine Learning Experiments

4.1. Models

Experiments were performed with the Transformer-based pre-trained language models which were shown to perform well in the automatic genre identification task in a monolingual as well as a cross-lingual setting (Repo et al., 2021). More specifically, two models were used, the base-sized massively multilingual XLM-RoBERTa model (Conneau et al., 2020), and the trilingual Croatian-Slovene-English CroSloEngual BERT model (Ulčar and Robnik-Šikonja, 2020). The XLM-RoBERTa model was chosen because it was revealed to be the best performing model in cross-lingual automatic genre identification based on the CORE dataset (Repo et al., 2021), and to be comparable to the Slovene monolingual model SloBERTa (Ulčar and Robnik-Šikonja, 2021) in experiments, performed on GINCO (Kuzman et al., 2022). The CroSloEngual BERT model was revealed to achieve results comparable to the XLM-RoBERTa model or to even outperform the latter model in common monolingual and cross-lingual NLP tasks (Ulčar et al., 2021). Thus, it was included in these experiments to explore whether it achieves similar results on the AGI task as well.

4.2. Experimental Setup

The datasets were split into 60:20:20 train, dev and test splits, stratified according to the label distribution. The models were trained on the train split, consisting of 20,350 texts in the case of EN-GINCORE, and of 486 texts in the case of SL-GINCORE and MT-GINCORE, and tested on the test split, i.e., 6,784 texts in the case of EN-GINCORE and 162 texts in the case of SL-GINCORE and MT-GINCORE. The dev split, which is of the same size as the test split, was used for testing the hyperparameter optimization. When splitting the datasets, it was assured that the splits of SL-GINCORE and MT-GINCORE contain the same instances, so that they differ only in the language of the content.

The Transformer models are available at the Hugging Face repository and were trained using the Simple Transformers library. To find the optimal number of epochs and the learning rate, the hyperparameter search was performed separately for CroSloEngual BERT and XLM-RoBERTa. The maximum sequence length was set to 512 tokens and other hyperparameters were set to default values. As the EN-GINCORE dataset is more than 40 times larger than the SL-GINCORE and MT-GINCORE datasets, separate hyperparameter searches for each dataset were performed. Optimum learning rate was revealed to be 10^{-5}, while the optimum number of epochs varies based on the training dataset and the model, i.e., the optimum number of epochs when training on the EN-GINCORE with a) XLM-RoBERTa is 9, and b) CroSloEngual BERT is 6; while the optimum number of epochs when training on the SL-GINCORE and MT-GINCORE with a) XLM-RoBERTa is 60, and b) CroSloEngual BERT is 90.

We performed monolingual in-dataset experiments and cross-lingual cross-dataset experiments[5]. The monolingual experiments, described in Section 4.3.1., are in-dataset experiments, which means that the models were trained and tested on splits from the same dataset. In contrast to this, in cross-dataset experiments, presented in Section 4.3.2., the models are trained on one dataset and tested on the other. At the same time, these experiments are cross-lingual, as the original datasets are in different languages.

Three runs of each experiment were performed and average results are reported. The models used in monolingual and cross-lingual setups were evaluated via micro F1 and macro F1 scores to measure the instance-level and the label-level performance.

4.3. Results

4.3.1. Monolingual In-dataset Experiments

First, the datasets are compared via monolingual in-dataset experiments where the models were trained and tested on the splits of the same dataset. In addition to this, a dummy classifier which predicts the majority class was implemented as an illustration of the lower bound. The results, presented in Table 1, show that the mapping of the original labels into a joint schema was successful and that it is possible to achieve good results when learning Transformer models on GINCORE datasets. Transformer models are shown to be very effective at this task, achieving micro and macro F1 scores that are higher than the scores of the dummy model for at least 30 points. XLM-RoBERTa, which was revealed to be the best performing model, achieved relatively high results, with micro and macro F1 scores ranging between 0.72 and 0.84, even when trained on the two smaller datasets, which consist of less than 1,000 instances.

The results show that in a monolingual setting, the massively multilingual XLM-RoBERTa model outperforms the trilingual CroSloEngual BERT model. While Ulčar et al. (2021) showed that the trilingual model is comparable to the XLM-RoBERTa model at NLP tasks which are focused on classification of words or multiword units, such as named-entity recognition and part-of-speech tagging, these results reveal that CroSloEngual BERT is not as suitable as XLM-RoBERTa for automatic genre identification.

Among all monolingual experiments, the best micro and macro F1 results were achieved when the XLM-RoBERTa was trained and tested on the machine-translated MT-GINCO dataset, reaching average micro and macro F1 scores of 0.81 and 0.84 respectively. At the same time, the

[5]The code for data preparation and machine learning experiments is available here: `https://github.com/TajaKuzman/Cross-Lingual-and-Cross-Dataset-Experiments-with-Genre-Datasets`.

Datasets		Majority classifier		XLM-RoBERTa		CroSloEngual BERT	
Trained on	Tested on	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
SL-GINCORE	SL-GINCORE	0.259	0.027	0.782±0.02	0.725±0.01	0.738±0.01	0.599±0.06
MT-GINCORE	MT-GINCORE	0.259	0.027	**0.807**±0.01	**0.841**±0.03	0.714±0.00	0.501±0.05
EN-GINCORE	EN-GINCORE	0.363	0.036	0.768±0.00	0.715±0.00	**0.761**±0.00	**0.706**±0.00
SL-GINCORE	EN-GINCORE	0.029	0.004	**0.639**±0.01	0.539±0.01	0.547±0.02	0.391±0.02
MT-GINCORE	EN-GINCORE	0.029	0.004	0.625±0.01	0.521±0.01	0.585±0.01	0.409±0.01
EN-GINCORE	SL-GINCORE	0.253	0.027	0.603±0.02	0.575±0.03	0.566±0.02	0.510±0.03
EN-GINCORE	MT-GINCORE	0.253	0.027	0.630±0.02	**0.663**±0.03	**0.630**±0.01	**0.543**±0.01

Table 1: Results of monolingual and cross-lingual experiments performed with XLM-RoBERTa and CroSloEngual BERT models, reported via micro and macro F1 scores (averaged over three runs). As a baseline, the scores of a majority classifier are added. The best scores for each of the two Transformer models for each of the two setups (in-dataset experiments and cross-dataset experiments) are shown in bold.

lowest scores, i.e., micro F1 of 0.71 and macro F1 of 0.50, were obtained on the same dataset in combination with the CroSloEngual BERT. Similarly, while XLM-RoBERTa achieved the worst results when trained and tested on the EN-GINCORE, CroSloEngual BERT achieved the best results on this dataset. The difference between the results on the same datasets shows the importance of analyzing the output of multiple models before reaching any conclusion regarding the datasets – if only XLM-RoBERTa would be used, one could assume that the EN-GINCORE dataset is less suitable for automatic genre identification experiments. However, after performing experiments with both models, we can see that no dataset consistently provides the best results.

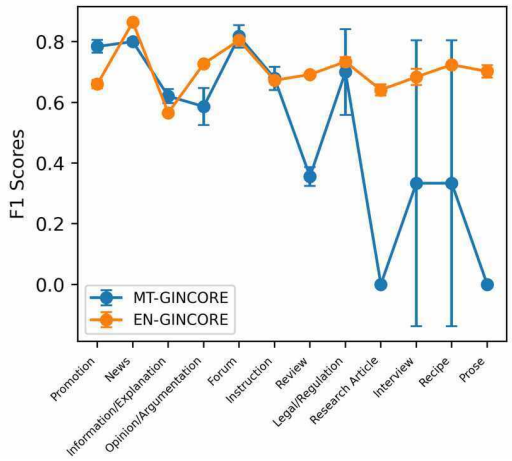


Figure 2: F1 scores per labels (averaged over three runs) in in-dataset experiments with MT-GINCORE and EN-GINCORE, performed with CroSloEngual BERT. Labels are ordered according to their frequency in the smallest of the two datasets, MT-GINCORE.

If we compare experiments, performed with the same model, we can observe that the largest differences between the datasets are in terms of macro F1 scores which are calculated on the level of labels. As shown in Figure 2, the biggest differences between the F1 scores per labels occur in cases of labels that are represented by a very small number of instances in the smaller datasets, SL-GINCORE and MT-GINCORE. Half of the labels, i.e., *Review*, *Legal/Regulation*, *Research Article*, *Interview*, *Recipe* and *Prose*, are represented by solely 4 instances or less in SL-GINCORE and MT-GINCORE test splits. One should be aware that this means that a correct or incorrect prediction of such a small number of instances per labels has a large impact on the macro F1 score. Furthermore, a correct prediction of labels with only one or two instances in the test split might happen due to chance or a similarity of texts in the train and test split. Thus, the F1 scores of these labels are not reliable. As shown in Figure 2, in the three runs, the F1 scores of *Interview* and *Recipe*, which are represented by only 1 instance in the SL-GINCORE and MT-GINCORE test sets, were either 0 or 1, which has a large impact on a macro F1. These results also show how important it is to repeat each experiment multiple times, to ascertain stability and reliability of results.

If we compare the three datasets based on micro F1 scores, there are small differences between them, i.e., a difference of 4 points between the lowest and highest scores when XLM-RoBERTa was used and a difference of 5 points when CroSloEngual BERT was used. Interestingly, although the EN-GINCORE is 40 times larger than the SL-GINCORE and MT-GINCORE, it does not provide higher results than the other two datasets when the XLM-RoBERTa model is used for training. Similar results were revealed in previous work (see Repo et al. (2021)) where they performed monolingual experiments with XLM-RoBERTa on the CORE dataset and three smaller genre-annotated datasets, Finnish FinCORE, French FreCORE and Swedish SweCORE datasets. Although the non-English datasets were annotated with the CORE schema, the annotation procedure and dataset collection methods are more similar to the GINCO approach than CORE. Their experiments showed that the XLM-RoBERTa and other Transformer models perform similarly or better when trained on datasets which consisted of 1,800 to 2,200 instances than when trained on the CORE dataset.

We have two hypotheses why this is the case: 1) It might be that due to high capacities of Transformer models, their performance on this task plateaus already at a few thousand instances and contributing bigger datasets does not significantly improve the results. 2) Or this could indicate that

the CORE dataset is less suitable for AGI machine learning experiments. The reason for that could be that as crowdsourcing was used for the annotation of the dataset, the assigned labels are less reliable and the classes are consequently fuzzier. Poor reliability of the dataset was also confirmed by low inter-annotator agreement. The authors of the dataset reported that there was no agreement between at least three of four annotators on the subcategory of 48.98% of texts (Egbert et al., 2015). When the schema and approach was used by Sharoff (2018) on another corpus, he reported nominal Krippendorff's alpha of 0.53 on the level of subcategories, which is below the acceptable threshold of 0.67, as defined by Krippendorff (2018). In contrast to this, the GINCO dataset was reported to achieve Krippendorff's alpha of 0.71, confirming much higher reliability of annotations.

4.3.2. Cross-lingual Cross-dataset Experiments

To assess comparability of the English CORE dataset and the Slovene GINCO dataset, we performed cross-lingual cross-dataset experiments by training the Transformer models on one dataset and testing them on another. In addition to experimenting with cross-lingual transfer from Slovene to English dataset and vice versa, we also explored whether translating the Slovene dataset into English with a machine translation system improves the results of cross-dataset experiments.

The results, shown in Table 1, reveal that the trilingual CroSloEngual BERT model performs worse than the massively multilingual XLM-RoBERTa model in the cross-lingual experiments with a difference of 12 points between the highest macro F1 scores obtained by the models and a much slighter difference between the highest micro F1 scores (0.009).

In general, results obtained in the cross-lingual experiments are significantly lower than the results from the monolingual experiments. If we compare experiments performed with XLM-RoBERTa, there are differences in 13–18 points in micro F1 and 5–32 points in macro F1 between testing the model on the same dataset as it was trained on (monolingual experiments) and on another dataset (cross-lingual experiments). In case of CroSloEngual BERT, the differences between testing on the same dataset versus testing on the other dataset were in 13–20 points in micro F1 and 9–20 points in macro F1.

Nevertheless, the XLM-RoBERTa scores, which range between 0.6–0.64 and 0.52–0.66 for micro and macro F1 respectively, are a promising indicator that cross-lingual transfer could be possible in this task for Slovene as well. Furthermore, the results are comparable to the results of cross-lingual experiments with the CORE corpora, reported by Repo et al. (2021). When they trained the XLM-RoBERTa model on the CORE corpus and tested it on Finnish, Swedish and French datasets, annotated with the CORE schema, the micro F1 scores ranged from 0.61 to 0.69. Here it needs to be noted that they used a large-sized model which was shown to significantly outperform the base-sized model used by us (Conneau et al., 2020), and that they used 8 labels, while we used 12. Considering this, the results of learning on CORE, mapped to the GINCORE schema, and testing on SL-GINCORE, which reached 0.60 micro F1 with the base-sized XLM-RoBERTa model, are promising, showing that mapping to the GINCORE schema gives comparable results to using the CORE schema.

To obtain a deeper insight into the comparability of the GINCO and CORE corpora, we can compare how the F1 scores per labels change when we test the model on another corpus versus when we test it on the same dataset. Figure 3 shows a comparison between the F1 scores per labels for in-dataset experiments with SL-GINCORE and cross-dataset experiments from SL-GINCORE to EN-GINCORE, performed with XLM-RoBERTa. An analysis of these experiments, performed with CroSloEngual BERT, confirmed that differences between label scores occur when learning with any of the two models, and do not depend on the model. The same differences in label scores were also observed in experiments where MT-GINCORE is used instead of SL-GINCORE, which indicates that the language of the dataset does not seem to have a large impact on the results per labels.

As shown in Figure 3, the F1 scores for *News* and *Opinion/Argumentation* are almost the same in both setups, which shows that in regard to these genres, the datasets are comparable enough for the model to generalize from one dataset to the other. The F1 scores are significantly lower in cross-lingual experiments in case of *Promotion*, *Information/Explanation*, *Forum* and *Instruction*. For the labels that are under-represented in the SL-GINCORE, i.e., labels that are on the right side of *Review* in the Figure, it is not possible to ascertain whether the differences between the scores are an indicator that the datasets are not comparable in regard to these labels or that the differences occurred due to chance.

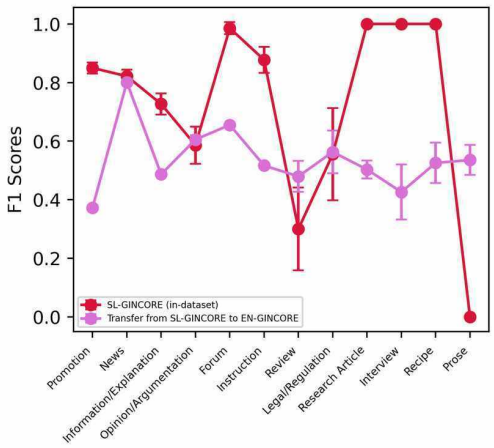


Figure 3: Comparison of average F1 scores per labels between in-dataset experiments and cross-dataset experiments with XLM-RoBERTa. The models were trained on SL-GINCORE, and tested on a) SL-GINCORE (in-dataset experiments) and b) EN-GINCORE (cross-dataset experiments). Labels are ordered according to their frequency in the smallest of the datasets, SL-GINCORE.

As in the in-dataset experiments, experiments with the two Transformer models show that while one dataset combination seems to achieve the best results with one model,

it performs differently with the other model. These results once again show the importance of using multiple models on multiple datasets in the experiments to see whether conclusions obtained from experiments with one model are still supported when using another, yet similar model, and how the performance of the models depends on the datasets. While results in terms of micro F1, achieved with XLM-RoBERTa, point to a conclusion that transfer from SL-GINCORE to EN-GINCORE achieves better results than the other direction, macro F1 scores, achieved with XLM-RoBERTa, and both F1 scores, achieved with CroSloEngualBERT, show transfer direction from English to Slovene to be better. However, although the EN-GINCORE dataset is 40 times larger than SL-GINCORE, the transfer from EN-GINCORE to SL-GINCORE does not achieve significantly higher results than the transfer in the other direction when the Slovene dataset is used.

In addition to this, the results show that machine-translating the dataset into English can in some cases improve the results of cross-lingual experiments. In cases where the model was trained on the GINCO datasets, i.e., SL-GINCORE or MT-GINCORE, and tested on the EN-GINCORE dataset, the setup with the machine-translated text achieved slightly lower results than the setup with the original Slovene dataset, SL-GINCORE, in case of XML-RoBERTa, and slightly better results in case of CroSloEngual BERT. However, when the transfer was applied in the other direction, that is, from EN-GINCORE to SL- or MT-GINCORE, machine translating the test instances from Slovene into English resulted in improvements of macro F1 scores, achieved with XLM-RoBERTa, and both micro and macro F1 scores, obtained with CroSloEngual BERT.

5. Conclusions

Following Repo et al. (2021) who showed that good levels of cross-lingual transfer can be achieved by training Transformer models on a large English genre dataset and applying them to datasets in other languages, the goal of this study was to explore whether it is possible to achieve similar results on the Slovene genre dataset. The results revealed to be promising, as despite using a smaller Transformer model and a different schema with more labels than previous work, the results are rather comparable, showing that the English CORE and Slovene GINCO datasets are comparable enough to allow cross-dataset experiments. The XLM-RoBERTa scores, which range between 0.6–0.64 and 0.52–0.66 in terms of micro and macro F1 scores respectively, are a promising indicator that cross-lingual transfer could be possible in the automatic genre identification task for Slovene as well. Furthermore, high F1 scores achieved with XLM-RoBERTa in monolingual experiments show that automatic genre identification is feasible already with a very small dataset, and that using the GINCORE schema on all datasets gives good results. Moreover, despite the fact that the CORE dataset is 40 times larger than the GINCO dataset, it did not provide consistently significantly better results than the GINCO dataset in either of the setups. We plan to analyze this further by exploring what results can be achieved when smaller portions of CORE are used for training, and by extending the GINCO dataset to analyze whether this further improves the results.

As recently developed trilingual Croatian-Slovene-English CroSloEngual model was shown to be comparable to massively multilingual XLM-RoBERTa model in numerous NLP tasks (see Ulčar et al. (2021)), both models were used in the experiments to analyze their performance in the AGI tasks. The results of both monolingual and cross-lingual experiments showed that despite achieving high results in other common NLP tasks, CroSloEngual BERT seems to be less suitable than XLM-RoBERTa for automatic genre identification.

To improve monolingual and cross-lingual results, we also experimented with translating the Slovene GINCO dataset into English, which is the main language on which the Transformer models were pre-trained. In regard to monolingual experiments, there were no consistent results which would confirm that using an English dataset improves classification. However, when the models were trained on the English EN-GINCORE and tested on MT-GINCORE, i.e., a Slovene dataset, machine-translated into English, this led to improvement of macro F1 scores, achieved with XLM-RoBERTa, and both micro and macro F1 scores for CroSloEngual BERT. This means that machine translating the dataset into the language of another dataset might be beneficial in cross-lingual cross-dataset experiments.

Although monolingual and cross-lingual experiments showed good results also when the models were trained on SL-GINCORE and MT-GINCORE, consisting of less than 1,000 instances, comparisons of F1 scores, reported for each label in different runs and setups, showed that some labels are represented by too few instances to provide reliable results. In the future, we plan to extend the GINCO dataset to assure more reliable results and to further improve the classifiers' performance.

In addition to this, recent work by Rönnqvist et al. (2021) showed that multilingual modeling, where the model was trained on CORE datasets in various languages, resulted in significant gains over cross-lingual modeling, where the model was trained solely on the English CORE dataset. As our research revealed that the CORE and GINCO labels can be successfully mapped to a joint schema, in the future, we plan to extend the experiments to multilingual modeling by training the model on a combination of all CORE datasets and the Slovene GINCO dataset.

Acknowledgments

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains. This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).

6. References

Noushin Rezapour Asheghi, Serge Sharoff, and Katja Markert. 2016. Crowdsourcing for web genre annotation. *Language Resources and Evaluation*, 50(3):603–641.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. Slovene web corpus MaCoCu-sl 1.0. Slovenian language resource repository CLARIN.SI.

Douglas Biber and Jesse Egbert. 2018. *Register variation online*. Cambridge University Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Mark Davies and Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide*, 36(1):1–28.

DeepL. n.d. Why DeepL? `https://www.deepl.com/en/whydeepl`.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Tomaž Erjavec and Nikola Ljubešić. 2014. The slWaC 2.0 corpus of the Slovene web. *T. Erjavec, J. Žganec Gros (ur.) Jezikovne tehnologije: zbornik*, 17:50–55.

Eugenie Giesbrecht and Stefan Evert. 2009. Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus. In: *Proceedings of the fifth Web as Corpus workshop*, pages 27–35.

GloWbe. n.d. Corpus of Global Web-Based English (GloWbE): Texts. `https://www.english-corpora.org/glowbe/`.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Taja Kuzman, Mojca Brglez, Peter Rupnik, and Nikola Ljubešić. 2021. Slovene web genre identification corpus GINCO 1.0. Slovenian language resource repository CLARIN.SI.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. In: *Proceedings of the Language Resources and Evaluation Conference*, pages 1584–1594, Marseille, France. European Language Resources Association.

Veronika Laippala, Roosa Kyllönen, Jesse Egbert, Douglas Biber, and Sampo Pyysalo. 2019. Toward multilingual identification of online registers. In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 292–297.

Veronika Laippala, Samuel Rönnqvist, Saara Hellström, Juhani Luotolahti, Liina Repo, Anna Salmela, Valtteri Skantsi, and Sampo Pyysalo. 2020. From web crawl to clean register-annotated corpora. In: *Proceedings of the 12th Web as Corpus Workshop*, pages 14–22.

Wanda J Orlikowski and JoAnne Yates. 1994. Genre repertoire: The structuring of communicative practices in organizations. *Administrative science quarterly*, pages 541–574.

Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk university Faculty of informatics, Brno, Czech Republic.

Liina Repo, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo, and Veronika Laippala. 2021. Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In: *16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, EACL 2021*, pages 183–191. Association for Computational Linguistics (ACL).

Samuel Rönnqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. 2021. Multilingual and zero-shot is closing in on monolingual web register classification. In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 157–165.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2022. American-British-variety Classifier. `https://github.com/macocu/American-British-variety-classifier`.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

Matej Ulčar and Marko Robnik-Šikonja. 2020. CroSloEngual BERT 1.1. Slovenian language resource repository CLARIN.SI.

Matej Ulčar and Marko Robnik-Šikonja. 2021. Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0. Slovenian language resource repository CLARIN.SI.

Matej Ulčar, Aleš Žagar, Carlos S Armendariz, Andraž Repar, Senja Pollak, Matthew Purver, and Marko Robnik-Šikonja. 2021. Evaluation of contextual embeddings on less-resourced languages. *arXiv:2107.10614*.

Marlies Van der Wees, Arianna Bisazza, and Christof Monz. 2018. Evaluation of machine translation performance across multiple genres and languages. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vedrana Vidulin, Mitja Luštrek, and Matjaž Gams. 2007. Using genres to improve search engines. In: *1st International Workshop: Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pages 45–51.

Ahmad Yulianto and Rina Supriatnaningsih. 2021. Google Translate vs. DeepL: A quantitative evaluation of close-language pair translation (French to English). *AJELP: Asian Journal of English Language and Pedagogy*, 9(2):109–127.

Slovenian Epistemic and Deontic Modals in Socially Unacceptable Discourse Online

Jakob Lenardič,[*] Kristina Pahor de Maiti[†]

[*]Faculty of Arts, University of Ljubljana
jakob.lenardic@ff.uni-lj.si
[†]CY Cergy Paris University
kristina.pahor-de-maiti@u-cergy.fr

Abstract

In this paper, we investigate the use of epistemic and deontic modal expressions in Slovenian Facebook comments. Modals are linguistic expressions that can be strategically used to fulfill the face-saving dimension of communication and to linguistically mask discriminatory discourse. We compile a list of modal expressions that have a tendency towards a single modal reading in order to enable robust corpus searches. Using this set of modals, we first show that deontic, but not epistemic, modals are significantly more frequent in socially unacceptable comments. In the qualitatve part of the paper, we discuss the use of modals expressing deontic and epistemic necessity from the perspective of discourse pragmatics. We explore how the communicative strategy of face-saving interacts with personal and impersonal syntax in the case deontic modals, and how hedging and boosting interacts with irony in the case of epistemic modals.

1. Introduction

Hate speech and other forms of socially unacceptable discourse have a negative effect on society (Delgado, 2019; Gelber and McNamara, 2016). For instance, calls to action targeting specific demographics on social media have been shown to lead to offline consequences such as real-world violence (Siegel, 2020). Linguistically, socially unacceptable attitudes are often disseminated in a dissimulated form, using pragmatic markers which superficially lessen the strength of intolerant claims or violent calls to action; nevertheless, the discursive markers of such dissimulated discourse are still not well known (Lorenzi-Bailly and Guellouz, 2019), especially outside of English social media.

In this paper, we look at the use of Slovenian modal expressions as key pragmatic contributors to the dissimulation of unacceptable discourse on social media. We first look at how the use of epistemic modals, which convey the speaker's truth commitment, and the use of deontic modals, which convey how the world should or must be according to a set of contextually determined circumstances, differ between unacceptable and acceptable discourse in the case of Slovenian Facebook comments obtained from the *FRENK* corpus (Ljubešić et al., 2021).

We then turn to a qualitative analysis of modals conveying logical necessity. We discuss how the meaning of deontic necessity, which corresponds to some kind of obligation that needs to be fulfilled by the agent of the modalised proposition, can have a secondary pragmatic meaning that is akin to face-saving observed with epistemic modals and that arises with syntactically impersonal modals. We then discuss how epistemic modals are used to achieve a face-saving effect, either as hedging or boosting devices or as the intensifiers of irony.

The paper is structured as follows. Section 2. presents the semantic and pragmatic properties of epistemic and deontic modals, while Section 3. presents some of the related corpus-linguistic work on modality in socially unacceptable discourse. Section 4. describes the make-up of the *FRENK* corpus in terms of the subtypes of socially unacceptable discourse and the criteria for the selection of the analysed modals. Section 5. presents the quantitative analysis, wherein epistemic and deontic modals are compared between the acceptable and unacceptable supersets in *FRENK*. Section 6. presents the qualitative analysis, where certain deontic and epistemic necessity modals are discussed in terms of their pragmatic functions. Section 7. concludes the paper.

2. Theoretical background

2.1. The semantics of epistemic and deontic modals

Modal expressions are semantic operators that interpret a prejacent proposition within the irrealis realm of possibility (Kratzer, 2012). There are two key semantic components to modals – one is the modal force, which corresponds to the logical strength of the modal expression and roughly ranges from possibility via likelihood to necessity, and the other is the type of modality,[1] according to which the evaluation of the possibility is tied to the actual world.[2]

There are two main types of modality – epistemic on the one hand and root on the other (Coates, 1983; Kratzer, 2012; von Fintel, 2006). Epistemic modals tie the evaluation of the possibility or necessity to the speaker's knowledge about the actual world. For instance, the possibility adverb *morda* in (1), taken from the *FRENK* corpus, has the reading which says that there is a possibility that the referents of the indefinite subject *nekaj jih* ("some of them")

[1]For formal semanticists viewing modals as quantifiers over possible worlds (von Fintel, 2006; Kratzer, 2012), there are actually three semantic components – *modal force, modal base*, and *the ordering source*; for ease of exposition, we conflate the modal base and ordering source under the simplified *modality type* component of meaning.

[2]The italics in the examples are always our own and used to highlight the modal under scrutiny.

will stay in the country. This possibility reading is epistemic as it conveys that the speaker is not sure whether the possibility of their staying will actually turn out to be the case.

(1) [N]ekaj jih bo *morda*$_{\text{EPISTEMIC}}$ ostalo v naših krajih.
"Some of them will *possibly* stay in our country."

Root modality, on the other hand, is not tied to the speaker's (un)certainty about the truth of the proposition. Rather, it ascribes the possibility to certain, usually unspecified, facts about the actual world. There are several subtypes of root modality, but the one we are interested in this paper is the deontic subtype, in which the evaluation of possibility or necessity is tied to some contextually determined authority, such as a set of rules, the law, or even the speaker (Palmer, 2001, 10). An example of a deontic modal is the verb *dovoliti* in example (2), again taken from *FRENK*. This verb also denotes possibility in terms of modal force, so the deontic possibility reading roughly translates to *they should not be given the possibility* (i.e., be allowed) *to change our culture*.

(2) [S]eveda se jim ne sme *dovoliti*$_{\text{DEONTIC}}$[,] da bi spremenil naso *(sic)* kulturo.
"They should not be *allowed* to change our culture."

Note that a single modal can have different readings in terms of modality type. This is, for instance, the case with the necessity modal *morati*, where the epistemic reading in (3a) conveys that the speaker is certain (i.e., epistemic necessity) that whomever they are referring to is a bonafide Slovenian. By contrast, the deontic reading in (3b) says that what needs to be necessarily done is preparing for the competition. Such readings are disambiguated contextually.

(3) a. Ta *mora*$_{\text{EPISTEMIC}}$ biti pravi Slovenec, ni dvoma.
"He *must* be a bonafide Slovenian, no doubt about it."

b. Pripraviti se bodo *morali*$_{\text{DEONTIC}}$ tudi na konkurenco, ki je zdaj še nimajo.
"They *must* also prepare for the competitors which they do not have."

(Roeder and Hansen, 2006, 163)

2.2. The pragmatics of epistemic and deontic modals

Modality expresses the speaker's subjective attitudes and opinions (Palmer, 2001), which is why the pragmatical aspects of the modalised utterance play an important role in discourse.

Epistemic modals fulfill what Halliday (1970) calls the interpersonal dimension of the utterance. In this sense, epistemic modals show the following three pragmatic uses (Coates, 1987) related to Brown et al. (1987)'s Politeness Theory. First, they are used as part of the negative politeness strategy to save the addressee's negative face, when for instance the speaker tries to facilitate open discussion by not assuming the addressee's stance on the conversational issue in advance. Second, epistemic modals can be used as an *addressee-oriented* positive politeness strategy, which involves the preservation of the positive image of the addressee and prevents them from feeling inferior to the speaker. Finally, they are used as part of a *speaker-oriented* positive politeness strategy, which involves the preservation of the positive image of the speaker by enabling the smooth withdrawal from a statement that can be perceived as a boast, threat, or similar.

Related to such politeness strategies, modals fulfil the conversational role of so-called hedging or boosting devices (Hyland, 2005). Epistemic modals function as hedges when the speaker uses them to reduce their commitment to the truth of the propositional content – i.e., to signal their hesitation or uncertainty in what is being expressed, which is a type of face-saving strategy in and of itself. (Gonzálvez García, 2000; Hyland, 1998). In terms of modal force, it is weak epistemic modals denoting possibility that typically correspond to hedges, though certain necessity modals can also acquire such a function in certain contexts, as we will show in the qualitative analysis.

Strong epistemic modals, which express certainty or high commitment of the speaker to the truth of the utterance, typically function as boosters and are used by the speaker to convince his or her audience, make his or her utterance argumentatively stronger, close the dialogue for further deliberation (Vukovic, 2014), stress the common knowledge and group membership (Hyland, 2005), and so forth. Such boosters can also be used manipulatively to boost a claim that is otherwise controversial or highly particular (Vukovic, 2014).

Deontic modality also fulfils interpersonal roles in communication. Because deontic modals express notions such as obligation and permission, they have to do with negotiating social power between an authority and the discourse participant to whom the permission is granted or obligation imposed upon (Winter and Gärdenfors, 1995). Deontic statements often involve a power imbalance between interlocutors (which is especially evident in case it is not in the interest of the agent to fulfil the obligation), so the use of deontic modals is often paired up with other pragmatic devices denoting politeness or face-saving. Politeness is thus "an overarching pragmalinguistic function that can be overtly or covertly marked in deontic and epistemic modal utterances" (Gonzálvez García, 2000, 127).

3. Related work on modality in hate speech

The linguistic and pragmatic characteristics of modality have not yet been extensively explored in the literature on online socially unacceptable discourse. One exception is the work done by Ayuningtias et al. (2021), who analyses YouTube comments related to the 2019 Christchurch mosque shootings. They find that clauses with deontic modals outnumber those with epistemic modals, and that the main discursive strategy of commenters in socially unacceptable comments is to use deontic modals to incite violent action against members of the New Zealand Muslim community.

Other corpus linguistic studies investigate modal markers from the perspective of stance. Chiluwa (2015), for example, analyses the stance expressed in the Tweets of two radical militant groups, Boko Haram and Al Shabaab.

Among other stance-related elements, she investigates the use of hedges (including weak epistemic modals) and boosters (including strong epistemic modals). The results show that boosters are more frequent than hedges although their overall frequency in the data was low. According to the author, the low frequency of hedges shows that radicalist discourse does not exhibit the tendency to mitigate commitment, which goes hand in hand with the slightly higher presence of boosters that are used as a rhetorical strategy to support (possibly unfounded) statements and to influence, radicalise and win over their readers by projecting assertiveness.

Another study on stance in the context is by Sindoni (2018), who looks at the verbal and multimodal construction of hate speech in British mainstream media. She analyses epistemic modal operators (among other related devices) in order to uncover the writer's stance and attitude towards the content conveyed in the news item. She finds that modality is strategically used to present the author's opinions as facts, while the opinions of others are reported as hypotheses and assumptions.

4. The *FRENK* corpus

4.1. Corpus make-up

Subcorpus	Tokens	
Acceptable	92,922	34%
Offensive	143,948	53%
Inappropriate	1,471	1%
Violent	8,789	3%
Not relevant	24,572	9%
Σ	271,702	100%

Table 1: The make-up of the *FRENK* corpus in terms of socially (un)acceptable discourse.

For this study, we have used *FRENK*, a 270,000-token corpus of Slovenian Facebook comments of mostly socially unacceptable discourse (Ljubešić et al., 2019). The Facebook comments in the *FRENK* corpus concern two major topics – migrants, generally in the context of the 2015 European migrant crisis, and the LGBTQ community, mostly in the context of their civil rights – and are manually annotated for several different kinds of discourse.[3] The annotations distinguish whether the discourse is aimed towards a target's personal background, such as sexual orientation, race, religion, and ethnicity, or their belonging to a particular group, such as political party. They also distinguish the type of the discourse itself, which falls into 4 broad categories, one being acceptable discourse and the others different kinds of socially unacceptable discourse (de Maiti et al., 2019, 38):

- Acceptable discourse
- Socially unacceptable discourse
  - Offensive discourse, which corresponds to abusive, threatening or defamatory speech that is targeted towards someone on the basis of their background or group participation.
  - Violent discourse, which contains threats or calls to physical violence and is often punishable by law (Fišer et al., 2017, 49).
  - Inappropriate speech, which contains offensive language but is not directed at anyone in particular.

[3]The annotations are performed on the comment level while also taking into account the features of the entire discussion thread.

Modal	Syntax	Modality	Force	AF
naj$_{\text{IND}}$	Adverb	Deontic	Likelihood	886
morati	Verb	Deontic	Necessity	489
treba	Adjective	Deontic	Necessity	306
smeti	Verb	Deontic	Possibility	150
verjetno	Adverb	Epistemic	Likelihood	123
mogoče	Adverb	Epistemic	Possibility	92
dovoliti	Verb	Deontic	Possibility	55
morda	Adverb	Epistemic	Possibility	46
najbrž	Adverb	Epistemic	Likelihood	29
ziher	Adverb	Epistemic	Necessity	25
zagotovo	Adverb	Epistemic	Necessity	16
potrebno	Adjective	Deontic	Necessity	4
Σ				2,221

Table 2: The analysed modals; AF stands for absolute frequency.

For our study, we have created two subsets of comments: the *acceptable subset* containing comments tagged as *acceptable*, and the *unacceptable subset* containing comments tagged as *offensive, violent* or *inappropriate*. This decision is based on the frequency distributions as shown in Table 1. We can observe that the FRENK subcorpora are uneven in terms of size, with the violent and inappropriate sets contain significantly fewer comments than the acceptable and offensive sets. Because violent discourse is generally less frequent than offensive discourse in linguistic corpora,[4] it is difficult to annotate automatically (Evkoski et al., 2022), so one of the crucial features of *FRENK* is the fact that the annotations into discourse type were done manually, employing 8 trained annotators per Facebook comment (Ljubešić et al., 2019, 9). Note that about 9% of the Facebook comments are marked as Not relevant, which refers to comments with incorrect topic classification (*ibid., 5*).

The latest, that is, version 1.1, of the *FRENK* corpus, which also includes texts in Croatian and English, is available for download from the CLARIN.SI repository (Ljubešić et al., 2021). However, the online version, which is accessible through CLARIN.SI's noSketch Engine concordancer and which we have used for the purposes of this paper,[5] is not yet available to the public.

4.2. The modals analysed in the study

Table 2 shows that there are 12 modal expressions used in the study. We have selected the modals using the following two criteria.

The first criterion is the modal's tendency towards a single modal reading. As discussed in Section 2.1., modals are in principle ambiguous in terms of their modality type. However, corpus data show that certain modals have an overwhelming preference for a single reading; for instance, while the modal auxiliary *morati* can theoretically have both the epistemic and the deontic interpretations (Roeder and Hansen, 2006, 162–163), as was shown in (3), the epistemic reading (3a) is actually extremely rare in attested usage, and in the case of the *FRENK* corpus completely nonexistent.[6] Similarly, whenever the adverb *naj* is used in the indicative rather than conditional mood (glossed with the subscript IND in Tables 2 and 4), its meaning is always some shade of the deontic reading (command, wish, etc.). Thus, all the modals in Table 2 are either unambiguously deontic or unambiguously epistemic, so they function as a robust set for testing how deontic and epistemic modality manifests itself in different types of discourse without confounding examples with unintended interpretations.

Second, some lexemes known to convey modal interpretations also frequently occur with a superficially similar propositional meaning that, however, is not modal. Such is the adverb *itak*, as in example (4), also taken from *FRENK*.

(4) Krscanstvo pa *itak* izvira iz istih krajev kot islam in juduizem *(sic)*.
"*Of course*, Christianity comes from the same place as Islam and Judaism."

This adverb differs from e.g. the certainty adverb *zagotovo* in that it does not convey the speaker's degree of certainty,[7] but rather simply intensifies whatever he or she knows to be actually the case (the historical-geographic source of Christianity). Because such non-modal readings are usually as frequent as the modal meaning in attested usage, we have omitted them from our study.

Lastly, note that in terms of part of speech, the modals in Table 2 do not constitute a syntactically homogenous set.

[4]This is also a result of the EU Code of conduct and terms of service of social media platforms, according to which content deemed illegal due to its hateful character needs to be taken down.

[5]https://www.clarin.si/noske

[6]The frequency counts were preformed on lemmas, as this is sufficient for distinguishing the part of speech as well; for instance, the lemma *mogoče* corresponds to the adverbial forms, whereas the lemma *mogoč* corresponds to the adjectival ones; however, the adjectival form when used predicatively is consistently ambiguous between the non-epistemic and epistemic interpretations, see Lenardič and Fišer (2021) for discussion and examples.

[7]*Zagotovo* has the synonym *gotovo*; we have excluded it from our overview because it is too frequently used in the non-modal sense, as in (1), which is mostly typical of non-standard Slovenian.

(1) Postrelit in *gotovo*.
"Shoot them all – that's the end of it."

	Acceptable		Unacceptable			
Modal	**AF**	**RF**	**AF**	**RF**	**A/U**	**U/A**
verjetno	52	559.6	66	428.0	1.3	0.8
morda	24	258.3	19	123.2	2.1	0.5
mogoče	29	312.1	55	356.7	0.9	1.1
najbrž	12	129.1	13	84.3	1.5	0.7
zagotovo	3	32.3	13	84.3	0.4	2.6
ziher	8	86.0	15	97.3	0.9	1.1
Σ	128	1,377.4	181	1,173.7	1.2	0.9

Table 3: The distribution of epistemic modals in the *FRENK* corpus; AF stands for absolute frequency and RF for relative frequency, normalised to a million tokens.

While most modals are syntactically adverbs (e.g., *morda, ziher*), some are verbs selecting for finite clausal complements, such as *dovoliti* in (2), verbs selecting for non-finite complements, such as *morati* in (3), and predicative adjectives (of the syntactic frame *It is necessary to*) selecting for non-finite complements, such as *treba* (see the examples in Section 6.1.). However, such syntactic differences have no bearing on the modal interpretation – in all cases, the modals remain sentential operators that take semantic scope over the proposition denoted by the clause.

5. Quantitative Analysis

Tables 3 and 4 show how the Slovenian modals are distributed between the acceptable and unacceptable subsets for the unambiguously epistemic and deontic modals, respectively. The unacceptable subset brings together the three subtypes – offensive, inappropriate, and violent – introduced in Section 4.1.. The acceptable and unacceptable sets contain $92,922$ and $154,208$ tokens, respectively.

In the epistemic set (Table 3), half of the modals – that is, the possibility modal *mogoče* and the necessity modals *ziher* and *zagotovo* – are more frequent in the corpus of unacceptable discourse, while the remaining 3 modals – that is, the possibility modal *morda* and the logically synonymous likelihood modals *najbrž* and *verjetno* – are more frequent in the subset of socially acceptable discourse. Overall, the six epistemic modals are 1.2 times more frequently used in acceptable discourse than they are in unacceptable discourse.

The distribution is reversed in the set of unambiguously deontic modals (Table 4). Here, all modals, save for the possibility verb *smeti* ("to allow"), are more characteristic of unacceptable rather than acceptable discourse, with the deontic necessity adjective *treba* and deontic likelihood adverb naj_{IND} showing the largest preference for the unacceptable set. Overall, the 6 deontic modals are 1.3 times more frequently used in socially unacceptable discourse than they are in acceptable discourse.

Statistically, we have tested the overall differences in frequency between the unacceptable and acceptable sets for both the epistemic (Table 3) and deontic (4) modals using the log-likelihood statistic. This statistic is used to "establish whether the differences [between pairwise frequencies in two corpora with different sizes] are likely to be due to chance or are statistically significant" (Brezina, 2018, 83–84). The formula for calculating the log likelihood statistic

	Acceptable		Unacceptable			
Modal	**AF**	**RF**	**AF**	**RF**	**A/U**	**U/A**
naj$_{\text{IND}}$	227	2,442.9	583	3,780.6	0.6	1.5
morati	151	1,625.0	292	1,893.6	0.9	1.2
treba	87	936.3	197	1,277.5	0.7	1.4
smeti	41	441.2	60	389.1	1.1	0.9
dovoliti	17	183.0	34	220.5	0.8	1.2
potrebno	1	10.8	3	19.5	0.6	1.8
Σ	524	5,639.1	1,169	7,580.7	0.74	1.3

Table 4: The distribution of deontic modals in the *FRENK* corpus.

is given in (5), where the observed values $O_{1,2}$ correspond to the absolute frequencies of a modal in the unacceptable and acceptable sets.

$$(5) \quad 2 \times \left(O_1 \times \ln \frac{O_1}{E_1} + O_2 \times \ln \frac{O_2}{E_2} \right)$$

It turns out that the overall greater occurrence of epistemic modals in the acceptable set (AF $=$ 128 tokens, RF $=$ $1,377.4$ tokens/million) than in the unacceptable set (AF $=$ 181 tokens, RF $=$ $1,173.7$ tokens/million) is statistically insignificant at $p < 0.05$; log likelihood $= 1.902, p = 0.165$. By contrast, the greater occurrence of deontic modals in the unacceptable set (AF $= 1,169$ tokens; RF $= 7,580.7$ tokens/million) than in the acceptable one (AF $=$ 524 tokens; RF $=$ $5,639.1$ tokens/million) is statistically significant at the same cut-off point; log likelihood $= 32.8, p = 9 \times 10^{-9}$.

Using the online tool *Calc* (Cvrček, 2021), we have also calculated the Difference Index (DIN) – an effect-size metric – for the overall difference between the acceptable and unacceptable deontic sets. The DIN value is -14.687, which indicates that the deontic modals' preference for the unacceptable set, although statistically significant, is relatively small (Fidler and Cvrček, 2015, 230). In addition, *Calc* automatically computes the confidence intervals for the relativised frequencies, which is $5,639.1 \pm 471.4$ for the overall acceptable RF and $7,580.7 \pm 426.9$ for the unacceptable RF at the 0.05 significance level. The fact that the intervals do not overlap further confirms that the difference is not accidental.

These findings are related to those in the literature (see Section 3.) as follows. Just like in Ayuningtias et al. (2021)'s work on socially unacceptable discourse in YouTube comments, our deontic modals significantly outnumber epistemic modals in both the acceptable and unacceptable sets (e.g., $1,169$ deontic modals vs. 181 epistemic modals under unacceptable). Second, both modals of epistemic necessity in Table 3 – that is, *zagotovo* and *ziher* ("certainly") – differ from most of the weaker modals, like *morda* ("possibly") and *najbrž* ("likely"), in that they are more frequent in unacceptable discourse; this is similar to the finding by Chiluwa (2015), who shows that strong epistemic modals are more frequent than weak ones in the case of Tweets by radical militant groups. However and in contrast to Chiluwa (2015), our statistically significant finding is not the difference in modal force, but rather the difference in modality type, as discussed above.

6. Qualitative analysis

6.1. Deontic modals in violent discourse

In Section 5., it was shown that deontic modals are more typical of unacceptable rather than acceptable discourse, a finding that was shown to be statistically significant.

To look at the pragmatics of deontic modals and their discursive role in relation to socially unacceptable discourse, let's first recall from Section 4.1. that the socially unacceptable discourse in the *FRENK* corpus is further subdivided into several subtypes. Here we focus on two – offensive discourse on the one hand and violent on the other. It turns out that all of the surveyed deontic modals, with the exception of the auxiliary *morati*, are actually more prominent in violent discourse than in offensive discourse; this is shown in Table 5, where for instance *treba* is almost four times as frequent in the violent-speech subset (RF $= 4437.3$ tokens per million) than it is in the offensive subset (RF $= 1083.7$ tokens per million).

What is interesting is that *treba* and *morati* are synonymous, possibly completely so, in terms of modal logic, as both entail necessities in terms of modal force and in most cases have a deontic reading that has to do with a contextually determined obligation.[8] However, despite the synonymy, *treba* is by far more frequent in violent speech than it is in offensive, while *morati* is the only deontic modal that is more prominent in offensive than in violent speech.

The difference in the distribution of the two synonymous modals can be tied to the fact that they vastly differ in their communicative function, which crucially is observable within the same subset. Put plainly, the chief difference is that *treba* occurs in considerably more hateful statements than *morati*, even though the statements all qualify as violent hate speech rather than offensive speech in that some kind of incitement towards violence is expressed in the modalised statement.

For instance, let's first consider some typical examples with *treba* from the violent subset:

(6) a. To golazen *treba* zaplinit, momentalno!!!!
"These vermin *must* be gassed at once!"

b. Pederčine je *treba* peljat nekam in postrelit.
"Faggots *must* be taken somewhere and shot."

c. Ni *treba* par tisoč Voltov, dovolj je 220, da ga strese in opozori, da bo čez par metrov stražar s puško.
"We don't *need* a couple of thousand Volts; 220 is enough to electrocute them and warn them that, a couple of metres further on, an armed guard is waiting."

[8] Note that in negated sentences with *treba*, negation takes scope over necessity, which means the interpretation is "it is **not** necessary" rather than "it is necessary **not**"; a more principled investigation into how this interaction affects the pragmatics of the modalised propositions is left for future work, though we note that negation in examples such as (6c) behaves in a similar manner to the so-called *metalinguistic negation* (Martins, 2020), as the commenter merely objects to the specific number of Volts, but still condones the violent action i.e. the electrocution of migrants.

Modal	Acceptable	Violent	Offensive
treba	936.3	4,437.4	1,083.7
potrebno	10.8	568.9	243.1
dovoliti	183.0	341.3	213.2
smeti	441.2	682.7	405.7
morati	1,625.0	1,479.1	1,910.4
naj_{IND}	2,442.9	6,371.6	3,647.2
Σ	5,639.2	13,881.0	7,503.3

Table 5: The distribution of deontic modals between the Offensive and Violent subsets of *FRENK*; the frequencies are relative and normalized to a million tokens.

The chief linguistic characteristic of the *treba* examples boils down to lexical choice. The most prominent nominal collocate in the violent subset for the *treba* examples, calculated on the basis of the Mutual Information statistic, is *golazen* "vermin", which can be seen in example (6a), where migrants are referred to as such. According to Assimakopoulos et al. (2017, 41) such metaphoric expressions "are an intrinsic part of the Othering process, and central to identity construction". In the case of animal metaphors such as MIGRANTS ARE VERMIN, migrants are conceptually construed and stereotyped as an invasive out-group that is maximally different from the in-group to which the speaker considers themselves to belong (*ibid.*). The other most prominent nominal collocate is *elektrika* ("electricity"); metaphors containing this lexeme or lexemes related to electricity (volts, to schock, etc.) often have implied reference, where the undergoers of the verbal event, i.e., migrants, are not directly mentioned, as shown in example (6c). Curiously, when the targets of violent speech are not migrants but members of the LGBT community, instead of metaphors like *golazen*, slurs such as *pedri* ("faggots") are used, as in example (6b).

Note that it is not only *treba* which patterns with such charged lexical items; for instance, the adverb *naj*, which denotes the speaker's desire in terms of deontic modality, also frequently occurs with the electricity metaphor, as in (7).

(7) Elektriko v žice spustit. *Naj* kurbe skuri!
"Electrify the fence wires! *May* it burn the whores!"

The examples with *morati*, on the other hand, are significantly less lexically charged, as shown in (8), and the statements framed in a more indirect way.

(8) a. Vse Evropske države bi *morale* bolj grobo udarit po migrantih.
"All European countries should *have to* more strictly strike back against migrants."

b. Kdo nas zaščitil[,] a *moramo* mi tud nabavit pištolo
"Who will protect us? Do we also *have to* buy a gun?"

c. Evropa bi *morala* stopiti skupaj hermeticno zapreti meje.
"Europe should *have to* come together and hermetically close the borders."

Even when the *morati* examples convey that it is necessary that some kind of action be taken against e.g. migrants, as in example (8a), the verbs used are such that they no longer convey explicit violent acts, such as *postreliti* ("to shoot"), *zapliniti* ("to gas"), and *stresti* ("to electrocute") in the *treba* examples (6), but express non-violent acts, as in the case of the verbal phrase *zapreti meje* "close the borders" in (8c). Indeed, the calls to violent action with *morati* are significantly more tentative, as many of the cases of deontic *morati* are embedded under the conditional mood clitic *bi*, which leads to a composite meaning where the deontic necessity is interpreted as a suggestion rather a direct command, as in examples (8a) and (8c), which also is not the case with *treba*.

To sum up the discussion so far, we have observed that while *treba* and *morati* both convey deontic necessity (roughly an obligation that needs to be met), they are paired up with quite substantially different statements in terms of hateful rhetoric in the case of the same type of unacceptable discourse, i.e., violent speech. Further, *morati* is also the only deontic modal which is less typical of violent speech than it is of offensive speech.

We suggest that the difference is tied to the way the pragmatics of deontic modals interact with their core syntactic and semantic properties. As discussed in Section 2.2., pragmatically deontic modals fulfil the interpersonal function in communication. The interpersonal dimension has to do with the fact that the deontic necessity, i.e., obligation, is ascribed by the speaker to whoever corresponds to the agent of the verbal event in the modalised proposition; concretely, in the case of example (8a), the speaker says that it is European countries that have the obligation to strike back against migrants.

The chief difference between the *treba* (6) and the *morati* (8) examples, manifested in the discussed lexical differences, lies in this interpersonal pragmatic dimension, which is crucially influenced by the syntax of the expressions. *Treba* is an impersonal predicative adjective which, in contrast to *morati*, syntactically precludes the use of a nominative grammatical subject that would be interpreted as the agent in the modalised proposition (Rossi and Zinken, 2016). Consequently, all the statements in the *treba* set of examples are such that the agent has an undefined, arbitrary reference – for instance, it is unclear who is expected to "gas the vermin" in example (6a). What happens pragmatically is that the subject-less syntax of the adjective *treba* allows the speaker to sidestep the ascription of obligation to a specific agent, thus largely obviating what is perhaps the core interpersonal aspect of deontic modality. This cannot be really avoided with *morati*, which is a personal verb that obligatorily selects for a grammatical subject in active clauses – in other words, because of its personal syntax, *morati* presents a bigger interpersonal burden on the speaker, as he or she needs to specifically name the person or institution that is required to fulfill the obligation.

Note that, in the violent subset, there is only one example where *morati* is used with the verb *dobiti* ("get"), which induces a passive-like interpretation (9). Here, the grammatical subject headed by *Vsak* ("everyone") is interpreted as the target of the violent action rather that the agent. It is

Modal	Acceptable	Violent	Offensive
morda	258.3	0.0	169.3
mogoče	312.1	113.8	555.8
verjetno	559.6	341.3	451.6
najbrž	129.1	0.0	90.3
ziher	86.0	113.8	97.3
zagotovo	32.3	113.8	83.4
Σ	1,377.4	682.7	1,447.5

Table 6: The distribution of epistemic modals between the Acceptable, Violent, and Offensive subsets of *FRENK*; the frequencies are relative and normalized to a million tokens.

telling that this is also the only example with *morati* which is closer in the use of lexically charged items (i.e., being "shot in the head" rather than "the closing of borders" in the previous examples) to the *treba* examples, as this passive-like construction also precludes the use of an agentive noun phrase (unless it is introduced by the Slovenian equivalent of the *by*-phrase, but there are no such examples in the corpus).

(9) [V]sak, ki se približa našim ženskam in otrokom, *mora* dobiti metek v čelo.
"Everyone who gets close to our women and children *must* be shot in the head."

In short, the interpersonal structure influences the degree of hateful rhetoric, in the sense that speakers are more ready to use degrading metaphors, slurs and violent verbal expressions when they can avoid ascribing the obligation to someone specific. We follow Luukka and Markkanen (1997) by suggesting that impersonality has a similar hedging effect to epistemic modals, in the sense that the unexpressed agent in impersonals introduces a degree of semantic vagueness to the proposition, as does uncertainty brought about by the epistemic reading. Thus, with *treba*, deontic imposition and epistemic face-saving meet in one and the same lexeme.

6.2. Epistemic modals in offensive and acceptable discourse

Epistemic modals are slightly more frequent in acceptable comments, although the difference is not statistically significant, as was shown in Section 5. In order to explore further the possible differences and similarities in the use of epistemic modals between different types of comments, we look at their distribution in three subcorpora, namely in acceptable, offensive and violent comments. The distribution is shown in Table 6. We find that epistemic modals are very infrequent in the violent comments (even unattested for *morda* "possibly" and *najbrž* "likely") in contrast to deontic modals, which are more frequent almost across the board in the violent set (Table 5). On the other hand, the epistemic modals show a similar distribution between acceptable and offensive comments in contrast to violent comments.

We now look at the pragmatics of the epistemic necessity modal *ziher* ("certainly"), as it exhibits the most comparable frequency between the acceptable and offensive subcorpora.

In offensive comments, *ziher* is used either as a booster (10) or a hedge (11), a discursive function which the commenter uses as part of the face-saving strategy. Boosting is shown in example (10).

(10) Begunca? Ekonomske migrante pa picke, ki se ne znajo borit za svoj kos zemlje *ZIHER* ne!!!!!!!
"Accepting a refugee? *CERTAINLY* not accepting economic migrants and cunts who don't know how to fight for their piece of land!!!!!!!"

In this example, the use of the modal conveys the lexical meaning of certainty and thus the full speaker's truth commitment to the propositional content. By being accompanied by excessive exclamatory punctuation, upper case letters and contemptuous argumentation, the modal pragmatically acts as a booster emphasizing the speaker's commitment. The face-saving dimension comes about because the assertiveness conveyed by the modal helps legitimize the speaker as a member of the in-group that is exclusionary of migrant out-group.

(11) [K]r k cerarju nej gredo *zihr* ma veliko stanovanje ... bedaki.
"They better go to the prime minister Cerar, he *surely* has a big flat ... assholes."

Contrary to the previous example, the modal in (11) pragmatically hedges the propositional content by invoking the presumed shared knowledge of the in-group, which concerns the size of the prime minister's home. Here, hedging is related to the fact that the modal activates the face-saving strategy which protects the speaker from the accusation of making an unfounded claim, as the modalised statement, despite entailing certainty, is still weaker than the unmodalised variant which would otherwise report that the speaker holds factual knowledge about the prime minister's apartment.

While the offensive comments predominantly feature *ziher* in such a hedging or boosting role, in the large majority of the acceptable comments, the modal conveys an additional figurative meaning – i.e., that of irony, which we also claim is related to face-saving and contributes an additional persuasive effect in terms of discourse pragmatics (Gibbs and Izett, 2005; Attardo, 2000).

Example (12) conveys a proposition whose ironic meaning is emphasized by the modal *ziher*.

(12) Itak, dejmo vsi lagat, to je *ziher* prav :)
"Of course, let's all lie, that's *certainly* the right thing to do :)"

The ironic reading of this example is suggested by the use of the intensifying adverb *itak* ("of course"), exaggeration by means of the collective reading of the plural pronoun *vsi* ("everyone"), the use of the verb in the first-person *dejmo* ("let's"), and the use of the emoticon. Finally, the face-saving strategy enacted in this example has two dimensions. The first is the protection of the speaker's face since the irony not only enables the speaker to capitalise on the use of a sophisticated rhetorical device, but also to claim group affiliation by clearly stating the values that the

group has in common. The second aspect is the protection of the addressee's face since the irony helps tone down the speaker's criticism – according to Gibbs and Izett (2005), ironic criticism is accepted better or in a friendlier way than direct critiques.

7. Conclusion

This paper has presented a corpus investigation of epistemic and deontic modal expressions in Slovenian Facebook comments in the *FRENK* corpus.

We have first proposed a set of Slovenian modals that show an overwhelming tendency towards a single modal reading. Because of such unambiguity, they constitute a robust set that allows for precise quantitative comparisons between different types of discourse without irrelevant confounding examples and for careful manual analysis of the corpus examples. Quantitatively, we have shown that deontic modals are a prominent feature of unacceptable discourse, and that they are especially prominent in discourse that concerns incitement to violent action, which is legally prosecutable.

In terms of discourse pragmatics, we have first shown that modals which are completely synonymous both in terms of force and modality type can nevertheless profoundly differ in the degree of hateful rhetoric in the same type of socially unacceptable discourse. We have shown that what makes a difference in such examples is the presence of impersonal syntax, which offers speakers the ability to linguistically obviate the ascription of the denoted obligation to a particular agent. We have suggested that this sort of face-saving strategy of ambiguity by way of impersonality correlates with the speaker's tendency to use dehumanising language, such as slurs or degrading metaphors. In the case of epistemic modals, we have shown that acceptable and offensive comments, which are highly similar at their surface linguistic level, differ pragmatically in relation to face-saving; while offensive comments use epistemic modals as simple hedging or boosting devices, acceptable comments use the modals to convey ironic statements in which the irony is emphasised by the modal. We have claimed that the irony also contributes to the face-saving pragmatics.

In future work, we intend to explore how deontic and epistemic modals also differ based on topic (migrants on the one hand and the LGBTQ community on the other). We also want to explore if and how the discourse differs if the unacceptable comments are either directed towards a person's individual background (e.g., race, ethnicity) or group affiliation (e.g., political party).

Acknowledgments

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), the DARIAH-SI research infrastructure, and the national research project *N6-0099: LiLaH: Linguistic Landscape of Hate Speech*.

8. References

Stavros Assimakopoulos, Fabienne H. Baider, and Sharon Millar. 2017. *Online hate speech in the European Union: A discourse-analytic perspective*. Springer Na- ture.

Salvatore Attardo. 2000. Irony markers and functions: Towards a goal-oriented theory of irony and its processing. *Rask*, 12:3–20.

Diah Ikawati Ayuningtias, Oikurema Purwati, and Pratiwi Retnaningdyah. 2021. The lexicogrammar of hate speech. In: *Thirteenth Conference on Applied Linguistics (CONAPLIN 2020)*, pages 114–120. Atlantis Press.

Vaclav Brezina. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge University Press.

Innocent Chiluwa. 2015. Radicalist Discourse: A study of the stances of Nigeria's Boko Haram and Somalia's al Shabaab on Twitter. *Journal of Multicultural Discourses*, 10(2):214–235.

Jennifer Coates. 1983. *The Semantics of the Modal Auxiliaries*. Croom Helm, London and Canberra.

Jennifer Coates. 1987. Epistemic modality and spo- ken discourse. *Transactions of the Philological Society*, 85(1):110–131.

Václav Cvrček. 2021. *Calc 1.03: Corpus Calculator*. Czech National Corpus. `https://www.korpus.cz/calc/`.

Kristina Pahor de Maiti, Darja Fišer, and Nikola Ljubešić. 2019. How haters write: Analysis of nonstandard language in online hate speech. *Social Media Corpora for the Humanities (CMC-Corpora2019)*, page 37.

Richard Delgado. 2019. *Understanding words that wound*. Routledge.

Bojan Evkoski, Andraž Pelicon, Igor Mozetič, Nikola Ljubešić, and Petra Kralj Novak. 2022. Retweet communities reveal the main sources of hate speech. *PloS one*, 17(3):e0265602.

Masako Fidler and Václav Cvrček. 2015. A data-driven analysis of reader viewpoints: Reconstructing the historical reader using keyword analysis. *Journal of Slavic linguistics*, pages 197–239.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 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*, pages 46–51.

Katharine Gelber and Luke McNamara. 2016. Evidencing the harms of hate speech. *Social Identities*, 22(3):324–341.

Raymond W Gibbs and Christin Izett. 2005. Irony as persuasive communication. *Figurative language comprehension: Social and cultural influences*, pages 131–151.

Francisco Gonzálvez García. 2000. Modulating grammar trough modality: A discourse approach. *ELIA, 1, 119-136*.

Michael A.K. Halliday. 1970. Functional diversity in language as seen from a consideration of modality and

mood in english. *Foundations of language*, pages 322–361.

Ken Hyland. 1998. *Hedging in scientific research articles*, volume 54. John Benjamins Publishing Company Amsterdam.

Ken Hyland. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse studies*, 7(2):173–192.

Angelika Kratzer. 2012. *Modals and conditionals: New and revised perspectives*, volume 36. Oxford University Press.

Jakob Lenardič and Darja Fišer. 2021. Hedging modal adverbs in slovenian academic discourse. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 9(1):145–180.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The frenk datasets of socially unacceptable discourse in slovene and english. In: *International conference on text, speech, and dialogue*, pages 103–114. Springer.

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, and Ajda Šulc. 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`.

Nolwenn Lorenzi-Bailly and Mariem Guellouz. 2019. Homophobie et discours de haine dissimulée sur twitter: celui qui voulait une poupée pour noël. *Semen. Revue de sémio-linguistique des textes et discours*, 47.

Minna-Riitta Luukka and Raija Markkanen. 1997. Impersonalization as a form of hedging. *Research in Text Theory*, pages 168–187.

Ana Maria Martins. 2020. Metalinguistic negation. In *The Oxford Handbook of Negation*. Oxford University Press.

Frank Robert Palmer. 2001. *Mood and modality*. Cambridge University Press.

Carolin F. Roeder and Björn Hansen. 2006. Modals in contemporary slovene. *Wiener Slavistisches Jahrbuch*, 52:153–170.

Giovanni Rossi and Jörg Zinken. 2016. Grammar and social agency: The pragmatics of impersonal deontic statements. *Language*, 92(4):e296–e325.

Alexandra A Siegel. 2020. Online hate speech. *Social media and democracy: The state of the field, prospects for reform*, pages 56–88.

Maria Grazia Sindoni. 2018. Direct hate speech vs. indirect fear speech. A multimodal critical discourse analysis of the sun's editorial '1 in 5 brit muslims' sympathy for jihadis". *Lingue e Linguaggi*, 28:267–292.

Kai von Fintel. 2006. Modality and language. In Donald M. Borchert, editor, *Encyclopedia of Philosophy – Second Edition*, pages 20–27. MacMillan Reference USA, Detroit.

Milica Vukovic. 2014. Strong epistemic modality in parliamentary discourse. *Open Linguistics*, 1(1).

Simon Winter and Peter Gärdenfors. 1995. Linguistic modality as expressions of social power. *Nordic Journal of Linguistics*, 18(2):137–165.

The ParlaSpeech-HR benchmark for speaker profiling in Croatian

Nikola Ljubešić,[*†] **Peter Rupnik**[*]

[*]Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ijs.si
peter.rupnik@ijs.si

[†]Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, SI-1000 Ljubljana

Abstract

Recent advances in speech processing have made speech technologies significantly more accessible to the research community. Beyond the most-popular task of automatic speech recognition, classifying speech acts by various criteria has also recently caught interest. In this paper we propose a benchmark constructed from a dataset of speeches given in the Croatian parliament, aimed at predicting the following speaker profile features: speaker identity, gender, age, and power position (whether the speaker is in the ruling coalition or opposition). We evaluate various pre-trained transformer models on our variables of interest, showing that speaker identification and power position prediction seem to rely mostly on language-specific features, while gender and age prediction rely more on generic speech features, available also in models not pre-trained on the target language. We release the benchmark to serve in measuring the strength of upcoming speech models on a lower-resourced language such as Croatian.

1. Introduction

Speech technologies have recently experienced a quantum leap in their development due to the successful application of the self-supervised pre-training of transformer models on speech data (Schneider et al., 2019). Due to this significant simplification of the development of speech technologies, their uptake has increased significantly (Fan et al., 2020; Pepino et al., 2021; Bartelds et al., 2022), which resulted also in the development of the first open dataset for training automatic speech recognition in Croatian (Ljubešić et al., 2022), based on data from the Croatian parliament. The parliamentary data are especially suited for speech experiments, not only because they are in the public domain, but also because they are rich in speaker metadata (Ljubešić et al., 2022).

In this work we are presenting a rather opportunistic benchmark for speaker profiling in Croatian, based on the ParlaSpeech-HR dataset and the available information on the speakers in that dataset. We define four tasks. In the first task, speaker identification, the task is to predict who of the possible 50 speakers is the speaker of a speech act. For the second task, male and female speakers are to be discriminated between. The third task is focused on discriminating between younger and older speakers, 49 years of age being the division point between the two age groups. In the fourth task we aim at discriminating the speech acts depending on whether they were given by MPs from the ruling coalition, or from the opposition.

We compare models pre-trained on the target language (Croatian) and models that were not pre-trained on this language, obtaining insights not only how well transformer models perform on these tasks, but also how much language-dependent these tasks are. While there have been many approaches to speaker profiling developed before the era of transformers, in this work, we limit ourselves on evaluating transformer models only, primarily due to their reported superior performance (Yang et al., 2021).

2. Related work

Various speech benchmarks, including speaker profiling tasks, exist, but mostly for the English language. The SUPERB benchmark (Yang et al., 2021) consists of tasks of speaker identification – identifying the speaker from a closed set of speakers, speaker verification – binary tasks of whether two utterances are spoken by the same speaker, and speaker diarization – predicting who is speaking when for each timestamp, where multiple speakers can also speak simultaneously.

Another recent benchmark, XTREME-S (Conneau et al., 2022), is focused on evaluating universal cross-lingual speech representations in many languages, the tasks based on speech classification covering spoken language identification among 104 languages, and intent classification from the e-banking domain.

A dataset used for speaker identification benchmarking is VoxCeleb (Nagrani et al., 2017), consisting of over 1,000 celebrities voice samples, obtained by applying facial recognition over YouTube videos.

A well known dataset used for benchmarking automatic speech recognition systems, but also used for speaker profiling is the TIMIT dataset (Garofolo et al., 1993), consisting of 630 speakers of 8 dialects of American English. It consists of speaker information on gender, age and height (Kalluri et al., 2020).

In this work we are not trying to build on top of the existing benchmarks due to two reasons. The main reason

is our interest in less-resourced languages, primarily South Slavic languages, for which there is little to no data available. The very recently released ParlaSpeech-HR dataset, on which this benchmark is based on, is the first openly available speech dataset for Croatian (Ljubešić et al., 2022). The second reason is the disruptive effect the speech transformers had on the field, drastically lowering the previous level of error (Yang et al., 2021), with significant improvements expected in the near future as well. This is why we opt for a new, very opportunistic benchmark on speakers from the Croatian parliament. Besides documenting the highly important data selection decisions, we are reporting first results on the current state-of-the-art technology. Given the current high pace of innovation in speech technologies, that is surely not to slow down soon, this benchmark will be highly useful in assessing what new technologies are and will be able to offer to a less-resourced language such as Croatian.

3. Benchmark construction

In this section we present the dataset our benchmark is constructed on, and the data selection protocols given the four variables of interest.

3.1. The dataset

The dataset this benchmark is based on is the ParlaSpeech-HR dataset (Ljubešić et al., 2022), aimed primarily at developing automatic speech recognition systems for Croatian. It consists of 1816 hours of speech obtained from 309 speakers. For each speaker metadata on age, gender, party affiliation, role in the parliament, and power status (opposition vs. coalition) is available. More details on the content and construction procedure of the ParlaSpeech-HR dataset can be found in the description paper (Ljubešić et al., 2022).

3.2. Data selection

For each of the four tasks a separate data selection procedure was set-up, given the limited data available, but also the different nature of the tasks. While most tasks are binary (gender, age, power status), the task of speaker identification is a 50-class task. Furthermore, while for the three binary tasks the training, development and testing subsets have to consist of different speakers, on the speaker identification task, in all three subsets the same speakers have to be present. Finally, in the tasks of age and power status prediction we decided to sample only from male speakers as there are too few female speakers in the dataset for a reasonable sampling that would not include any unwanted bias.

Additionally, in each of the four tasks we only selected instances that were at least 8 seconds in duration. While most of the ParlaSpeech-HR dataset consists of such instances (voice activity detection was set-up in such a fashion), there is a small number of instances, mostly coming from endings of audio files, that are shorter than 8 seconds.

We also discarded speakers producing more than 3,000 instances or less than 200 instances. While the speakers with a small production might complicate the data selection procedure as we want each selected speaker to be equally represented in a sample, the most prolific speakers were left out of the sampling procedures due to their very specific roles in the parliament, which quite likely carries different unwanted biases in their speech production.

In the four following subsections we describe the specific sampling criteria applied for each of our four tasks.

3.2.1. Speaker identification

For the task of speaker identification, 25 speakers per binary gender were sampled. Per speaker, 100 instances were included in the training subset, 10 in the development subset, and 10 in the test subset. Checks were performed to assure that for no speaker instances from the same video appear in more than one subset. With this sampling procedure, each of the three subsets consist of the same 50 speakers, the training subset having 5,000 instances, while the development and testing subsets of 500 instances each.

3.2.2. Gender prediction

For each of the two binary genders, male and female, 25 speakers were selected for the training subset, every speaker being represented with 20 instances. For each of the two genders, 5 speakers (that were not already in the training subset) were taken for the development split, and 5 speakers for the test split. Every speaker in the development and testing subset was represented with 200 instances. With this we assured three subsets of distinct speakers, the training subset consisting of 1,000 instances, and the development and testing subset of 2,000 instances.

3.2.3. Age prediction

Given that there are very few distinct female speakers in the ParlaSpeech-HR dataset, and that controlling for gender while performing any data split is necessary due to the likely strong signal coming from the gender of the speaker as a potential confounder, after some metadata analyses, we decided to setup the age prediction task on male speakers only.

The age distribution of male speakers showed a rather narrow and normal distribution around the median of 49 years of age. The age distribution is far from a uniform and wide distribution, that would allow for a diverse age prediction task, being set-up as a regression task, or a classification task with many categories. This is why we decided to define this as a binary task, predicting whether a speaker is below or above the median age. For the training portion of the task, 60 speakers were selected, with 20 instances per speaker. For the development and test set, 20 speakers were selected for each subset, each speaker being represented by 50 instances. While performing the split, additional checks were put in place to ensure that the age distribution in each of the subsets is as-close-as-possible to the distribution in the full dataset. Additional checkups were also performed to ensure that no speaker leakage existed between the three subsets. With this data selection, the training subset consists of 1,200 instances, while the development and testing subsets consist of 1,000 instances each. Given that the median was chosen as the classification boundary, the final dataset is balanced regarding the two levels of the age variable.

3.2.4. Power status prediction

We decided to wrap up the benchmark with a quite likely less acoustic task, and a more semantic task. Given that we are currently proposing a shared task on predicting whether a transcript of a speech was given by the ruling coalition or opposition, we decided to add that task in this benchmark as well, but performed on speech and not on text transcripts. The ParlaSpeech-HR data come from a single term of the Croatian parliament, which means that the ruling coalition members are mostly from the right political spectrum, while the opposition members are mostly from the left side of the political spectrum. Disentangling the party affiliation or political orientation, and the power status was rather impossible here, which has to be taken into account while analysing the results.

Similar to the task of age prediction, we, again, sampled only among male speakers as the number of female speakers was too low for well-stratified samples. Similar as with age, given the high predictability of gender, we did not want to allow for gender to become a confounder of our primary prediction task, which is power status in this case. We sampled 25 speakers per each power status for train, each speaker being represented by 50 instances. For the development and test sets we selected 9 speakers for each subset, again, representing each speaker with 50 instances. Additional checks were performed that there is no speaker leakage between the three subsets. With this, the size of the training subset is 2,500 instances, while the development and test subsets consist of 900 instances each. For simplicity of evaluation, the division of instances regarding the power status variable is balanced, with 50% instances coming from each side of the political power spectrum.

The benchmark is made available for reproducibility and further benchmarking to the public via the GitHub repo `https://github.com/clarinsi/parlaspeech-hr-benchmark/`.

4. Experimental setup

In this section we give a short description of the setup of the experiments performed on the newly constructed benchmark.

We perform all our experiments with transformer models (Vaswani et al., 2017) that were pre-trained on spoken data. We use the Transformers library (Wolf et al., 2019) and retrieve pre-trained models from the Huggingface model repository.

We use the model pre-trained on Croatian that has proven to perform best on the task of automatic speech recognition (ASR) (Ljubešić et al., 2022), namely the `Slavic` model.[1] We compare the performance of the pre-trained-only model to the model that was additionaly fine-tuned on the ASR task (`Slavic-asr`[2]) to investigate whether fine-tuning the model on the same data, but another task, improves performance.

We compare the performance of the model pre-trained on Croatian to the model that was pre-trained on an unrelated language, in our case English (`English-asr`[3]). We have decided to use the English model fine-tuned for ASR as the non-finetuned model[4] was giving random results after fine-tuning to any of our four tasks. This suspiciously bad result is probably to be followed back to a technical issue in the model, rather than the fact that the model was not fine-tuned on ASR before, as will be seen in the comparison between the performance of the `Slavic` and the `Slavic-asr` model.

The overview of models used in our experiments, together with a short description on the type and amount of data the models were pre-trained and fine-tuned on, is given in Table 1. The non-finetuned Croatian model was pre-trained on around 99 thousand hours of raw recordings of speeches in various Slavic languages that were given in the European parliament. The fine-tuned Croatian model was additionally fine-tuned on the ASR task on around 300 hours of the ParlaSpeech-HR dataset. The English model was pre-trained on 53 thousand hours of raw speech material obtained from audio books and was fine-tuned for the ASR task on 960 hours of similar material.

Regarding hyperparameter optimization, we investigate only the number of epochs required for performance improvements to stall, which is performed by training on the training portion and evaluating on the development portion. For the first two tasks of speaker identification and gender prediction, two epochs were shown to be enough, while for the tasks of age prediction and power status prediction, 15 epochs over the training subset were chosen as optimal.

We evaluate each model on our test subset by reporting both the accuracy and macro F1 metric. Given that all our tasks consist of datasets with a balanced distribution of the response variable, our random baseline lies at 0.5 in case of the binary classification schema, and 0.02 in case of the 50-class speaker identification schema.

For the less challenging tasks of speaker identification and gender prediction, we perform two types of evaluation, on full instances, and on the first 2 seconds of each instance only.

5. Results

5.1. Speaker identification

The results on the speaker identification task are presented in Table 2. The results show for task to be quite easy for the `Slavic` and `Slavic-asr` models applied on full instances. The model fine-tuned on ASR seems to perform slightly better in the full-data scenario, keeping an even score on instances clipped to two seconds, while in that case the non-finetuned model experiences a significant drop of 20 points. This result seems to show how important it is for the model to experience the exact speakers it is supposed to differentiate between, even on another task such as ASR. We do not believe that transfer has occurred between the ASR task and the speaker identification task directly (the model exploiting what people are saying while deciding on the speaker identity) but rather that its parameters

[1] `https://huggingface.co/facebook/wav2vec2-large-slavic-voxpopuli-v2`
[2] `https://huggingface.co/classla/wav2vec2-large-slavic-parlaspeech-hr`
[3] `https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self`
[4] `https://huggingface.co/facebook/wav2vec2-large`

model name	short name	pre-training	ASR fine-tuning
facebook/wav2vec2-large-slavic-voxpopuli-v2	Slavic	Slavic (99k hours)	-
classla/wav2vec2-large-slavic-parlaspeech-hr	Slavic-asr	Slavic (99k hours)	Croatian (300 hours)
facebook/wav2vec2-large-960h-lv60-self	English-asr	English (53k hours)	English (960 hours)

Table 1: List of models used in our experiments, with amount and type of pre-training and fine-tuning data.

model	clipped	accuracy	macro F1
Slavic	no	0.998	0.998
	2 sec	0.806	0.784
Slavic-asr	no	1.000	1.000
	2 sec	1.000	1.000
English-asr	no	0.334	0.275
	2 sec	0.106	0.048

Table 2: Speaker identification results.

were previously adapted to focus better at the peculiarities of the 50 speakers in question.

For the English model, it shows interestingly to perform rather badly, with predictions over the full length of each instance (between 8 and 20 seconds) being correct only in 33% of cases. This is still quite far apart from the random baseline of 2%, but also very far from the stellar performance of the models pre-trained on Croatian. Predicting only on 2 seconds of speech further deteriorates the results to an accuracy of 10%. For the speaker identification task the pre-training language seems to be very important, as the model quite likely models phonetic peculiarities of each speaker, rather than only acoustic features for which any speech transformer should be useful.

To investigate which speakers get confused between by the `Slavic` model, when only two seconds are available for prediction, we present the confusion matrix in Figure1. The matrix shows that speakers of the same gender are being confused between each other, e.g. Arsen Bauk, Davor Bernardić and Božo Petrov being confused for Žarko Katić, or Sunčana Glavak and Ljubica Lukačić being misclassified as Ivana Ninčević-Lesandrić.

5.2. Gender prediction

The results on task of gender prediction are presented in Table 3. On this task all three models, regardless of the language they are pre-trained on, achieve very good performance, the lowest result being accuracy of 98.5%, and the difference in the length of test instances not having a strong impact. Interestingly, the `Slavic-asr` model that performed perfectly on the speaker identification task is the one that performs the worse on the gender prediction task.

Investigating what type of confusion occurs on this task we analyse the output of the `Slavic` model on 2-second instances. We represent the results via a confusion matrix in Figure 2, showing that male instances are sometimes confused for female instances, but not vice versa. Investigating further what speakers are being confused most of the time, it shows that it is a limited number of speakers whose voice has, at least in some occasions, a higher pitch.

The results on gender prediction show that transformer

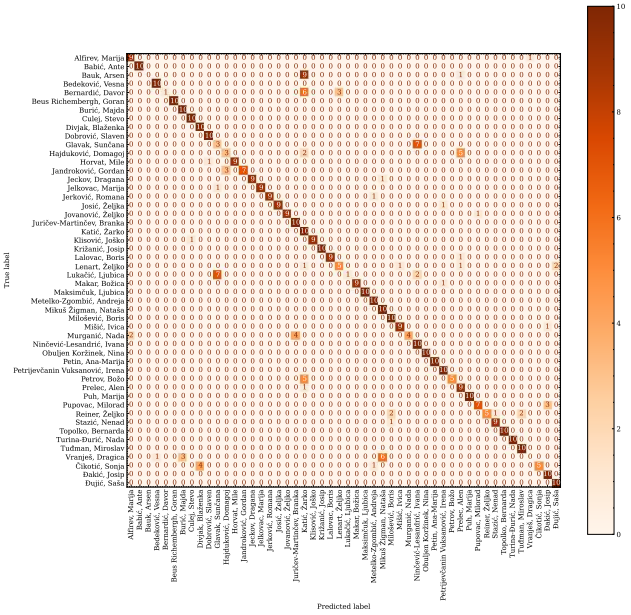


Figure 1: Confusion matrix for speaker identification with the `Slavic` model on instances clipped to two seconds.

eval split	clipped	accuracy	macro F1
Slavic	no	0.997	0.997
	2 sec	0.989	0.989
Slavic-asr	no	0.985	0.985
	2 sec	0.985	0.985
English-asr	no	0.999	0.999
	2 sec	0.994	0.994

Table 3: Gender prediction results.

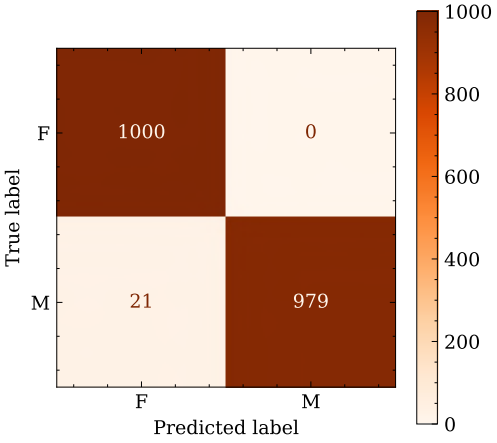


Figure 2: Confusion matrix for speaker gender prediction of the `Slavic` model on 2-second test instances.

model	clipped	accuracy	macro F1
Slavic	no	0.694	0.690
Slavic-asr	no	0.722	0.722
English-asr	no	0.678	0.672

Table 4: Age prediction results.

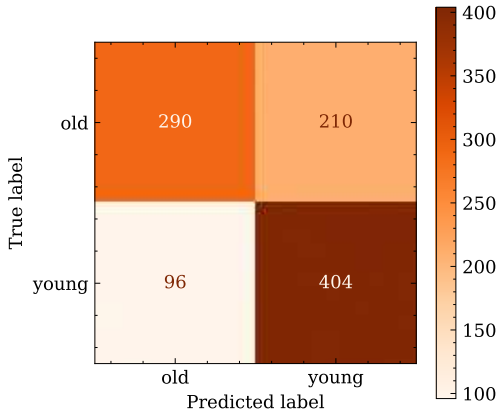


Figure 3: Confusion matrix for speaker age classification by the `Slavic` model.

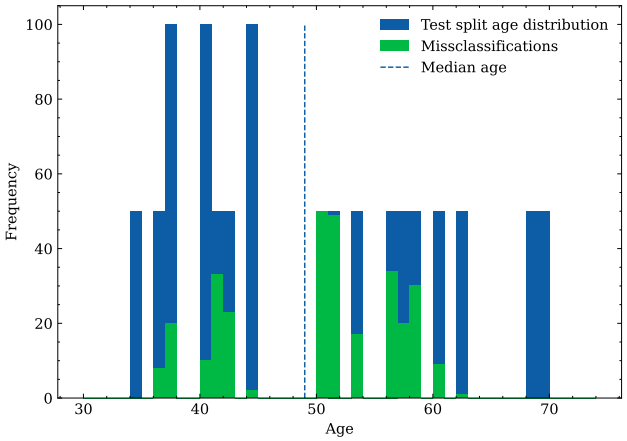


Figure 4: Distribution of age in our test subset, along with misclassifications by the `Slavic` model.

model	clipped	accuracy	macro F1
Slavic	no	0.590	0.587
Slavic-asr	no	0.627	0.626
English-asr	no	0.549	0.531

Table 5: Power status identification.

models do not rely on language-specific features, but quite likely on the pitch of a speaker's voice, with best results being reported by the English model, with almost perfect results even on 2-second test instances.

5.3. Age prediction

The results on age prediction, guessing whether a speaker is younger or older than 49 years, which is the median speaker age in the dataset, are given in Table 4. Here we do not perform experiments on speech samples clipped to two seconds as the task is already demanding enough on full-length instances. The `Slavic-asr` model seems to perform best, with accuracy of 72%, 50% being a random result. The `Slavic` and `English-asr` model seem to be suspiciously close in performance, only with a point and a half difference, which shows that the age prediction task does not rely on language-specific features, but rather general acoustic features.

To investigate the confusion patterns between the two age groups, we plot a confusion matrix of the `Slavic` model in Figure3. The confusion matrix shows clearly that more frequently older speakers tend to be misclassified as younger speakers than vice versa.

Given that we have divided the speakers by age on the median point, and that the speaker age is rather normally distributed, we wanted to additionally check whether most of the prediction errors occur on users who are close to the class boundary. To investigate this, we plot an instance-level age histogram in Figure 4, encoding the correctly and incorrectly classified instances by the `Slavic` model with different colour. The histogram shows that most misclassifications happen, as expected, close to the median class boundary, with almost all instances of speakers of 50 and 51 years of age being misclassified as younger speakers. Classifications on the youngest (35 years) and oldest speakers (68 and 69 years) show to be perfectly performed by the model.

This insight might motivate us to organise the age prediction task in the future as a classification task into three categories, the middle category, around the median age, being considered hard, and discarded in the easier setup of the classification task.

5.4. Power status prediction

The results of our final task, power status prediction, are given in Table 5. The results show to be, as expected, the lowest of all four tasks defined in this benchmark. The `Slavic-asr` model performs best, with the difference to the non-finetuned model being 2.7 accuracy points. The model that was not pre-trained on Croatian achieves a significantly lower result, 5 points lower than any model pre-trained on Croatian, showing that for solving this task mostly language-specific features are used.

Which features exactly are actually used is hard to identify. The only attempt we perform in this direction is a per-speaker analysis of correct and incorrect classifications by the `Slavic-asr` model, which we present in Figure 5. The results show that people in power seem to be easier to identify than those who are in opposition, as the speakers having the lowest percentage of correctly classified instances are mostly from the opposition. The error also seems to be rather speaker-dependent, with eight of the speakers having accuracy above 80%, and the five worst-performing speakers having accuracy below 40%.

Analysing the five worst-performing speakers, a trend can be observed, with the two speakers in power being two of the most fine-mannered speakers, while two out of three speakers from the opposition are rather known for their harsh speech. This analysis has also shown that the signal the classifier has caught on is quite likely based on the polit-

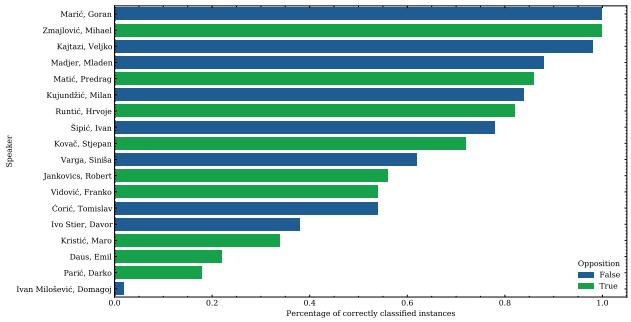


Figure 5: Per-speaker accuracy level with the `Slavic-asr` model on the power status prediction task.

ical orientation rather than power status itself. For performing modelling of the power status in speech, the training and evaluation data should consist of multiple-terms data, with the same political orientations having speeches given while in power and while in opposition.

6. Conclusion

In this paper we have presented a benchmark for speaker profiling in Croatian, based on the recordings of the Croatian parliament. We have carefully selected the speakers and instances to be used in the benchmark, paying special attention to any type of bias or confounders that might be included in the tasks.

We have performed initial experiments with transformer models pre-trained on speech, obtaining interesting insights. The task of speaker identification seems to be rather language-dependent, and can be further improved if the model has seen speakers to be identified before the final fine-tuning process. Gender prediction seems to be the least language specific, obtaining very good results regardless of the model, and quite likely relying simply on the pitch of the speaker. Age prediction, in our case set up as a binary task, with the boundary being the age median, shows to be hard, but very feasible on instances that are further away from the classification boundary. The task shows to use language-specific features to a small amount, but the model that has experienced the same speakers before the final fine-tuning still performing visibly better than the model that has not. Power status prediction is the hardest of all four tasks, and shows to rely on language-specific features, again profiting additionally from experiencing the speakers prior to the final fine-tuning. Analysing the accuracy by speaker shows that the power status model seems to have caught on the political orientation rather than the language of power itself. For working on modelling that phenomenon, a dataset controlling for political orientation should be constructed, which requires a much wider data range than is currently available.

We are releasing the benchmark definitions, to be coupled with the full ParlaSpeech-HR dataset (Ljubešić et al., 2022) in a GitHub repository.[5]

[5] `https://github.com/clarinsi/parlaspeech-hr-benchmark/`

Acknowledgements

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341 (MaCoCu project). This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).

7. References

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. Neural representations for modeling variation in speech. *Journal of Phonetics*, 92:101137.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, et al. 2022. Xtreme-s: Evaluating cross-lingual speech representations. *arXiv preprint arXiv:2203.10752*.

Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. 2020. Exploring wav2vec 2.0 on speaker verification and language identification. *arXiv preprint arXiv:2012.06185*.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, and David S. Pallett. 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403.

Shareef Babu Kalluri, Deepu Vijayasenan, and Sriram Ganapathy. 2020. Automatic speaker profiling from short duration speech data. *Speech Communication*, 121:16–28.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, Ivo-Pavao Jazbec, Vuk Batanović, Lenka Bajčetić, and Bojan Evkoski. 2022. ASR training dataset for Croatian ParlaSpeech-HR v1.0. Slovenian language resource repository CLARIN.SI, `http://hdl.handle.net/11356/1494`.

Nikola Ljubešić, Danijel Korzinek, Peter Rupnik, and Ivo-Pavao Jazbec. 2022. ParlaSpeech-HR – a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. In: *Proceedings of the Third ParlaCLARIN Workshop*, Marseille, France.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng- I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Cross-Level Semantic Similarity in Newswire Texts and Software Code Comments: Insights from Serbian Data in the AVANTES Project

Maja Miličević Petrović,[*] Vuk Batanović,[†] Radoslava Trnavac,[‡] Borko Kovačević[‡]

[*] Department of Interpreting and Translation, University of Bologna
Corso della Repubblica 136, 47121 Forlì
maja.milicevic2@unibo.it
[†] Innovation Center of the School of Electrical Engineering, University of Belgrade
Bulevar kralja Aleksandra 73, 11120 Belgrade
vuk.batanovic@ic.etf.bg.ac.rs
[‡] Faculty of Philology, University of Belgrade
Studentski trg 3, 11000 Belgrade
radoslava.trnavac@fil.bg.ac.rs, borko.kovacevic@fil.bg.ac.rs

Abstract

This paper presents the Serbian datasets developed within the project *Advancing Novel Textual Similarity-based Solutions in Software Development* – AVANTES, intended for the study of Cross-Level Semantic Similarity (CLSS). CLSS measures the level of semantic overlap between texts of different lengths, and it also refers to the problem of establishing such a measure automatically. The problem was first formulated about a decade ago, but research on it has been sparse and limited to English. The AVANTES project aims to change this through the study of CLSS in Serbian, focusing on two different text domains – newswire and software code comments – and on two text length combinations – phrase-sentence and sentence-paragraph. We present and compare two newly created datasets, describing the process of their annotation with fine-grained semantic similarity scores, and outlining a preliminary linguistic analysis. We also give an overview of the ongoing detailed linguistic annotation targeted at detecting the core linguistic indicators of CLSS.

1. Introduction

One of the central meaning-related tasks in Natural Language Processing (NLP) is Semantic Textual Similarity (STS; Agirre et al., 2012). The goal of STS is to establish the extent to which the meanings of two short texts are similar to each other, which is typically encoded as a numerical score on a Likert scale. The similarity scores can subsequently be used in more complex tasks, such as Question Answering (Risch et al., 2021) or Text Summarisation (Mnasri et al., 2017).

In the related task of Cross-Level Semantic Similarity (CLSS) the goal is to contrast texts of non-matching size, such as a phrase and a sentence, or a sentence and a paragraph. CLSS was first formulated as a *SemEval* shared task by Jurgens et al. (2014), who saw it as a generalisation of STS to items of different lengths. Clearly, the length discrepancy brings an additional level of complexity, as longer texts tend to carry a greater amount of salient information than shorter texts, so CLSS can be understood as aiming to measure how well the meaning of the longer text is summarised in the shorter one.

Previous work on CLSS has generally been sparse and, to the best of our knowledge, focused entirely on English. In addition, there is a large discrepancy between the NLP models, which are based on linguistically opaque text properties, and linguistic analyses of semantic similarity. The main aim of this paper is to describe the first non-English annotated CLSS datasets, *CLSS.news.sr* and *CLSS.codecomments.sr*, developed within the project *Advancing Novel Textual Similarity-based Solutions in Software Development* – AVANTES. Both datasets comprise phrase-sentence and sentence-paragraph text pairs in Serbian and both are (being) manually annotated for CLSS. After providing some background, we describe the dataset creation and CLSS annotation, outline a preliminary linguistic analysis, and explain how the linguistic properties identified as relevant for recognising different similarity levels are being annotated further, with a view to improving linguistic descriptions of semantic similarity and testing linguistically informed NLP models.

2. Related work

Previous studies of CLSS are few. The NLP task was introduced by Jurgens et al. (2014, 2016), who provided the first annotated datasets for English, composed of text pairs of different lengths (paragraph to sentence, sentence to phrase, phrase to word, and word to sense), in genres including newswire, travel, scientific, review, and others. The initial datasets were re-used in subsequent work on developing and evaluating CLSS methods at different specific levels (e.g., Rekabsaz et al., 2017 for sentence to paragraph), or regardless of text length (e.g., Pilehvar and Navigli, 2015). Among related tasks, Conforti et al. (2018) dealt with the problem of cross-level stance detection, where the stance target is a sentence, and the text to be evaluated is a long document.

In Serbian, previous work on semantic similarity has been relatively limited. Batanović et al. (2011) and Furlan et al. (2013) introduced *paraphrase.sr*, a corpus of Serbian newswire texts manually annotated with binary similarity judgments; they also used it to train and evaluate several paraphrase identification approaches. Batanović et al. (2018) extended this dataset with fine-grained similarity scores, using the resulting *STS.news.sr* corpus to compare several automatic models. Finally, Batanović (2020) showed that multilingual pre-trained models such as *multilingual BERT* (Devlin et al., 2019) outperform all traditional methods, while Batanović (2021) obtained even better results using BERT's counterpart for Serbian and other closely related languages, *BERTić* (Ljubešić and Lauc, 2021).

In terms of linguistic analysis, semantic similarity is not systematically defined and described, and the contributing phenomena tend to be explored in isolation from each other

(e.g., synonymy in lexical semantics, diathesis alternations in morphosyntax). A somewhat more integrated approach is found with regard to the neighbouring notion of *paraphrase*, intended as a relation of (near-)equivalence of meaning between phrases and/or sentences (Mel'čuk, 2012: 46), i.e. as an instance of high semantic similarity (albeit a non-symmetrical one). According to Milićević (2007), paraphrases can be of different types based on the nature of information that underlies equivalence (linguistic vs. extra-linguistic), the level of linguistic representation involved (morphology, lexicon, semantics, syntax), and the depth of relation. A detailed typology of changes involved in paraphrase has been proposed by Vila Rigat (2013) and Vila et al. (2014) in view of the NLP task of automatic paraphrase detection. This typology combines several criteria and multiple levels of granularity into a taxonomy that will be presented in more detail in Section 4.2, as the basis for our linguistic analysis of CLSS.

3. Datasets and CLSS annotation

The corpora of phrase-sentence and sentence-paragraph text pairs presented in this paper are developed within the AVANTES project. The aim of this project is to support the analysis of correspondences between blocks of source code, written in a programming language, with an analysis of the level of semantic similarity between their respective documentation comments, written in a natural language (English or Serbian), with the goal of detecting code similarity and clones. A CLSS setup is highly appropriate for the textual similarity task due to arbitrary comment length, which can range from single words to phrases, sentences and entire paragraphs. Since the language used in comments is known to diverge from the standard language, for instance in being syntactically incomplete (Zemankova and Eastman, 1980), we add to our study setup CLSS in standard language, choosing newswire texts as its representative.

In the context of the project, comparative analyses are planned both between text domains and between languages. For this reason, it was important to establish a common methodology for the creation and annotation of datasets. Since the only pre-existing CLSS dataset was the *SemEval* one for English, we adopted the approach of Jurgens et al. (2014) as a (partial) model for our work. We retained their five-point similarity scale, with scores ranging from 0 to 4, as well as their definitions for each score: 0 – unrelated, 1 – slightly related, 2 – somewhat related but not similar, 3 – somewhat similar, 4 – very similar. However, we altered the method of text pair construction. Namely, while Jurgens et al. (2014) provided annotators with a longer text and asked them to generate a shorter one with a designated similarity score in mind, we pre-prepared numerous text samples of different lengths (phrases, sentences, and paragraphs), and asked the annotators to combine these texts into phrase-sentence and sentence-paragraph pairs, aiming for a balanced score distribution for the pairs they construct. The main motivation for this choice was that the generation of texts by annotators would have been very difficult to implement in the domain of source code comments, given the highly technical and often project-specific terminology encountered in them. At the same time, our approach prevented a potential paraphrasing bias that the annotators could inadvertently introduce.

3.1. *CLSS.news.sr*

The initial texts for the *CLSS.news.sr* dataset were obtained from the Serbian news aggregator website *naslovi.net*. This website provides a headline and an introductory paragraph for each news report; a subhead is frequently included too. We treated the headlines as source material for phrases, subheads as source material for sentences, and introductory paragraphs as source material for paragraphs for our corpus, exploiting the journalistic convention that the beginning sections in an article commonly provide a summary of its content; our approach was the same one used in the construction of multiple other newswire STS and paraphrasing corpora (Dolan et al., 2004). Since news item are commonly reported differently by different media outlets, cross-linking the texts of different reports allowed for the creation of text pairs with varying degrees of semantic similarity. Close to 18,000 news reports, published between June and August 2021, were scraped using the *scrapy* Python library,[1] to ensure the annotators had a sufficient quantity of raw text available for creating adequate pairs. To ensure comparability with the *SemEval* dataset, our target dataset size was 1,000 phrase-sentence and 1,000 sentence-paragraph pairs.

The construction of the 2,000 text pairs was divided between five annotators, who were either trained linguists or had previous experience with text annotation for the closely related STS task. Even though they received text samples pre-classified based on length, they were instructed to evaluate whether an item in a certain category really was a phrase, a sentence, or a paragraph, and were allowed to change the categorisation. Paragraphs were defined as text containing a minimum of two sentences (where only complete sentences were to be taken into account). A sentence had to contain at least one finite verb form, whereas a phrase was not allowed to contain finite verbs (non-finite forms such as infinitives and participles were allowed, as were deverbal nouns).

The annotators were provided with the similarity score definitions and *SemEval* examples to help them interpret each score. Since these examples proved insufficient to ensure high annotation consistency, the outputs were calibrated by having all annotators create a smaller set of five to six representative pairs for each similarity score and each length pairing. These pairs were reviewed by project researchers and feedback was provided regarding any issues encountered. The following step was the compilation of a detailed set of examples, three per similarity score and length pairing, using the agreed upon representative pairs from all annotators. This set, the score definitions and general instructions became an integral part of the final annotation guidelines for our task, available in the dataset repository in Serbian (original) and English (translation).[2] A subset of examples is shown in Table 1.

The annotators were subsequently asked to construct a total of 200 pairs for each text length combination, trying to include both pairs clearly corresponding to a specific score, and less clear-cut ones. The resulting 2,000 cross-level text pairs were labelled with semantic similarity scores by all five annotators, using the *STSAnno* tool (Batanović et al., 2018). The final score for each pair was calculated by averaging the scores of all individual annotators. Obtaining multiple parallel annotations and

[1] http://scrapy.org/

[2] http://vukbatanovic.github.io/CLSS.news.sr/

averaging them out was chosen instead of relying on an adjudicated double annotation (used for the *SemEval* dataset) in order to minimise individual annotator's biases. In addition, while Jurgens et al. (2014) allowed finer-grained score distinctions using multiples of 0.25, in our setup with five annotators this was not necessary.

Score	Examples
4	Veliki požar na železničkoj stanici u Londonu *A large fire at a London railway station*
	Veliki požar izbio je danas na metro stanici u centralnom delu Londona. *A large fire broke out today at an underground station in central London.*
3	Novi nacionalni praznik: Džuntint *A new national holiday: Juneteenth*
	Američki Kongres usvojio je predlog zakona prema kojem je 19. jun proglašen praznikom u znak sećanja na kraj ropstva i odlazak poslednjih robova 1865. godine u državi Teksas. *The American Congress passed a Draft law declaring 19 June a holiday to commemorate the end of slavery and the liberation of the last slaves in 1865 in the state of Texas.*
2	Veliki problem za Portugal *A major problem for Portugal*
	Loše vesti stižu za Portugal pred start Evropskog prvenstva. *Bad news arrives for Portugal just before the start of the European Championship.*
1	Svađa pred svadbu *A pre-wedding argument*
	Mirko Šijan i Bojana Rodić uskoro očekuju svoje prvo dete, a uveliko se sprema i njihova svadba. *Mirko Šijan and Bojana Rodić are expecting their first child soon, and their wedding is being prepared.*
0	Otvaranje silosa u Zrenjaninu *A silo opening in Zrenjanin*
	Maja Žeželj, voditeljka, ispričala je kako je svojevremeno jedva izvukla živu glavu. *Maja Žeželj, TV presenter, told the story of how some time ago she nearly died.*

Table 1: Guideline examples of phrase-sentence pairs in the newswire dataset for each similarity score.

The final *CLSS.news.sr* dataset comprises 30 thousand tokens in the phrase-sentence subset, and 86 thousand tokens in the sentence-paragraph subset. The average sentence length is ~22 tokens in the sentence-paragraph pairs and ~23 tokens in the phrase-sentence ones. The average phrase length is ~6 tokens, while the average paragraph length is ~64 tokens. The average similarity scores are close to the scale's mean value of 2: 1.91 in the sentence-paragraph subset, and 1.96 in the phrase-sentence subset. The distribution of different scores is fairly uniform, especially for the phrase-sentence pairs; the peaks include a marked one around 0, and a less evident one around 3. The annotation (self-)agreement levels are very high. For the phrase-sentence subset, the average binary agreement between each annotator and the mean of other annotators' scores yields a Krippendorff's alpha coefficient of $\alpha = 0.929$, while the Pearson and the Spearman correlation coefficients are equal, $r = \rho = 0.938$. In the case of sentence-paragraph pairs these values are $\alpha = 0.922$, $r = 0.937$ and $\rho = 0.934$. More details and a comparison with the English *SemEval* dataset are reported in Batanović and Miličević Petrović (2022).

3.2. *CLSS.codecomments.sr*

A particularly innovative part of the work conducted in the AVANTES project is the creation of a corpus of software code comments, to be made publicly available for download and use in testing NLP models once the annotation of semantic similarity is completed. The sources that the code comment dataset was drawn from include public repositories such as GitHub, student projects, coursework and teaching materials from various computing courses at the School of Electrical Engineering of the University of Belgrade and other academic institutions in Serbia, as well as software projects developed at the Computing Center of the School of Electrical Engineering. In order to prevent our work from being focused on the specificities of a single programming language or programming paradigm, we opted to collect comments from eight programming languages: C, C++, C#, Java, JavaScript/TypeScript, MATLAB, Python, and SQL.

We focused on manually pre-selecting only those code comments that describe the functionality of particular sections of code, ranging from individual code lines, to methods and functions, to classes and entire modules. To do so, we relied on a newly designed taxonomy for differentiating between types of code comments (Kostić et al., 2022), which includes the following code comment categories: Code, Functional-Inline, Functional-Method, Functional-Module, General, IDE, Notice and ToDo. The initial data collection and pre-selection were performed by master's degree students at the School of Electrical Engineering of the University of Belgrade, as part of their course project for the Natural Language Processing course. In total, after all duplicate entries were removed, 9,395 code comments belonging to the Functional categories were identified. These include 6,455 Functional-Inline comments, which describe the functionality of individual code lines or code passages, 1,829 Functional-Method comments, which address the functionality of functions and class methods, and 1,111 Functional-Module comments, which are related to the functionality of entire code modules and classes.

In order to construct text pairs, the comments were first roughly divided into candidates for phrases, sentences, and paragraphs on the basis of a set of heuristics. Using whitespace tokenisation, we treated all texts with up to six tokens as candidates for phrases. All texts containing more than six tokens, but limited to a single sentence, were treated as candidates for sentences, while those with more than one sentence were considered paragraph candidates. The number of sentences was determined using a regular expression that treated question marks, exclamation marks, and periods outside of URLs and decimal numbers as sentence boundaries. Using this procedure, the text set was divided into 4,880 phrase candidates, 3,592 sentence candidates, and 923 paragraph candidates.

Due to the high domain specificity of code comments, we entrusted the creation of CLSS pairs to two experienced

programmers. They used the provided candidate texts to form the pairs, but were instructed to carefully evaluate whether each sample truly belonged to its automatically assigned length grouping. Such an evaluation was necessary because complete standard sentences and paragraphs were rarely encountered in the data. Instead, we found that despite having a sentence-like function in the comment, many texts are not true sentences in the linguistic sense – they do not follow any punctuation rules and they lack a predicate, or possess it only implicitly (e.g., *@author Tim 2* or *Naziv komponente* 'Component name' within a paragraph item). Similarly, paragraphs in the code comment domain are often separated into units not via standard punctuation, but rather by using visual boundaries, such as moving to a new line in the source file, or (repeatedly) using special characters (e.g., * or ###). Limiting our text selection to a rigid definition of sentences and paragraphs would thus not only have reduced the size of the dataset, but it would also have led to the exclusion of numerous domain-specific phenomena, significantly impacting our linguistic analyses of code comments. We therefore decided to count as paragraphs texts consisting of at least two clearly identifiable units, even if those units were not true sentences. Similarly, we expanded the sentence set with texts containing an implicit predicate, as well as with those containing subordinate clauses without a main clause (e.g., relative clauses such as: *Metode koje se odnose na simulaciju procesa* 'Methods that refer to process simulation').

Score	Examples
4	Računanje površine pravougaonika *Calculating the area of a rectangle*
	Površina pravougaonika po formuli je a * b *The area of a rectangle according to the formula is a * b*
3	POMOCNA FUNKCIJA *AUXILIARY FUNCTION*
	Fajl koji pruza pomocne funkcije *A file that provides auxiliary functions*
2	ubrzano kretanje *accelerated movement*
	Zelimo da se ogranicimo od mogucnosti da se ubrzano krece. *We want to limit the possibility of accelerated movement.*
1	Update dokumenta *Document update*
	Ovaj program formira html dokument *This program forms an html document*
0	izracunavanje faktorijela *calculating the factorial*
	Azurira rotaciju kamere preko pomeraja misa *Updates the camera rotation via mouse movement*

Table 2: Guideline examples of phrase-sentence pairs in the code comment dataset for each similarity score.

This allowed us to construct a code comment dataset of the same size as *CLSS.news.sr*. The *CLSS.codecomments.sr* dataset therefore includes 1,000 phrase-sentence pairs, comprising 14 thousand tokens, and 1,000 sentence-paragraph pairs, comprising 39 thousand tokens. The average sentence length is ~10 tokens in both the sentence-paragraph and the phrase-sentence pairs. The average phrase length is ~3 tokens, while the average paragraph length is ~29 tokens. Overall, the code comments are approximately half the length of the newswire text items.

Although our initial aim was again to construct a dataset balanced across the range of similarity scores, this proved to be impossible with our selection of source texts, since they pertained to a wide range of programming projects with different purposes and implemented using diverse programming paradigms and languages. This made the construction of pairs with high similarity scores very problematic. We therefore abandoned the goal of obtaining a balanced score distribution, but still instructed the programmers to compile as many highly similar pairs as possible with the given source content. Each programmer was tasked with the construction and scoring of 500 pairs of each length.

The similarity scoring of the text pairs was performed on the basis of guidelines similar to the ones used in the newswire domain, but with a new set of three examples per score and length pairing, drawn from the code comment domain; a subset of phrase-sentence pair examples is shown in Table 2. After the code comment text pairs were constructed, they were forwarded to the same annotators who worked on the *CLSS.news.sr* dataset, in order to obtain multiple parallel annotations. Since this work is still in progress, our linguistic analyses of *CLSS.codecomments.sr* in this paper will be based on the individual similarity scores assigned by the two programmers who constructed the text pairs.

4. Linguistic analysis

The NLP algorithms used in automatic treatment of semantic similarity rely on different types of information, including linguistic features. While state-of-the-art models such as *multilingual BERT* and *BERTić* reach performances that correlate highly with human scores, with coefficients $r,\rho > 0.9$ for CLSS on Serbian newswire texts (Batanović and Miličević Petrović, 2022), they lack linguistic transparency and are of limited help in understanding the relative contributions of different levels of language structure and different specific features. Since one of the aims of the AVANTES project is to combine NLP with linguistic knowledge, we conduct two types of linguistic analyses on the datasets. A preliminary qualitative analysis is performed to gain initial insight into the data and help decide on the specifics of detailed annotation of semantic similarity indicators (to be followed by a quantitative analysis of the annotated datasets).

4.1. A qualitative overview

A qualitative linguistic analysis was performed on a random sample of ten text pairs per score, for both *CLSS.news.sr* and *CLSS.codecomments.sr*, and for both phrase-sentence and sentence-paragraph pairs. In the case of newswire texts, items that received the same score by all annotators were selected; an approach focused on clear-cut cases was deemed useful as a first step in the analysis given its goals of verifying both the linguistic relevance of the similarity scores and the taxonomy for more detailed linguistic annotation. For comments, the initial scores

assigned by programmers were used for selection. The analysis consisted in a comparison of information content between the pairs' components, as well as a study of vocabulary overlaps (or lack thereof). Its goal was to get an initial grasp of the data and help define a taxonomy to base a more elaborate analysis on.

For both corpora and both types of comparisons, the pairs marked 4 are characterised by the occurrence of the same distinctive vocabulary items: personal names and/or numbers (newswire), or specialised terms (comments). The form is often not identical, but the items involved are clearly relatable on morphological grounds (e.g., they are inflectional forms of the same noun, as in *Kragujevcu*.LOC – *Kragujevca*.GEN 'Kragujevac', *parametre*.ACC – *parametrima*.INS 'parameters', or a noun and a denominal adjective, as in *Vlasotincu*.N – *vlasotinačkom*.ADJ '(of) Vlasotince')[3]. The shared numbers are mostly large and either quite specific or used in a collocation (e.g., *100.620*, or *3.000 dinara* '3000 dinars'). Overlaps in common lexical words are also frequently based on morphologically related rather than identical forms (e.g., *stiglo*.PAST.PART – *stići*.INF 'arrive', *novozaraženih* 'newly infected' – *novih slučajeva zaraze* 'new cases of infection', *filtriranje* 'filtering' – *filtar* 'filter'). A number of synonyms are found (*potvrda – sertifikat* 'certificate', *promenljiva – varijabla* 'variable'), sometimes involving a Serbian and an English word (*mreža – grid* 'grid'), and sometimes within different collocations based on the same term (e.g., *toplotni talas – talas vrućina* 'heat wave', *zoom levela – stepena zoom-a* 'zoom level'). Overall, most lexical words from the smaller unit are present in the larger one, which also contains other elements that describe the situation in more detail, but without adding entirely new topics (*u Londonu* 'in London' – *u centralnom delu Londona* 'in central London'; *funkcija sa parametrima* 'a function with parameters' – *funkcija koja nije f(void), vec prima parametre* 'a function that is not f(void), but accepts parameters').

Score 3 items are distinguished by similar properties in terms of shared lexis and especially personal names and specialised terms, but with entirely new information in the longer item, and/or partly different information in the components of the pair, leading to a less marked overall vocabulary overlap (e.g., *Neuralna mreza* 'neural network' – *vanila neuralna mreza koja se obucava pomocu genetskog algoritma* 'vanilla neural network which is trained via a genetic algorithm'). Near-synonyms appear to be more common in score 3 pairs (*reč* 'word' – *termin* 'term', *nov ugovor* 'new contract' – *produžetak saradnje* 'extension of collaboration'). In both score 4 and score 3 items, the head noun of the phrase tends to appear as the subject or the object of the sentence predicate, or it is a deverbal noun that corresponds to the predicate (*unos*.N – *unosi*.V 'input'). The predicate is typically the same in sentence-paragraph pairs, with additional predicates in the paragraph item.

Among less similar pairs, those marked 2 are somewhat mixed, as they either contain different personal names/ specialised terms and similar common vocabulary, or vice versa (*Tropski pakao u Beogradu* 'tropical hell in Belgrade' – *I sutra će u Novom Sadu biti veoma toplo* 'It will again be very warm in Novi Sad tomorrow'; *prekid rekurzije* 'interruption of recursion' – *ako ima decu onda idemo rekurzivni poziv* 'if it has children then we do a recursive call'). The predicate of the sentence item is typically not related to the head noun of the phrase item. The pairs marked 1 and 0 contain barely any overlapping personal names or specialised terms. Score 1 items do share some common lexical words, but synonyms, near-synonyms, and terms from the same wider semantic field are more present than words that are identical or morphologically closely related (e.g., *tragedija* 'tragedy' – *nesreća* 'accident', *pljuskovi* 'showers' – *kiša* 'rain'). Items marked 0 typically do not share any lexical words.

When it comes to differences between the two corpora, in *CLSS.news.sr* it is often the case that the relatedness of lexical items in the pair is based on real world knowledge (largely about something happening at the time of writing) rather than on linguistic information (e.g., *vakcinacija* 'vaccination' – *virus korona* 'corona virus', *Tokio* 'Tokio' – *Olimpijske igre* 'Olympic games'), especially in items assigned a score below 3. *CLSS.codecomments.sr*, on the other hand, is characterised by various non-standard features, such as inconsistent spelling (*popup* vs. *pop-up*), missing diacritics (*cita* for *čita* 'reads'), inflectional endings on English words inconsistently spelt with/without a dash (*zoom-a*, *workspace-u* vs. *levela*), non-standard abbreviations (*f-ja* for *funkcija* 'function)', or phonetic transcription of English terms (*eksepšn* 'exception').[4]

4.2. Linguistic annotation

Using the preliminary analysis outlined above and the existing paraphrase typologies (primarily Vila Rigat, 2013; Vila et al., 2014; also Milićević, 2007; Mel'čuk, 2012), we propose a taxonomy of semantic similarity types and indicators, shown and illustrated in Table 3; most examples are taken directly or adapted from our corpora (examples for two clear indicators are omitted to save space). The initial focus is on the nature of information that similarity is based on, and a core distinction is made between linguistic, quasi-linguistic and extralinguistic similarity types. This is at the same time one of the main points of divergence between our approach and the one by Vila Rigat (2013) and Vila et al. (2014), who acknowledge the existence of non-linguistic paraphrase, but do not include it in their core typology; we rely on Milićević (2007) and Mel'čuk (2012) for these types. Another difference with respect to previous work is that our taxonomy makes reference to similarity *indicators*, while *changes* are invoked in previous work, due to paraphrase being perceived as involving a source and a target item.

Linguistic similarity is based on language-internal information at the word/lexical unit level (i.e., the morpho-lexicon), the level of structural organisation, and the level of meaning (i.e., semantics). The first two types have two subtypes each: morphology- and lexicon-based and syntax- and discourse-based indicators respectively; the indicator types and subtypes thus follow the classical organisation in formal levels of linguistic analysis. Finally, the indicator names in the last column of Table 3 denote specific mechanisms through which semantic similarity is established. Following Vila et al. (2014), our assumption is that the indicators reveal what triggers semantic similarity at the micro level. In other words, unlike the similarity

[3] Abbreviations used: LOC – locative; GEN – genitive; ACC – accusative; INS – instrumental; ADJ – adjective; N – noun, PAST.PART – past participle; INF – infinitive; V – verb.

[4] Many of the features found in code comments are shared with computer-mediated communication in Serbian (see Miličević Petrović et al., 2017).

scores assigned to pairs of items as wholes (i.e., to entire phrases, sentences, or paragraphs), the linguistic taxonomy targets individual phenomena that cumulatively contribute to the overall score, where such individual elements are not mutually exclusive and several can be co-present.

Looking more closely at the indicator subtypes, morphology-based indicators concern the morphological form of words, capturing complete equivalence, as well as inflectional and derivational relations, i.e. different forms of the same word or changes of category via derivational

Similarity type	**Indicator type**	**Indicator subtype**	**Indicator (*example*)**
Linguistic	Morpholexicon-based	Morphology-based	- Identical (*požar – požar* 'fire') - Inflectional (*parametre*.ACC – *parametrima*.INS 'parameters') - Derivational (*Vlasotincu*.N – *vlasotinačkom*.ADJ '(of) Vlasotince')
		Lexicon-based	- Spelling and format (*pop-up – popup*) - Synthetic/analytic (*novozaraženih* 'newly infected' – *novih slučajeva zaraze* 'new cases of infection') - Same polarity -- Synonymy (*potvrda – sertifikat* 'certificate) -- Near-synonymy (*reč* 'word' – *termin* 'term') -- Hyponymy (*škoda* 'Škoda' – *automobil* 'car') -- Meronymy (*Vašington* 'Washington' – *SAD* 'USA') - Opposite polarity (*izgubio* 'lost' – *nije uspeo da pobedi* 'failed to win') - Converse (*pogibija dva pešaka* 'death of two pedestrians' – *usmrtio pešake* 'killed the pedestrians')
	Structure-based	Syntax-based	- Diathesis alternations (*opljačkali su stan* 'robbed the flat' – *stan je opljačkan* 'the flat was robbed') - Coordination changes - Subordination and nesting changes
		Discourse-based	- Punctuation (*Potpis dana - Aleksandar Kolarov!* 'Signature of the day - Aleksandar Kolarov!' – *Aleksandar Kolarov potpisao novi ugovor* 'Aleksandar Kolarov signed a new contract') - Direct/indirect style (*Bilčik ocenjuje da vežbe ne pomažu* 'Bilčík states that the military exercises do not help' – *Bilčik ukazuje da vesti o vežbi "nisu od pomoći"* 'Bilčík points out that the news of a military exercise "is not helpful"') - Sentence modality (*maske više nisu obavezne?* 'masks no longer compulsory?' – *neće biti obavezne zaštitne maske* 'protective masks will not be compulsory')
	Semantics-based		(*Tropski pakao* 'tropical hell' – *biti veoma toplo* 'be very warm')
	Miscellaneous		- Change of order (*klasa singleton – Singleton patern* 'singleton class/pattern') - Addition/deletion (*funkcija za sortiranje* 'sorting function' – *metoda koja sortira uzetu matricu* 'the method that sorts the given matrix')
Quasi-linguistic	Pragmatic		(*Scattered showers are very likely – Bring your umbrella*; Mel'čuk, 2012: 60)
Extralinguistic	Situational		(*Besplatno kroz Severnu Makedoniju od danas* 'Free travel through North Macedonia from today' – *Novina od 15. juna* 'New rules from 15 June')
	Encyclopaedic		(*Italija* 'Italy (the team)' – *ekipa sa Apenina* 'the team from the Apennine Mountains')
	Logical		(*Još pola dinara za veknu hleba* 'Half a dinar more for a loaf of bread' – *Cena hleba visa za 20%* 'The price of bread higher by 20%'; Milićević, 2007: 145)

Table 3: Overview of the taxonomy of semantic similarity (the examples are drawn from *CLSS.news.sr/CLSS.codecomments.sr*, or from the literature).

affixes. The *identical* indicator is not present under the morphology heading in Vila Rigat (2013) and Vila et al. (2014), who categorise it as a "paraphrase extreme", which is a special type in their taxonomy, capturing longer chunks of text; we add it based on the preliminary analysis presented in Section 4.1, which revealed that identical individual words are common in highly similar items in CLSS. Additional information that could prove useful concerns parts of speech, the distinction between personal and common nouns, as well as information on general vs. specialised vocabulary. Given that the identification of specialised terminology would require work that goes beyond the scope of the current project, we are still evaluating the possibility of including it in the analysis.

Lexicon-based indicators are somewhat more varied, ranging from different spellings of the same words, to syntactic and analytic expressions of the same meaning, and to lexical semantic relations in the narrow sense. Same polarity items constitute the most complex group of lexical relations, comprising synonymy as a similarity relation *par excellence*, near-synonymy, hyponymy (the relationship between superordinate/more general and subordinate/more specific lexical items), and meronymy (a part-whole relation). Opposite polarity relations are based on antonym pairs with opposite comparative words, or with one of the components negated. Finally, a converse relation captures complementary actions whose arguments are inversed.

Syntax-based indicators capture those relations that imply a syntactic reorganisation in the sentence; they can be found within single sentences, or in the way multiple sentences are connected. Specific cases include instances of diathesis alternations (such as the active/passive alternation), coordination (where coordinated units are present in one member of the pair, but not in the other), and subordination or nesting (where subordinate/nested elements are present in only one item). The second subtype of structural changes, discourse-based indicators, do not affect the sentential arguments, but are instead related to elements such as punctuation and formatting (beyond single lexical units), affirmative vs. interrogative sentence modality, and direct vs. indirect speech.

The semantics-based subtype is also distinguished by going beyond the level of individual lexical items, as it concerns phrase/sentence-level meaning. No subtypes of specific indicators are singled out, as this level of analysis refers generally to the distribution of semantic content across lexical units, and it can involve multiple and varied formal changes that lead to different lexicalisations of the same meaning units. The boundaries between semantics-based similarity and lexicon-based similarity indicators are not always clear-cut, but it is generally the case that lexicon-based indicators concern individual words or multiword units, while semantics-based similarity relies on multiple lexical items.

The last type of linguistic indicators is classified as miscellaneous, given that it captures phenomena that do concern the linguistic structure of items, but do not clearly belong to a single level of linguistic analysis. Change of order and addition/deletion are found here as specific indicator types, the former involving units with the same content expressed using different word orders, and the latter based on added or omitted information. Both indicators concern at least syntax and discourse; given the cross-level setup, the latter is particularly important for our datasets.

Beyond the linguistic structure, the quasi-linguistic domain captures inference-based similarity that relies on pragmatic information. The core linguistic meanings and the extralinguistic referents are different in this case, but the meaning of one element in the pair can still be inferred from the meaning of the other. Given the nature of our texts, this type of similarity is expected to be infrequent, and we have so far not identified any examples; however, we leave this category in our taxonomy to possibly be applied in the annotation phase. The extralinguistic domain also entails inequality of linguistic meaning, but it involves information equivalence between two texts, i.e. reference to the same real-world situation. It requires knowledge external to language for similarity to be recognised; this knowledge can be situational (containing elements such as *today* or *here*), encyclopaedic (involving general knowledge), or logical (requiring calculations or other similar operations). Based on the initial analyses of our datasets, this is a common type of similarity, especially in newswire texts.

Keeping the above definitions in mind, the outlined taxonomy will be applied to the *CLSS.news.sr* and *CLSS.codecomments.sr* corpora. Detailed guidelines are currently being developed, and the texts (initially from *CLSS.news.sr*) are being prepared for word/segment-level annotation with semantic similarity indicators, within the identified pairs. The annotation will be performed by the project researchers, first as a double procedure on a smaller sample, and then individually once a satisfactory level of agreement is reached. The initial phase will at the same time enable us to verify the appropriateness of the taxonomy, and adapt it should the need arise. The annotated datasets will be used for empirically validating the taxonomy, for gaining a better understanding of the linguistic factors that carry the most weight in cross-level semantic similarity in different text genres, and for learning how this kind of information can be taken into account in NLP models. Based on previous work on paraphrase and a preliminary exploration of our data at text level (with entire pairs marked for indicator presence/absence), morphological indicators, addition/deletion and same polarity items are expected to be particularly prominent.

5. Concluding remarks

In this paper, we have described the first non-English CLSS corpora, *CLSS.news.sr* and *CLSS.codecomments.sr*. The focus was on the methodology used to construct and annotate the data, as well as on their initial linguistic analysis. We believe these two datasets to be an important resource for Cross-Level Semantic Similarity research, not only in virtue of representing a new language, but also due to introducing an underexplored text genre (source code comments), and due to dedicating substantial attention to the linguistic properties of the datasets.

Our planned next steps are to complete the CLSS annotation of code comments, implement the proposed linguistic taxonomy of semantic similarity in the annotation of both datasets, conduct a more extensive linguistic analysis based on the annotated data, and examine the impact of linguistic traits on the performances of automatic CLSS models. Another goal is to compare the results to those obtained on similar datasets for English, using the *SemEval* dataset for newswire, and our own dataset (which is currently being created) for source code comments.

6. Acknowledgements

The AVANTES project (*Advancing Novel Textual Similarity-based Solutions in Software Development*) is supported by the Science Fund of the Republic of Serbia, grant no. 6526093, within the "Program for Development of Projects in the Field of Artificial Intelligence". The authors would like to thank Jelica Cincović and Dušan Stojković for constructing the code comment text pairs, as well as Bojan Jakovljević, Lazar Milić, Marija Lazarević, Ognjen Krešić, and Vanja Miljković for annotating the corpora with semantic similarity scores.

7. References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez- Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 385–393, Montreal, Canada. Association for Computational Linguistics.

Vuk Batanović. 2020. *A Methodology for Solving Semantic Tasks in the Processing of Short Texts Written in Natural Languages with Limited Resources*. Ph.D. thesis, University of Belgrade.

Vuk Batanović. 2021. Semantic similarity and sentiment analysis of short texts in Serbian. In: *Proceedings of the 29th Telecommunications forum (TELFOR 2021)*, Belgrade, Serbia, IEEE.

Vuk Batanović and Maja Miličević Petrović. 2022. Cross-Level Semantic Similarity for Serbian Newswire Texts. In: *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France. European Language Resources Association.

Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. 2018. Fine-grained Semantic Textual Similarity for Serbian. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1370–78, Miyazaki, Japan, European Language Resources Association.

Vuk Batanović, Bojan Furlan, and Boško Nikolić. 2011. A software system for determining the semantic similarity of short texts in Serbian. In: *Proceedings of the 19th Telecommunications forum (TELFOR 2011)*, pages 1249–52, Belgrade, Serbia, IEEE.

Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Towards automatic fake news detection: Cross-level stance detection in news articles. In: *Proceedings of the First Workshop on Fact Extraction and VERification*, pages 40–49, Brussels, Belgium, Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACL-HLT 2019*, pages 4171–86, Minneapolis, Minnesota, USA, Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora. In: *Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–56, Geneva, Switzerland, Association for Computational Linguistics.

Bojan Furlan, Vuk Batanović, and Boško Nikolić. 2013. Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems*, 55(3):710–19.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. SemEval-2014 Task 3: Cross-Level Semantic Similarity. In: *Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)*, pages 17–26, Dublin, Ireland. Association for Computational Linguistics.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Cross Level Semantic Similarity: An Evaluation Framework for Universal Measures of Similarity. *Language Resources and Evaluation*, 50(1):5–33.

Marija Kostić, Aleksa Srbljanović, Vuk Batanović, and Boško Nikolić. 2022. Code Comment Classification Taxonomies. In: *Proceedings of the Ninth IcETRAN Conference*, Novi Pazar, Serbia.

Nikola Ljubešić and Davor Lauc. 2021. BERTić – The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In: *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2021)*, pages 37–42, Kiev, Ukraine, Association for Computational Linguistics.

Igor A. Mel'čuk. 2012. *Semantics. From Meaning to Text*. John Benjamins, Amsterdam.

Maja Miličević Petrović, Nikola Ljubešić, and Darja Fišer. 2017. Nestandardno zapisivanje srpskog jezika na Tviteru: mnogo buke oko malo odstupanja? *Anali Filološkog fakulteta* 29(2):111–36.

Jasmina Milićević. 2007. *La paraphrase*. Peter Lang, Bern.

Maâli Mnasri, Gaël de Chalendar, and Olivier Ferret. 2017. Taking into account Inter-sentence Similarity for Update Summarization. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 204–209, Taipei, Taiwan. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Roberto Navigli. 2015. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128.

Navid Rekabsaz, Ralf Bierig, Mihai Lupu, and Allan Hanbury. 2017. Toward optimized multimodal concept indexing. In: N. Nguyen, R. Kowalczyk, A. Pinto, and J. Cardoso, eds., *Transactions on Computational Collective Intelligence XXVI*, pages 144–61, Cham, Springer International Publishing.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In: *Proceedings of the Third Workshop on Machine Reading for Question Answering*, pages 149–57, Punta Cana, Dominican Republic, Association for Computational Linguistics.

Marta Vila Rigat. 2013. *Paraphrase Scope and Typology. A Data-Driven Approach from Computational Linguistics*. Ph.D. thesis, University of Barcelona.

Marta Vila, M. Antonia Martí, and Horacio Rodríguez 2014. Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4:205–18.

Marie Zemankova and Caroline M. Eastman. 1980. Comparative lexical analysis of FORTRAN code, code comments and English text. In: *Proceedings of the 18th annual Southeast regional conference*, pages 193–97, Tallahassee, Florida, USA, Association for Computing Machinery.

The ParlaSent-BCS Dataset of Sentiment-annotated Parliamentary Debates from Bosnia and Herzegovina, Croatia, and Serbia

Michal Mochtak,[*] Peter Rupnik,[†] Nikola Ljubešić[†‡]

[*]Institute of Political Science
University of Luxembourg
2 avenue de l'Université, L-4366 Esch-sur-Alzette
michal.mochtak@uni.lu

[†] Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
peter.rupnik@ijs.si
nikola.ljubesic@ijs.si

[‡]Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, SI-1000 Ljubljana

Abstract

Expression of sentiment in parliamentary debates is deemed to be significantly different from that on social media or in product reviews. This paper adds to an emerging body of research on parliamentary debates with a dataset of sentences annotated for detection of sentiment polarity in political discourse using sentence-level data. We sample the sentences for annotation from the proceedings of three Southeast European parliaments: Croatia, Bosnia and Herzegovina, and Serbia. A six-level annotation schema is applied to the data with the aim of training a classification model for the detection of sentiment in parliamentary proceedings. Krippendorff's alpha measuring the inter-annotator agreement ranges from 0.6 for the six-level annotation schema to 0.75 for the three-level schema and 0.83 for the two-level schema. Our initial experiments on the dataset show that transformer models perform significantly better than those using a simpler architecture. Furthermore, regardless of the similarity of the three languages, we observe differences in performance across different languages. Performing parliament-specific training and evaluation shows that the main reason for the differing performance between parliaments seems to be the different complexity of the automatic classification task, which is not observable in annotator performance. Language distance does not seem to play any role neither in annotator nor in automatic classification performance. We release the dataset and the best-performing models under permissive licences.

1. Introduction

Emotions and sentiment in political discourse are deemed as crucial and influential as substantive policies promoted by the elected representatives (Young and Soroka, 2012). Since the golden era of research on propaganda (Lasswell, 1927; Shils and Janowitz, 1948), a number of scholars have demonstrated the growing role of emotions on affective polarization in politics with negative consequences for the stability of democratic institutions and the social cohesion (Garrett et al., 2014; Iyengar et al., 2019; Mason, 2015). With the booming popularity of online media, sentiment analysis has become an indispensable tool for understating the positions of viewers, customers, but also voters (Soler et al., 2012). It has allowed all sorts of entrepreneurs to know their target audience like never before (Ceron et al., 2019). Experts on political communication argue that the way we receive information and how we process them play an important role in political decision-making, shaping our judgment with strategic consequences both on the level of legislators and the masses (Liu and Lei, 2018). Emotions and sentiment simply do play an important role in political arenas and politicians have been (ab)using them for decades.

Although there is a general agreement among political scientists that sentiment analysis represents a critical component for understanding political communication in general (Young and Soroka, 2012; Flores, 2017; Tumasjan et al., 2010), the empirical applications outside the English-speaking world are still rare (Rauh, 2018; Mohammad, 2021). This is especially the case for studies analyzing political discourse in low-resourced languages, where the lack of out-of-the-box tools creates a huge barrier for social scientists to do such research in the first place (Proksch et al., 2019; Mochtak et al., 2020; Rauh, 2018). The paper, therefore, aims to contribute to the stream of applied research on sentiment analysis in political discourse in low-resourced languages. The goal is to present a new annotated dataset compiled for machine-learning applications focused on the detection of sentiment polarity in the political discourse of three Southeast European (SEE) countries: Bosnia and Herzegovina, Croatia, and Serbia. We further use the dataset to train different classification models for the sentiment analysis applying different schemas and settings to demonstrate the benefits and limitations of the dataset and the trained models. We release the dataset and the best-performing models under permissive licenses to facilitate

further research and more empirically oriented projects. In general, the paper, the dataset, and the models contribute to an emerging community of research outputs on parliamentary debates with a focus on sentence-level sentiment annotation with future downstream applications in mind.

2. Dataset construction

2.1. Focus on sentences

The dataset we compile and then use for training different classification models focuses on a sentence-level data and utilizes sentence-centric approach for capturing sentiment polarity. The strategy goes against the tradition in mainstream research applications in social sciences which focus either on longer pieces of text (e.g. utterance of "speech segment" or whole documents (Bansal et al., 2008; Thomas et al., 2006)) or coherent messages of shorter nature (e.g. tweets (Tumasjan et al., 2010; Flores, 2017)). The approach, however, creates certain limitations when it comes to political debates in national parliaments where speeches range from very short comments counting only a handful of sentences to long monologues having thousands of words. Moreover, as longer text may contain a multitude of sentiments, any annotation attempt must generalize them, introducing a complex coder bias which is embedded in any subsequent analysis. The sentence-centric approach attempts to refocus the attention on individual sentences capturing attitudes, emotions, and sentiment positions and using them as lower-level indices of sentiment polarity in a more complex political narrative. Although sentences cannot capture complex meanings as paragraphs or whole documents do, they usually carry coherent ideas with relevant sentiment affinity. This approach stems from a tradition of content analysis in political science which focuses both on the political messages and their role in political discourse in general (Burst et al., 2022; Hutter et al., 2016; Koopmans and Statham, 2006).

Unlike most of the literature which approaches sentiment analysis in political discourse as a proxy for position-taking stances or as a scaling indicator (Abercrombie and Batista-Navarro, 2020b; Glavaš et al., 2017; Proksch et al., 2019), a general sentence-level classifier we aim for in this paper has a more holistic (and narrower) aim. Rather than focusing on a specific policy or issue area, the task is to assign a correct sentiment category to sentence-level data in political discourse with the highest possible accuracy. Only when a good performing model exists, a downstream task can be discussed. We believe it is a much more versatile approach which opens a wide range of possibilities for understanding the context of political concepts as well as their role in political discourse. Furthermore, sentences as lower semantic units can be aggregated to the level of paragraphs or whole documents which is often impossible the other way around (document $\rightarrow$ sentences). Although sentences as the basic level of analysis are less common in social sciences research when it comes to computational methods (Abercrombie and Batista-Navarro, 2020b), practical applications in other areas exist covering topics such as validation of sentiment dictionaries (Rauh, 2018), ethos mining (Duthie and Budzynska, 2018), opinion mining (Naderi and Hirst, 2016), or detection of sentiment carrying sentences (Onyimadu et al., 2013).

2.2. Background data

In order to compile a dataset of political sentiment for manual annotation and then use it for training the classification models for real world applications, we sampled sentences from three corpora of parliamentary proceedings in the region of former Yugoslavia – Bosnia and Herzegovina (Mochtak et al., 2022c),[1] Croatia (Mochtak et al., 2022a),[2] and Serbia (Mochtak et al., 2022b).[3] The Bosnian corpus contains speeches collected on the federal level from the official website of the Parliamentary Assembly of Bosnia and Herzegovina (Parlamentarna skupština BiH, 2020). Both chambers are included – House of Representatives (Predstavnički dom / Zastupnički dom) and House of Peoples (Dom naroda). The corpus covers the period from 1998 to 2018 (2nd – 7th term) and counts 127,713 speeches. The Croatian corpus of parliamentary debates covers debates in the Croatian parliament (Sabor) from 2003 to 2020 (5th – 9th term) and counts 481,508 speeches (Hrvatski sabor, 2020). Finally, the Serbian corpus contains 321,103 speeches from the National Assembly of Serbia (Skupština) over the period of 1997 to 2020 (4th – 11th term) (Otvoreni Parlament, 2020).

2.3. Data sampling

Each speech was processed using the CLASSLA-Stanza tool (Ljubešić and Dobrovoljc, 2019) with tokenizers available for Croatian and Serbian in order to extract individual sentences as the basic unit of our analysis. In the next step, we filtered out only sentences presented by actual speakers, excluding moderators of the parliamentary sessions. All sentences were then merged into one meta dataset. As we want to sample what can be understood as "average sentences", we further subset the sentence meta corpus to only sentences having the number of tokens within the first and third frequency quartile (i.e. being within the interquartile range) of the original corpus ($\sim$3.8M sentences). Having the set of "average sentences", we used the Croatian gold standard sentiment lexicon created by (Glavaš et al., 2012), translated it to Serbian with a rule-based Croatian-Serbian translator (Klubička et al., 2016), combined both lexicons, and extracted unique entries with a single sentiment affinity, and used them as seed words for sampling sentences for manual annotation. The final pool of seed words contains 381 positive and 239 negative words (neutral words are excluded). These seed words are used for stratified random sampling which gives us 867 sentences with negative seed word(s), 867 sentences with positive seed word(s), and 866 sentences with neither positive nor negative seed words (supposedly having neutral sentiment). We sample 2600 sentences in total for manual annotation. The only strata we use is the size of the original corpora (i.e. number of sentences per corpus). With this we sample 1,388 sentences from the Croatian parliament, 1059

[1] https://doi.org/10.5281/zenodo.6517697
[2] https://doi.org/10.5281/zenodo.6521372
[3] https://doi.org/10.5281/zenodo.6521648

sentences from the Serbian parliament, and 153 sentences from the Bosnian parliament.

2.4. Annotation schema

The annotation schema for labelling sentence-level data was adopted from Batanović et al. (Batanović et al., 2020) who propose a six-item scale for annotation of sentiment polarity in a short text. The schema was originally developed and applied to SentiComments.SR, a corpus of movie comments in Serbian and is particularly suitable for low-resourced languages. The annotation schema contains six sentiment labels (Batanović et al., 2020: 6):

- +1 (`Positive` in our dataset) for sentences that are entirely or predominantly positive
- –1 (`Negative` in our dataset) for sentences that are entirely or predominantly negative
- +M (`M_Positive` in our dataset) for sentences that convey an ambiguous sentiment or a mixture of sentiments, but lean more towards the positive sentiment in a strict binary classification
- –M (`M_Negative` in our dataset) for sentences that convey an ambiguous sentiment or a mixture of sentiments, but lean more towards the negative sentiment in a strict binary classification
- +NS (`P_Neutral` in our dataset) for sentences that only contain non-sentiment-related statements, but still lean more towards the positive sentiment in a strict binary classification
- –NS (`N_Neutral` in our dataset) for sentences that only contain non-sentiment-related statements, but still lean more towards the negative sentiment in a strict binary classification

The different naming convention we have applied in our dataset serves primarily practical purposes: obtaining the 3-way classification by taking under consideration only the second part of the string (if an underscore is present).

Additionally, we also follow the original schema which allowed marking text deemed as sarcastic with a code "sarcasm". The benefit of the whole annotation logic is that it was designed with versatility in mind allowing reducing the sentiment label set in subsequent processing if needed. That includes various reductions considering polarity categorization, subjective/objective categorization, change of the number of categories, or sarcasm detection. This is important for various empirical tests we perform in the following sections.

2.5. Data annotation

Data were annotated in two waves, with 1300 instances being annotated in each. Annotation was done via a custom online app. The first batch of 1300 sentences was annotated by two annotators, both being native speakers of Croatian, while the second batch was annotated only by one of them.

parliament	positive	neutral	negative
all	470	772	1358
HR	261	433	694
BS	27	42	84
SR	182	297	580

Table 1: Distribution of the three-class labels in the whole dataset, as well as across each of the three parliaments.

The inter-annotator agreement (IAA) measured using Krippendorff's alpha in the first round was 0.599 for full six-item annotation scheme, 0.745 for the three-item annotation schema (positive/negative/neutral), and 0.829 for the two-item annotation schema focused on the detection of only negative sentiment (negative/other). The particular focus on negative sentiment in the test setting is inspired by a stream of research in political communication which argues that negative emotions appear to be particularly prominent in the context of forming the human psyche and its role in politics (Young and Soroka, 2012). More specifically, political psychologists have found that negative political information has a more profound effect on attitudes than positive information as it is easier to recall and is more useful in heuristic cognitive processing for simpler tasks (Baumeister et al., 2001; Utych, 2018).

Before the second annotator moved to annotate the second batch of instances, hard disagreements, i.e. disagreements pointing at a different three-class sentiment, where +NS and -NS are considered neutral, were resolved together by both annotators through a reconciliation procedure.

The final distribution of the three-class labels in the whole dataset, as well as along specific parliaments, is given in Table 1. The presented distributions show that, regardless of a lexicon-based sampling, the negative class is still by far the most pervasive category, which might be even more the case in a randomly sampled dataset, something we leave for future work.

2.6. Dataset encoding

The final dataset, available through the CLARIN.SI repository, contains the following metadata:

- `sentence` that is annotated
- `country` of origin of the sentence
- annotation `round` (first, second)
- annotation of `annotator1` with one of the labels from the annotation schema presented in Section 2.4.
- annotation of `annotator2` following the same annotation schema
- annotation given during `reconciliation` of hard disagreements
- the three-way `label` (positive, negative, neutral) where +NS and -NS labels are mapped to the neutral class

- the `document_id` the sentence comes from
- the `sentence_id` of the sentence
- the `date` the speech was given
- the `name`, `party`, `gender`, `birth_year` of the speaker
- the `split` (train, dev, or test) the instance has been assigned to (described in more detail in Section 3.1.

The final dataset is organized in a JSONL format (each line in the file being a JSON entry) and is available under the CC-BY-SA 4.0 license.[4]

3. Experiments

3.1. Data splits

For performing current and future experiments, the dataset was split into the train, development and test subsets. The development subset consists of 150 instances, while the test subset consists of 300 instances, both using instances from the first annotation round, where two annotations per instance and hard disagreement reconciliations are available. The training data consists of the remainder of the data from the first annotation round and all instances from the second annotation round, summing to 2150 instances.

While splitting the data, stratification was performed on the variables of three-way sentiment, country, and party. With this we can be reasonably sure that no specific strong bias regarding sentiment, country or political party is present in any of the three subsets.

3.2. Experimental setup

In our experiments we investigate the following questions: (1) how well can different technologies learn our three-way classification task, (2) what is the difference in performance depending on which parliament the model is trained or tested on, and (3) is the annotation quality of the best performing technology high enough to be useful for data enrichment and analysis.

We investigate our first question by comparing the results on the following classifiers: fastText (Joulin et al., 2016) with pre-trained CLARIN.SI word embeddings (Ljubešić, 2018), the multilingual transformer model XLM-Roberta (Conneau et al., 2019),[5] the transformer model pre-trained on Croatian, Slovenian and English cseBERT (Ulčar and Robnik-Šikonja, 2020),[6], and the transformer model pre-trained on Croatian, Bosnian, Montenegrin and Serbian BERTić (Ljubešić and Lauc, 2021).[7] Our expectation is for the last model to perform best given that it was pre-trained on most data from the three languages. However, this assumption has to be checked given that for some tasks even models pre-trained on many languages obtain performance that is comparable to otherwise superior models pre-trained on one or few languages (Kuzman et al., 2022).

model	macro F1
classla/bcms-bertic	**0.7941 ± 0.0101****
EMBEDDIA/crosloengual-bert	0.7709 ± 0.0113
xlm-roberta-base	0.7184 ± 0.0139
fasttext + CLARIN.SI embeddings	0.6312 ± 0.0043

Table 2: Results of the comparison of various text classification technologies. We report macro-F1 mean and standard deviation over 6 runs with the model-specific optimal number of training epochs. The distributions of results of the two best performing models are compared with the Mann-Whitney U test (** $p < 0.01$).

While comparing the different classification techniques, each model was optimized for the epoch number hyperparameter on the development data, while all other hyperparameters were kept default. For training transformers, the simpletransformers library[8] was used.

The second question on parliament specificity we answer by training separate models on Croatian sentences only and Serbian sentences only, evaluating each model both on Croatian and on Serbian test sentences. We further evaluate the model trained on all training instances separately on instances coming from each of the three parliaments.

For our third question on the usefulness of the model for data analysis, we report confusion matrices, to inform potential downstream users of the model's per-category performance.

4. Results

4.1. Classifier comparison

We report the results of our text classification technology comparison in Table 2. The results show that transformer models are by far more capable than the fasttext technology relying on static embeddings only. Of the three transformer models, the multilingual XLM-RoBERTa model shows to have a large gap in performance to the two best-performing models. Comparing the cseBERT and the BERTić model, the latter manages to come on top with a moderate improvement of 1.5 points in macro-F1. The difference in the results of the two models is statistically significant regarding the Mann-Whitney U test (Mann and Whitney, 1947), with a p-value of 0.0053.

4.2. Parliament dependence

We next investigate the dependence of the results on from which parliament the training and the testing data came. Our initial assumption was that the results are dependent on whether the training and the testing data come from the same or a different parliament, with same-parliament results being higher. We also investigate how the model trained on all data performs on parliament-specific test data.

[4] http://hdl.handle.net/11356/1585

[5] https://huggingface.co/xlm-roberta-base

[6] https://huggingface.co/EMBEDDIA/crosloengual-bert

[7] https://huggingface.co/classla/bcms-bertic

[8] https://simpletransformers.ai

4.2.1. Impact of training data

We perform this analysis on all three transformer models from Section 4.1., hoping to obtain a deeper understanding of parliament dependence on our task. We train and test on data from the Croatian and the Serbian parliament only as the Bosnian parliament's data are not large enough to enable model training.

In Table 3 we report the results grouped by model and training and testing parliament. To our surprise, the strongest factor shows not to be whether the training and testing data come from the same parliament, but what testing data are used, regardless of the training data. This trend is to be observed regardless of the model used.

The results show that Serbian test data seem to be harder to classify, regardless of what training data are used, with a difference of ∼9 points in macro-F1 for the BERTić and the XLM-RoBERTa models. The difference is smaller for the cseBERT model, ∼7 points, but still shows the same trend as the two other models.

We have additionally explored the possibility of a complexity bias of Serbian test data in comparison to Serbian training data by performing different data splits, but the results obtained were very similar to those presented here. Serbian data seem to be harder to classify in general, which is observed when performing inference over Serbian data. Training over Serbian data still results in a model comparably strong to that based on Croatian training data. Important to note is that the Croatian data subset is 30% larger than the Serbian one.

To test whether the Serbian data complexity goes back to challenges during data annotation, or whether it is rather the models that struggle with inference over Serbian data, we calculated the Krippendorff IAA on data from each parliament separately. The agreement calculation over the ternary classification schema resulted in an IAA for Bosnian data of 0.69, Croatian data of 0.733, and Serbian data of 0.77. This insight proved that annotators themselves did not struggle with Serbian data as these had the highest IAA. We also tested whether there is excessive sarcasm in Serbian data, which might affect the model's performance. The dataset contains two sarcastic instances from the parliament of Bosnia and Herzegovina and 16 for both Croatia and Serbia, which means sarcasm can hardly explain the overall lower performance on Serbian test data. Lastly, we checked the type-token ratio (TTR) on samples of Croatian and Serbian sentences to estimate the lexical richness of each subset, a higher lexical richness of Serbian (via a higher type-token ratio) possibly explaining the lower results obtained on Serbian test data. By calculating the type-token ratio on 100 tokens selected from random sentences, and repeating the process 100 times in a bootstrapping manner, we obtained a result of 0.833 for Serbian and 0.839 for Croatian. This result shows for the Croatian part of the dataset to be just slightly more lexically rich (83.9 different tokens among 100 tokens on average) than Serbian (83.3 different tokens among 100 tokens), which does not explain the difference in performance of various classifiers on Serbian data.

The complexity of Serbian data that can be observed in the evaluation is due to some effect that we did not manage to identify at this point, but that will have to be taken under consideration in future work on this dataset.

XLM-RoBERTa		
train \test	HR	SR
HR	0.7296 ± 0.0251	0.6128 ± 0.0341
SR	0.7323 ± 0.0282	0.6487 ± 0.0203
cseBERT		
train \test	HR	SR
HR	0.7748 ± 0.0174	0.7146 ± 0.0175
SR	0.7762 ± 0.0114	0.6989 ± 0.0275
BERTić		
train \test	HR	SR
HR	0.8147 ± 0.0083	0.7249 ± 0.0105
SR	0.7953 ± 0.0207	0.7130 ± 0.0278

Table 3: Comparison of the three transformer models when trained and tested on data from the Croatian or Serbian parliament. Average macro-F1 and standard deviation over 6 runs is reported.

test	ternary	binary
all	0.7941 ± 0.0101	0.8999 ± 0.0120
HR	0.8260 ± 0.0186	0.9221 ± 0.0153
BS	0.7578 ± 0.0679	0.9071 ± 0.0525
SR	0.7385 ± 0.0170	0.8660 ± 0.0150

Table 4: Average macro-F1 and standard deviation of 6 runs of the BERTić model, trained on all training data, and evaluated on varying testing data.

4.2.2. Impact of testing data

In the next set of experiments, we compare the performance of BERTić classifiers trained over all training data, but evaluated on all and per-parliament testing data. Beyond this, we train models over the ternary schema that we have used until now (positive vs. neutral vs. negative), but also the binary schema (negative vs. rest), given our special interest in identifying negative sentences, as already discussed in Section 2.5.

We report results on test data from each of the three parliaments, including the Bosnian one, which, however, contains only 18 testing instances, so these results have to be taken with caution.

The results presented in Table 4 show again that the Serbian data seem to be the hardest to classify even when all training data are used. Bosnian results are somewhat close to the Serbian ones, but caution is required here due to the very small test set. This level of necessary caution regarding Bosnian test data is also visible from the five times higher standard deviation in comparison to the results of the two other parliaments. Croatian data seem to be easiest to classify, with an absolute difference of 9 points between the performance on Serbian and Croatian test data. Regarding the binary classification results, these are, as expected, higher than those of the ternary classification schema with an macro-F1 of 0.9 when all data are used. The relationship between specific parliaments is very similar to that observed using the ternary schema.

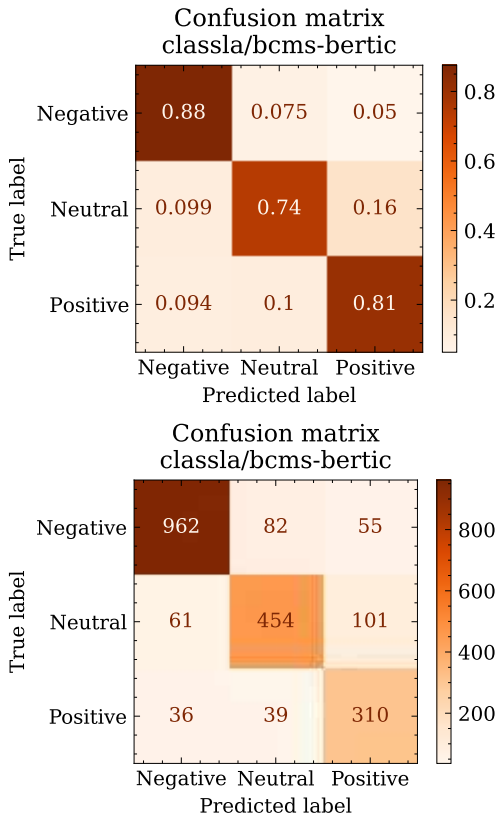


Figure 1: Row-normalised and raw-count confusion matrix of the BERTić results on the ternary schema.

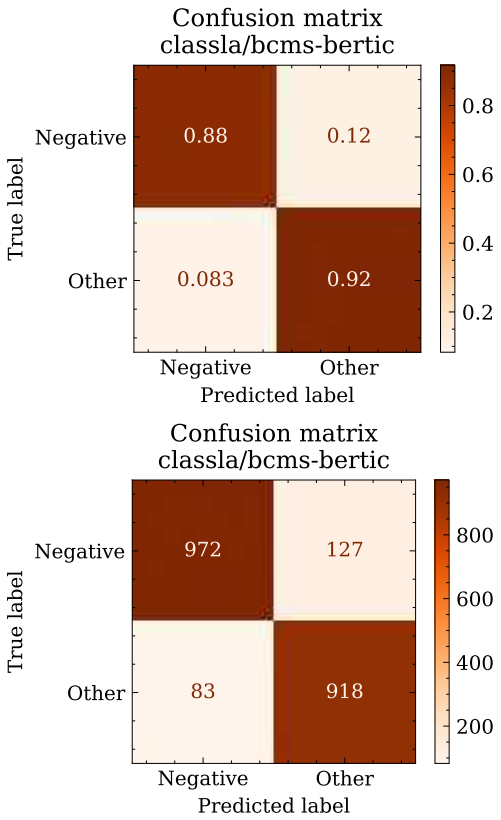


Figure 2: Row-normalised and raw-count confusion matrix of the BERTić results on the binary schema.

4.3. Per-category analysis

Our final set of experiments investigates the per-category performance both on the ternary and the binary classification schema. We present the confusion matrices on the ternary schema, one row-normalized, another with raw counts, in Figure 1. As anticipated, the classifier works best on the negative class, with 88% of negative instances properly classified as negative. Second by performance is the positive class with 81% of positive instances being labelled like that, while among the neutral instances 3 out of 4 instances are correctly classified. Most of the confusion between classes occurs, as expected, between the neutral and either of the two remaining classes.

The binary confusion matrices, presented in Figure 2 show for a rather balanced performance on both categories. On each of the categories recall is around 0.9, with a similar precision given the symmetry of the confusions.

When comparing the output of the ternary and the binary model, the ternary model output mapped to a binary schema performs slightly worse than the binary model, meaning that practitioners should apply the binary model if they are interested just in distinguishing between negative and other sentences.

Although any direct comparisons are hard to make, the few existing studies which performed text classification on sentence-level data, report much worse results. Rauh (2018) found that when three annotators and three sentiment dictionaries were compared on a ternary classification task (positive/negative/neutral), they agreed only in one-quarter of the 1,500 sentences. Using heuristic classifiers based on the use of statistical and syntactic clues, Onyimadu et al. (2013) found that on average, only 43% of the sentences were correctly annotated for their sentiment affinity. The results of our experiments are therefore certainly promising. Especially when it comes to the classification of negative sentences, the model has 1 in 10 sentence error rate which is almost on par with the quality of annotation performed by human coders.

5. Conclusion

The paper introduces a sentence-level dataset of parliamentary proceedings, manually annotated for sentiment via a six-level schema. The good inter-annotator agreement is reported, and the first results on the automation of the task are very promising, with a macro-F1 of ∼0.8 on the ternary schema and ∼0.9 on the binary schema. The difference in performance across the three parliaments is observed, but visible only during inference, Serbian data being harder to make predictions on, while for modelling, all parliaments seem to be similarly useful. One limitation of our work is the following: our testing data have been sampled as the whole dataset, with a bias towards mid-length sentences, and sentences containing sentiment words. Future work should consider preparing a sample of random sentences, or, even better, consecutive sentences, so that the potential issue of lack of a wider context during manual data annotation is successfully mitigated as well.

In general, the reported results have several promising implications for applied research in political science. First of all, it allows a more fine-grained analysis of political concepts and their context. A good example is a combination of the KWIC approach with sentiment analysis, with a focus on examining the tone of a message in political discourse. This is interesting for both qualitatively and quantitatively oriented scholars. Especially the possibility of extracting numeric assessment of the classification model (e.g. class probability) is particularly promising for all sorts of hypothesis-testing statistical models. Moreover, sentence-level analysis can be combined with the findings of various information and discourse theories for studying political discourse focused on rhetoric and narratives (e.g. beginning and end of a speech are more relevant than what comes in the middle). Apart from the concept-driven analysis, the classification model can be used for various research problems ranging from policy position-taking to ideology detection or general scaling tasks (Abercrombie and Batista-Navarro, 2020a; Glavaš et al., 2017; Proksch et al., 2019). Although each of these tasks requires proper testing, the performance of the trained models for such applications is undoubtedly promising.

As a part of our future work, we plan to test the usefulness of the predictions on a set of downstream tasks. The goal is to analyze the data from all three parliaments (Bosnia and Herzegovina, Croatia, and Serbia) in a series of tests focused on replication of the results from the existing research using mostly English data. Given the results we obtained, we aim to continue our research using the setup with the model trained on cross-country data. Furthermore, the three corpora we have used in this paper will be extended as a part of ParlaMint II project.

We make the ternary and binary BERTić models trained on all available training available via the HuggingFace repository[9][10] and make the dataset available through the CLARIN.SI repository (Mochtak et al., 2022d).

Acknowledgements

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341 (MaCoCu project). This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).

6. References

Gavin Abercrombie and Riza Batista-Navarro. 2020a. ParlVote: A corpus for sentiment analysis of political debates. In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France. European Language Resources Association.

Gavin Abercrombie and Riza Batista-Navarro. 2020b. Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *Journal of Computational Social Science*, 3(1):245–270.

Mohit Bansal, Claire Cardie, and Lillian Lee. 2008. The power of negative thinking: Exploiting label disagreement in the Min-cut classification framework. In: *Coling 2008: Companion volume: Posters*, pages 15–18, Manchester, UK. Coling 2008 Organizing Committee.

Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. 2020. A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts. *PLOS ONE*, 15(11):e0242050.

Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. Bad is Stronger than Good. *Review of General Psychology*, 5(4):323–370.

Tobias Burst, Werner Krause, Pola Lehmann, Jirka Lewandowski, Theres Matthieß, Nicolas Merz, Sven Regel, and Lisa Zehnter. 2022. Manifesto corpus.

Andrea Ceron, Luigi Curini, and Stefano M Iacus. 2019. *Politics and Big Data: Nowcasting and Forecasting Elections with Social Media*. Routledge, Abingdon, New York.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Rory Duthie and Katarzyna Budzynska. 2018. A deep modular rnn approach for ethos mining. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4041–4047. AAAI Press.

René D. Flores. 2017. Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 Using Twitter Data. *American Journal of Sociology*, 123(2):333–384.

R. Kelly Garrett, Shira Dvir Gvirsman, Benjamin K. Johnson, Yariv Tsfati, Rachel Neo, and Aysenur Dal. 2014. Implications of pro- and counterattitudinal information exposure for affective polarization. *Human Communication Research*, 40(3):309–332.

Goran Glavaš, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Semi-supervised acquisition of Croatian sentiment lexicon. In: *International Conference on Text, Speech and Dialogue*, pages 166–173. Springer.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Unsupervised cross-lingual scaling of political texts. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 688–693, Valencia, Spain. Association for Computational Linguistics.

Hrvatski sabor. 2020. eDoc. http://edoc.sabor.hr/.

Swen Hutter, Edgar Grande, and Hanspeter Kriesi. 2016. *Politicising Europe: Integration and mass politics.* Cambridge University Press, Cambirdge.

[9] https://huggingface.co/classla/bcms-bertic-parlasent-bcs-ter

[10] https://huggingface.co/classla/bcms-bertic-parlasent-bcs-bi

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. 2019. The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 22(1):129–146.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Filip Klubička, Gema Ramírez-Sánchez, and Nikola Ljubešić. 2016. Collaborative development of a rule-based machine translator between croatian and serbian. In: *Proceedings of the 19th Annual Conference of the Eu- ropean Association for Machine Translation*, pages 361– 367.

Ruud Koopmans and Paul Statham. 2006. Political Claims Analysis: Integrating Protest Event and Political Discourse Approaches. *Mobilization: An International Quarterly*, 4(2):203–221.

Taja Kuzman, Peter Rupnik, and Nikola Ljubesic. 2022. The ginco training dataset for web genre identification of documents out in the wild. *ArXiv*, abs/2201.03857.

Harold Dwight Lasswell. 1927. *Propaganda Technique in the World War*. Peter Smith, New York.

Dilin Liu and Lei Lei. 2018. The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election. *Discourse, Context & Media*, 25:143–152.

Nikola Ljubešić and Kaja Dobrovoljc. 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*, pages 29–34, Florence, Italy, August. Association for Computational Linguistics.

Nikola Ljubešić and Davor Lauc. 2021. BERTić – the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In: *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine, April. Association for Computational Linguistics.

Nikola Ljubešić. 2018. Word embeddings CLARIN.SI-embed.hr 1.0. Slovenian language resource repository CLARIN.SI.

Henry B. Mann and Donald R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Lilliana Mason. 2015. "I Disrespectfully Agree": The Differential Effects of Partisan Sorting on Social and Issue Polarization. *American Journal of Political Science*, 59(1):128–145.

Michal Mochtak, Josip Glaurdić, and Christophe Lesschaeve. 2020. Talking War: Representation, Veterans and Ideology in Post-War Parliamentary Debates. *Government and Opposition*, 57(1):148–170.

Michal Mochtak, Josip Glaurdić, and Christophe Lesschaeve. 2022a. CROCorp: Corpus of Parliamentary Debates in Croatia (v1.1.1). https://doi.org/10.5281/zenodo.6521372.

Michal Mochtak, Josip Glaurdić, and Christophe Lesschaeve. 2022b. SRBCorp: Corpus of Parliamentary Debates in Serbia (v1.1.1). https://doi.org/10.5281/zenodo.6521648.

Michal Mochtak, Josip Glaurdić, Christophe Lesschaeve, and Ensar Muharemović. 2022c. BiHCorp: Corpus of Parliamentary Debates in Bosnia and Herzegovina (v1.1.1). https://doi.org/10.5281/zenodo.6517697.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2022d. The sentiment corpus of parliamentary debates ParlaSent-BCS v1.0. Slovenian language resource repository CLARIN.SI.

Saif M. Mohammad. 2021. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. https://arxiv.org/abs/2005.11882.

Nona Naderi and Graeme Hirst. 2016. Argumentation mining in parliamentary discourse. In: Matteo Baldoni, Cristina Baroglio, Floris Bex, Floriana Grasso, Nancy Green, Mohammad-Reza Namazi-Rad, Masayuki Numao, and Merlin Teodosia Suarez, editors, *Principles and Practice of Multi-Agent Systems*, pages 16–25, Cham. Springer.

Obinna Onyimadu, Keiichi Nakata, Tony Wilson, David Macken, and Kecheng Liu. 2013. Towards sentiment analysis on parliamentary debates in Hansard. In: *Revised Selected Papers of the Third Joint International Confer- ence on Semantic Technology – Volume 8388*, JIST 2013, page 48–50, Berlin, Heidelberg. Springer-Verlag.

Otvoreni Parlament. 2020. Početna. https://otvoreniparlament.rs/.

Parlamentarna skupština BiH. 2020. Sjednice. https://www.parlament.ba/?lang=bs.

Sven-Oliver Proksch, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative Studies Quarterly*, 44(1):97–131.

Christian Rauh. 2018. Validating a sentiment dictionary for German political language—a workbench note. *Journal of Information Technology & Politics*, 15(4):319–343.

Edward A. Shils and Morris Janowitz. 1948. Cohesion and Disintegration in the Wehrmacht in World War II. *Public Opinion Quarterly*, 12(2):315.

Juan M. Soler, Fernando Cuartero, and Manuel Roblizo. 2012. Twitter as a tool for predicting elections results. In: *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1194–1200.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney. Association for Computational Linguistics.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment.

Proceedings of the International AAAI Conference on Web and Social Media, 4(1).

Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In: P. Sojka, I. Kopeček, K. Pala, and A. Horák, eds., *Text, Speech, and Dialogue TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*. Springer.

Stephen M. Utych. 2018. Negative Affective Language in Politics. *American Politics Research*, 46(1):77–102.

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.

Fine-grained human evaluation of NMT applied to literary text: case study of a French-to-Croatian translation

Marta Petrak,* Mia Uremović,* Bogdanka Pavelin Lešić*

* Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb
mpetrak@ffzg.hr
uremovic.mia@gmail.com
bpavelin@ffzg.hr

Abstract

Even though neural machine translation (NMT) has demonstrated phenomenal results and has shown to be more successful than previous MT systems, there is not a large number of works dealing with its application to literary text. This results from the fact that literary texts are deemed to be more complex than others because they involve more specific elements such as idiomatic expressions, metaphor, a specific author's style, etc. Regardless of this fact, there is a growing body of research dealing with NMT applied to literary texts, and this case study is one of them. The goal of the present paper is to conduct an in-depth, fine-grained evaluation of a novel translated by Google Translate (GT) in order to reach detailed insights into NMT performance on literary text. In addition, the paper aims to include for the first time, to the best of our knowledge, the French-Croatian language combination.

1. Introduction

Numerous studies have demonstrated that neural machine translation (NMT) outperforms previous MT systems (e.g. Bentivogli et al., 2016; Burchardt et al., 2017; Klubička et al., 2018; Hansen, 2021). This has been demonstrated for a number of various text types, among which literary texts are the least represented due to their specificities such as lexical richness, metaphorical and idiomatic elements (e.g. Toral and Way, 2018). Literary translation is also usually considered to be more complex than technical translation because it includes elements such as writer's individual style (Hadley, 2020).

Due to these facts, literary texts are still perceived to be "the greatest challenge for MT" (Toral and Way, 2018). Some more pessimistic authors even claim that "there is no prospect of machines being useful at (assisting with) the translation of [literary texts]" (Toral and Way, 2018). While the use of machine translation followed by the post-editing phase is a widespread practice generally speaking, it has not yet become a permanent fixture in literary translation (Besacier, 2014).

In spite of this fact, there has been a growing interest in applying MT to literature, which can be seen, for example, in the fact that there is a workshop on computational linguistics for literature organised by ACL since 2012[1]. Moreover, the French-speaking world has seen the creation of an observatory for MT (*Observatoire de la traduction automatique*) by the ATLAS[2] association in December 2018 to follow the development of MT application to literary text[3].

Even though studies that analyse the application of MT to literary text are less numerous than those applying MT to other types of text, they are not inexistant. Hansen's (2021) paper brings a detailed and up-to-date overview of the works dealing with MT of literary texts. The first literary text translated by MT was done by Besacier (2014), and it comprised an essay translated from English to French. A number of languages have already been covered by various studies of MT to literary text, among which Slavic (e.g. Slovene, Kuzman et al., 2019), Romance (e.g. Catalan, Toral and Way, 2018, French, Besacier, 2014), Hansen, 2021), Germanic (English, in a number of papers; German, Matusov, 2019); Scottish Gaelic and Irish, Ó Murchú, 2019), etc.

2. Goal of the paper

The goal of this case study is to go beyond the overall performance of NMT on literary text and to provide an extensive, in-depth human analysis of its results. In order to do so, we will, firstly, produce a MT of a French novel and, secondly, compare that translation with a human translation of the same text. The human translation will be done by a student in translation from French into Croatian as part of her Master's thesis, and the analysis will be carried out by two human evaluators, the student and an experienced professional translator.

In addition to providing an in-depth analysis of the translation of a literary text done by MT, our case study is the first one to pair, to the best of our knowledge, a large Romance language, French, with Croatian[4], a smaller scale language rich in morphology.

The rest of the paper is structured as follows: in Section 3 we describe the methodology used. Section 4 is the central part of the paper, as it sums up the results of our analysis combined with a number of specific

[1] Cf. e.g. https://aclanthology.org/events/clfl-2020/.
[2] ATLAS stands for *Association pour la promotion de la traduction littéraire* (Association for the promotion of literary translation), https://www.atlas-citl.org/.
[3] https://www.atlas-citl.org/lobservatoire-de-la-traduction-automatique/
[4] Croatian is the official language of the Republic of Croatia and of the EU., but is also spoken in Bosnia and Herzegovina, Montenegro, etc. It has approximately 5.6 million native speakers worldwide. Cf. https://www.european-language-grid.eu/ncc/ncc-croatia/.

examples from the corpus. In Section 5 we bring some concluding remarks and recommend some further steps.

3. Methodology

In order to conduct our analysis, we have chosen a novel, which is "arguably the most popular type of literary text" (Toral and Way, 2018). Our corpus comprises the first eight chapters of the novel *La traduction est une histoire d'amour* (*Translation is a Love Affair*) written by Jacques Poulin, a contemporary Canadian author. It comprises a total of 8,347 words. The original text, written in French, is first translated by GT, and subsequently by a human translator. The MT is analysed in detail by two evaluators, after which the two translations are compared.

Hansen (2021) argues that evaluation of texts produced by MT still remains a major obstacle. More precisely, if BLEU (Papineni et al., 2002) is the most widely used automatic metric, it has to be taken with caution in case of literary texts (*ibid*.). Papineni et al. (2002) argue that human evaluations of MT are "extensive" and therefore usually more fine-grained than automatic ones, but the authors also point to their expensiveness.

In our case study, we present a quantitative and qualitative analysis of errors. We base our methodology on the one developed by Pavlović (2016). Pavlović (*ibid.*) also argues that in the literature there is not a single classification of translation errors that all authors would agree upon, so she makes her own classification based upon extant ones by a number of previous authors and some specificities of the corpus. Her study (2016) included only non-literary texts, newspaper reports, public opinion reports and EU legal documents (opinions and decisions), a total of 3,406 words. Still, Pavlović's (2016) methodology was developed with the goal of comparing MT done by GT and human translation, and it takes into account some specificities of the Croatian language such as a rather free word order, abundance of inflection and morphological complexity. It should be emphasized that Pavlović's (2016) study was conducted before GT used NMT for Croatian, which is available today [5] and is the technology used for the analysis presented in this case study.

The analysis of errors conducted for this paper follows that given by Pavlović (2016), with only minor alterations. For example, the sub-category (D.c), 'numbers', is not present in the machine translation of the chosen text and is hence not part of this analysis.

4. Results and analysis

4.1. Fine-grained human evaluation

Our analysis has demonstrated that GT has provided a very satisfactory translation generally speaking, and some of its solutions were even better than the ones provided by the human translation in the cases where there was a possible choice between a general word and its more suitable or literary synonym.

Below we first bring a table with a general presentation of errors found in the MT.

Error category	%
Morphosyntax	55.3
Lexicon	32.1
Spelling	7
Other	5.6

Table 1: Classification of general error types produced by MT.

Table 1 demonstrates that morphosyntactic errors visibly make the most frequent error type in our corpus, i.e. more than half of the total number of errors. These are followed by errors in lexical choice. In Table 2 (below) we bring a detailed list of error types found in our corpus.

Error type	%
C.a. congruence	39.3
B.a. lexical choice	18.8
C.c. word order / order of phrase constituents	10.9
B.c. idiomatic expressions	7.5
B.b. term or title	5.8
C.b. verbal forms / tenses	5.2
A.a. punctuation	4.5
A.b. capital letters	2
D.a. not translated	2
D.b. omissions	1.9
D.d. format, etc.	1.6

[5] Cf. https://translate.google.com/intl/hr/about/languages/.

A.c. other spelling errors	0,5
D.c. numbers	0

Table 2: Detailed breakdown of error types found in the corpus.

4.1.1. Morphosyntactic errors

According to our analysis, the most common errors done by GT are morphosyntactic errors, more specifically congruence errors, representing 39.3%. This type of errors most frequently have to do with grammatical gender. Here is an example:

original	GT	human translation
*La meilleure **traductrice** du Québec*	*Najbolji **prevodilac** u Quebecu*	*Najbolja **prevoditeljica** u Québecu.*

Table 3: Example of congruence error.

In the above example, *traductrice* 'female translator' is translated by GT as *prevoditelj* 'male translator' even though both French and Croatian are marked for gender, and even though there is a ready-made solution in Croatian, *prevoditeljica* 'female translator'. The problem here is probably the fact that GT uses English as a sort of pivot or intermediate language (e.g. Ljubas, 2018) when translating between French and Croatian[6], that do not share as large a corpus of texts as they do with English individually.

This is a frequent error produced by GT in the corpus, i.e. not marking whatever has to do with the narrator, who is a woman, as female, but leaving male nouns, adjectives etc., which we also attribute to translating via English: e.g. *Je raccrochai* is translated as *Spustio* (masc.) *sam slušalicu* instead of *Spustila* (fem.) *sam slušalicu / Poklopila* (fem.) *sam*.

In other words, it can be said generally that our analysis has demonstrated that GT had no problems, for example, with the Croatian rich nominal case system and general subject-verb or noun-adjective agreement. This is in line with findings from the literature that neural systems have been found to make fewer morphological, lexical and word-order errors (e.g. Burchardt, 2017). What was a problem, however, in the category of morphosyntactic errors is recognizing the narrator as a female, and consequently translating all her attributes and making all the agreements in the feminine gender. This is a feature of the text that extends beyond sentence level and permeates the entire discourse of the novel. In some French sentences, this difference between masculine and feminine gender cannot be seen, for example in the present tense or in the past tense (*passé compose*) formed with the auxiliary verb to have (*avoir*). In Croatian, the same goes for the present tense, but the past tense always shows agreement with the subject in gender. The large number of errors in this category undoubtedly stems from the use of English as a pivot language.

4.1.2. Lexical errors

The next most represented category are lexical errors (32.1%), listed in the table below.

original	GT	human translation
*Eh bien, c'était le **portrait tout craché** de ma mère.*	*Pa, to je bila **pljuvačka slika** moje majke.*	*E pa to je **pljunuti portret** moje majke.*
*Les **ouaouarons**, affolés, …*	*Uplašeni **bikovi žabe** …*	***Žabe** su se prenerađene …*
*Je suis sur la route parce que ma **maîtresse** ne peut plus s'occuper de moi, (…)*	*Na putu sam jer se moja **ljubavnica** više ne može brinuti o meni*	*Na ulici sam jer se moja **vlasnica** više ne može brinuti o meni, (…)*
*Ma mère et ma grand-mère **reposaient** derrière l'église …*	*Moja majka i baka **odmarale su se** iza crkve …*	*Moja majka i baka **bile su pokopane** iza crkve …*
*… dans l'herbe jonchée de **feuilles mortes**.*	*… u travi posutoj **mrtvim lišćem**.*	*… travi prekrivenoj **suhim lišćem**.*
*J'étais très heureuse, **presque sur un nuage**, (…)*	*Bio sam vrlo sretan, skoro **na devetom oblaku**, (…)*	*Bila sam sretna, **gotovo u sedmom nebu**, (…)*
*Les maudites **algues**…*	*Proklete morske **alge**…*	*Proklete **alge**…*

Table 3: Examples of lexical choice errors.

[6] This has been claimed generally as a feature of GT that it uses when translating between any pair of languages. A Google spokesperson has admitted that Google Translate uses English for „bridging" between languages with fewer resources. See https://algorithmwatch.org/en/google-translate-gender-bias/; cf. https://www.circuitmagazine.org/chroniques-126/sur-le-vif-126/google-uses-english-as-a-pivot-language.

Errors is this category concern the following: 1) single-word polysemy, 2) idiomatic expressions, 3) calques from English.

With respect to single-word polysemy, GT has, for instance, erroneously translated *maîtresse* 'owner' (of a cat) as 'lover'. It also translated *reposaient* 'rested' as *odmarale su se* 'were having a rest' instead of *bile su pokopane*, which is used in the context of the dead buried in a graveyard. Furthermore, it translated *algues* as *morske alge* 'sea algae', which is an incorrect specification stemming from the fact that algae are usually related to the sea, but algae in the story, however, come from a pond.

As for idiomatic expressions (7% of total errors), GT rendered *le portrait craché* 'spitting image' as **pljuvačka slika* instead of *pljunuti portret*. It clearly calqued the expression *être sur un nuage* 'be on cloud nine' on English and translate dit as **biti na devetom oblaku*, which does not exist in Croatian, and should be translated as *na sedmom nebu* 'lit. on seventh sky'. The noun phrase *feuilles mortes* is litterally translated as **mrtvo lišće* instead of *suho lišće* 'lit. dry leaves', etc.

There are several instances of calquing from English, such as in the example of *ouaouarons*, animals known in English as American bullfrogs, which are litterally translated as *bikovi žabe* 'bulls-frogs', and for which we would suggest the translation *žabe* due to the fact that the particular species is irrelevant to the plot.

4.1.3. Other errors

In the category of capital letters, GT had difficulties rendering street names, which appeared in the text several times. Examples such as *609, rue Richelieu* were rendered by GT as *609, ulica Richelieu*, where all the individual elements are correctly translated, but the street name as a whole should be written as *Ulica Richelieu 609*, which is a conventional way of writing street names in Croatian.

Another interesting error concerns proper names. Let us cite two examples: *Marine* and *Chaloupe*. Marine, the name of the main character and narrator, is sometimes translated by GT as *marinac* 'Marine, i.e. member of an elite US fighting corps'. In addition to the same form, the English word is always capitalised, so that could be another reason for such a translation. *Chaloupe*, on the other hand, is the name of the cat that appears several times in the text. It is derived from the common noun *chaloupe* denoting a type of boat. GT translated the noun as *čamac* 'boat', making it a common noun and even leaving out the capital letter.

Bentivogli et al. (2016) and Toral and Sánchez Cartagena (2017) found that NMT improves notably on reordering and inflection than PBMT. In the case of Poulin's novel translated and analysed in this paper, there were generally very few problems with inflection, and word / constituent order represented only 10% of all the errors. What our analysis seems to point to is the fact that using English as a pivot language is the source of a large number of errors, and that using language-pair specific language corpora could arguably give better results in translating between two languages of which neither is English. This would also probably have a positive effect on the translation of culturally specific elements such as spelling and writing of toponyms (e.g. street names). Furthermore, our analysis also demonstrates that more improvement should be done in the detection and translation of polysemy and idiomatic expressions.

4.2. BLEU evaluation

In addition to a fine-grained human translation, BLEU score was also calculated using the interactive BLEU score evaluator[7] available via the Tilde platform. BLEU score is based on the correspondence of the MT output and the reference human translation.

Overall cumulative BLEU score for the literary text analysed in our case study was 5.49, which would suggest very poor MT quality. As a reference, BLEU scores of 30 to 40 are considered to be "understandable to good translations", while those of 40 to 50 are "high quality translations"[8]. Here is the breakdown of the BLEU score:

Type	1-gram	2-gram	3-gram	4-gram
Individual	21.92	5.86	2.79	2.54
Cumulative	21.92	11.33	7.10	5.49

Table 4: Results of automatic BLEU evaluation.

In other available case studies dealing with MT of a literary text, BLEU scores show significant variation. In the case of a translation of a literary essay from English into French (Besacier and Schwartz, 2015), BLEU score was around 30. In another case study dealing with English literary texts translated into Slovene, BLEU scores varied from 1.73 to 30 depending on the texts on which the MT model was trained (Kuzman et al., 2019). Toral and Way (2018) obtained BLEU scores of around 30 for English-to-Catalan translations of 12 English

[7] https://www.letsmt.eu/Bleu.aspx.
[8] https://cloud.google.com/translate/automl/docs/evaluate

novels by PBSMT and NMT systems, where NMT outperformed PBSMT.

Unlike the results obtained by Kuzman et al. (2019) in their study of a literary translation from English into Slovene, a language genetically very close to Croatian, where "there were no sentences that would not need postediting", in our case study there were a number of sentences entirely correctly rendered by GT, i.e. that would be publication ready.

In any case, it should be borne in mind that BLEU automatic evaluation metric was calculated with respect to a single human translation, and that it cannot represent the "real quality" of MT output. In that sense, Hansen (2022) notes, for instance, that two MT models used in his case study had a similar BLEU score in spite of the fact that the first one produced correctly translated words in incomprehensible sentences, while the second one generated correct sentences with words that semantically did not correspond the lexical field of the translated literary text. This is one of the reasons why we would not entirely agree that the translation provided by GT analysed in this paper is irrelevant or "useless", as it would be classified due to its BLEU score inferior to 10 (cf. footnote n° 8).

In addition, it should be noted that some authors claim that morphological richness of the Croatian language could raise problems for BLEU evaluation due to the fact that each Croatian noun has approximately 10 different word forms, which are considered by BLEU to be 10 different words, and not 10 different word forms of a single lemma (cf. Seljan et al., 2012). This could result in lower BLEU scores.

5. Conclusion

This case study is a contribution to a growing number of papers dealing with applying (N)MT to literary text, which has been thought of until only recently as a domain that could not be translated by MT. Various authors have, however, demonstrated the usefulness of using MT in literary translation. Some (e.g. Besacier and Schwartz, 2015) even argue that MT of literary text may even be of interest for all participants of the translation chain from editors, through readers to authors and translators.

Our analysis has demonstrated that there was a total of 738 errors in the text produced by GT, largely falling into two groups: morphosyntactic (around 55%) and lexical choice (around 32%) errors. While the morphosyntactic errors largely concerned errors in congruence stemming probably from the usage of English as a pivot language between French and Croatian, the lexical choice errors had mostly to do with polysemy, idiomatic expressions and calques.

Let us now compare our results with those from other existing works on MT of literary texts involving either of the two languages from this case study, Croatian or French. Hansen (2022), who analysed English-to-French translations of fantasy books, observed that, generally speaking, the MT output was rather literal and it produced mostly lexical errors, as well as errors related to determiners and syntax. While Hansen (*ibid.*) does not provide further details, we can generally say that in our French-to-Croatian literary translation morphosyntactic errors were by 20% more present than lexical errors, which is different than what he found in the English-French language pair. Furthermore, Hansen (*ibid*.) was surprised to note that the specific vocabulary related to the fantasy series in question was respected almost entirely, which is probably due to the training of the MT model on texts written by the same author. This is one of the reasons why Hansen (2022) suggests that personalized MT systems should be introduced in literary translation for translating specific authors' styles.

In another paper, involving Slovene, a language closely related to Croatian, and analysing translation of literary texts from English, Kuzman et al. (2019) observe that "error analysis (…) revealed various punctuation errors, wrong translations of prepositions and conjunctions, inappropriate shifts in verb mood, wrong noun forms and co-reference changes". The authors emphasize the presence of numerous semantic errors, "especially in connection with idioms and ambiguous words". In this case, more detailed data is also lacking, but we can generally conclude that this study also differs from ours in that semantic errors are definitely not the leading error type in our French-to-Croatian translation. Interestingly, Kuzman et al. (2019) also found that GNMT assigned the wrong gender to the main character, just as happened in our case, as mentioned in 4.1.1.

We can conclude that in the French-to-Croatian GT of the novel analysed in this text, morphosyntactic errors (55.3%) are the most represented ones, followed by various lexical errors (32.1%). These results are somewhat different from what was observed in earlier extant studies dealing with MT of literary texts from English to French and English to Slovene.

Even though BLEU score was only 5.49, indicating very poor translation quality which should be deemed as useless, we believe that the GT output would be useful to some extent to translators translating Poulin's novel from scratch. Further analyses should be made however in order to analyse whether GT trained on French and Croatian corpora would amount to better results than GT that uses English as pivot. Furthermore, it should also be studied how much post-processing effort is needed to correct errors of GT in comparison to translation from scratch in the French-to-Croatian language combination.

6. References

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In: J. Siu, K. Duh and X. Carreras, eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Association for Computational Linguistics, Austin, Texas.

Laurent Besacier and Lane Schwarz. 2015. Automated Translation of a Literary Work: A Pilot Study. In: A. Feldman, A. Kazantseva, S. Szpakowicz and C. Koolen, eds., *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 114–122. Association for Computational Linguistics, Denver, Colorado.

Laurent Besacier. 2014. Traduction automatisée d'une œuvre littéraire : une étude pilote. In: P. Blanche, F. Béchet and B. Bigi, eds., *Actes du 21ème Traitement Automatique des Langues Naturelles*, pages 389-94. Association pour le Traitement Automatique des Langues, Marseille. https://hal.inria.fr/hal-01003944

Marija Brkić, Sanja Seljan and Maja Matetić. 2011. Machine Translation Evaluation for Croatian-English and English-Croatian Language Pairs. In: B. Sharp, M. Zock, M. Carl, A. L. Jakobsen, eds., *Proceedings of the 8th International NLPCS Workshop: Human-Machine Interaction in Translation,* pages 93-104. Copenhagen: Copenhagen Business School.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A Linguistic Evaluation of Rule- Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, 108:159–170.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020.

Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level. In: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Delreck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis, eds., *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 3790–3798, European Language Resources Association, Marseille.

James Hadley. 2020. Traduction automatique en littérature : l'ordinateur va-t-il nous voler notre travail. *Contrepoint*, 4:14–18. https://www.ceatl.eu/wp-content/uploads/2020/12/Contrepoint_2020_04_article_04.pdf

Damien Hansen. 2022. La traduction littéraire automatique : Adapter la machine à la traduction humaine individualisée. https://hal.archives-ouvertes.fr/hal-03583562/document

Damien Hansen. 2021. Les lettres et la machine : un état de l'art en traduction littéraire automatique. In: P. Denis, N. Grabar, A. Fraisse, R. Cardon, B. Jacquemin, E. Kergosien and A. Balvet, eds., *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*, Vol. 1, pages 61–78. ATALA, Lille.

Filip Klubička, Antonio Toral and Víctor M. Sánchez- Cartagena. 2017. Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 108:121-132. https://arxiv.org/abs/1706.04389

Taja Kuzman, Špela Vintar and Mihael Arčan. 2019. Neural Machine Translation of Literary Texts from English to Slovene. In: J. Hadley, M. Popović, H. Afli, and A. Way, eds., *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9, European Association for Machine Translation, Dublin. https://aclanthology.org/W19-7301

Rudy Loock. 2018. Traduction automatique et usage linguistique : une analyse de traductions anglais-français réunies en corpus. *Meta* 63(3):786–806. https://doi.org/10.7202/1060173ar

Sandra Ljubas. 2018. Prijelaz sa statističkog na neuronski model: usporedba strojnih prijevoda sa švedskoga na hrvatski jezik. *Hieronymus*, 5:72–79. https://www.bib.irb.hr/978980

Evgeny Matusov. 2019. The Challenges of Using Neural Machine Translation for Literature. In: J. Hadley, M. Popović, H. Afli, and A. Way, eds., *Proceedings of the Qualities of Literary Machine Translation, Machine Translation*, pages 10-19, European Association for Machine Translation, Dublin. https://aclanthology.org/W19-7302.pdf

Eoin P. Ó Murchú. 2019. Using Intergaelic to pre-translate and subsequently post-edit a sci-fi novel from Scottish Gaelic to Irish. In: J. Hadley, M. Popović, H. Afli & A. Way, eds., *Proceedings of the Qualities of Literary Machine Translation*, pages 20–25, European Association for Machine Translation, Dublin. https://aclanthology.org/W19-7303

Kishore Papineni, Salim Roukos, Todd Ward and Wei- Jing Zhu. 2016. Bleu: a Method for Automatic Evaluation of Machine Translation. In: P. Isabelle, E. Charniak and D. Lin, eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for computational linguistics, Philadelphia, Pennsylvania, USA. doi.org/10.3115/1073083.1073135

Nataša Pavlović. 2017. Strojno i konvencionalno prevođenje s engleskoga na hrvatski: usporedba pogrešaka. In: D. Stolac and A. Vlastelić, eds., *Jezik kao predmet proučavanja i jezik kao predmet poučavanja*, pages 279–295, Srednja Europa, Zagreb.

Jacques Poulin. 2006. *La traduction est une histoire d'amour*. Leméac/Actes Sud, Montreal.

Sanja Seljan, Marija Brkić and Tomislav Vičić. 2012. BLEU Evaluation of Machine-Translated English-Croatian Legislation. In: N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, eds., *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey.

Antonio Toral and Andy Way. 2018. What Level of Quality Can Neural Machine Translation Attain on Literary Text? In: J. Moorkens, S. Castilho, F. Gaspari, S. Doherty, eds., *Translation Quality Assessment. Machine Translation: Technologies and Applications*, Vol 1, pages 263-287. Springer, Cham. https://doi.org/10.1007/978-3-319-91241-7

A Bilingual English-Ukrainian Lexicon of Named Entities Extracted from Wikipedia

Aleksandar Petrovski

Faculty of Informatics
International Slavic University
Marshal Tito 77 Sv. Nikole, North Macedonia
aleksandar.petrovski@msu.edu.mk

Abstract

This paper describes the creation of a bilingual English - Ukrainian lexicon of named entities, with Wikipedia as a source. The proposed methodology provides a cheap opportunity to build multilingual lexicons, without having expertise in target languages. The extracted named entity pairs have been classified into five classes: PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC (miscellaneous). It has been achieved using Wikipedia metadata. Using the presented methodology, a huge lexicon has been created, consisting of 624,168 pairs. The classification quality has been checked manually on 1,000 randomly selected named entities. The results obtained are 97% for precision and 90% for recall.

1. Introduction

The term named entity (NE) refers to expressions describing real world objects, like persons, locations, and organizations. It was first introduced to the Natural Language Processing (NLP) community at the end of the 20th century. Named entities are often denoted by proper names. They can be abstract or have a physical existence. Some other expressions, describing money, percentage, time, and date might also be considered as named entities. Examples of named entities include: *United States of America*, *Paris*, *Google*, *Mercedes Benz*, *Microsoft Windows*, or anything else that can be named.

The role of named entities has become more and more important in NLP. Their information is crucial in information extraction. As recent systems mostly rely on machine learning techniques, their performance is based on the size and quality of given training data. This data is expensive and cumbersome to create because experts usually annotate corpora manually to achieve high quality data. As a result, these data sets often lack coverage, are not up to date, and are not available in many languages. To overcome this problem, semi-automatic methods for resource construction from other available sources were deployed. One of these sources is Wikipedia.

The method presented here has been used to build a Python application which extracts the English - Ukrainian pairs from Wikipedia and classifies them using the English Wikipedia category system. Since both English and Ukrainian are among languages with most articles on Wikipedia, the result is a huge lexicon.

The goal of this paper is to present a method of extracting multilingual lexicons of classified named entities from Wikipedia. The method has been implemented to build a huge English - Ukrainian lexicon of named entities.

2. Related work

Building multilingual lexicons from Wikipedia has been a subject of research for more than 10 years. Schönhofen et al. (Schönhofen et al., 2007) exploited Wikipedia hyperlinkage for query term disambiguation. Tyers and Pienaar (Tyers and Pienaar, 2008) described a simple, fast, and computationally inexpensive method for extracting bilingual dictionary entries from Wikipedia (using the interwiki link system) and assessed the performance of this method with respect to four language pairs. Yu and Tsujii (Yu and Tsujii, 2009) proposed a method using the interlanguage link in Wikipedia to build an English-Chinese lexicon. Knopp (Knopp, 2010) showed how to use the Wikipedia category system to classify named entities. Bøhn and Nørvåg (Bøhn and Nørvag, 2010) described how to use Wikipedia contents to automatically generate a lexicon of named entities and synonyms that are all referring to the same entity. Halek et al. (Hálek et al., 2011) attempted to improve machine translation from English of named entities by using Wikipedia. In (Ivanova, 2012), the author evaluated a bilingual bidirectional English-Russian dictionary created from Wikipedia article titles. Higashinaka et al. (Higashinaka et al., 2012) aimed to create a lexicon of 200 extended named entity (ENE) types, which could enable fine-grained information extraction. Oussalah and Mohamed (Oussalah and Mohamed, 2014) demonstrated how to use info-boxes in order to identify and extract named entities from Wikipedia.

3. Wikipedia

Wikipedia is a free online encyclopedia, made and maintained as an open coordinated effort venture by a network of volunteer editors, utilizing a wiki – based editing system. Hosted and supported by the Wikimedia Foundation, since its start in 2001, the site has grown in both popularity and size. At the time of writing this paper (March 2022), Wikipedia contained over 58 million articles in 323 languages; its English version has over 6 million articles. The richness of information and texts continuously makes it an object of special research interest among the NLP (Natural Language Processing) community. By attracting approximately 6 billion visitors per month (Statista, 2021), it is the largest and most popular general reference work on the World Wide Web.

3.1. Wikipedia as a source

Even though Wikipedia isn't made and maintained by linguists, metadata about articles, for instance, translations, disambiguations, or categorizations are accessible. Its structural features, size, and multilingual availability give a reasonable base to derive specialized resources, like multilingual lexicons (Bøhn and Nørvag, 2010). Researchers have found that around 74% of Wikipedia pages describe named entities (Nothman et al., 2008), a clear indication of Wikipedia's high coverage for named entities. Each Wikipedia article associated with a named entity is identified with its title, which is itself a named entity. That is a perfect opportunity to build parallel lexicons of named entities between them.

Wikipedia is a very cheap resource of multilingual lexicons of named entities. Its database dump can be freely downloaded in sql and XML formats. But, taking into account the fact that Wikipedia articles have been written by millions of contributors, a question arises: What is the quality of these lexicons, and how reliable are they for using, e.g., in machine translation?

3.2. English and Ukrainian Wikipedias

The English Wikipedia is the English language edition of the Wikipedia online encyclopedia. English is the first language in which Wikipedia was written. It was started on 15 January 2001 (Wikimedia Foundation, 2022b), but versions of Wikipedia in other languages were quickly developed. Among these versions, there is one in Ukrainian language. The Ukrainian Wikipedia (Wikimedia Foundation, 2022c), written in the Cyrillic alphabet, was initiated in the year 2004.

A list of all Wikipedias is published regularly on the Internet, along with several parameters for each language (Wikimedia Foundation, 2022a). Four parameters are important: number of articles, the total number of pages (articles, user pages, images, talk pages, project pages, categories, and templates), number of active users (registered users who performed at least one change in the last thirty days), and depth (a rough indicator of the quality of Wikipedia, which shows how often articles are updated).

As shown in Table 1, as of 26 March 2022, the English Wikipedia contains 6,473,638 articles and 55,472,454 pages. There are 127,722 active users. The depth value is 1,110. It is by far the largest edition of Wikipedia. The Ukrainian Wikipedia contains 1,144,596 articles and 3,992,549 pages. There are 2,702 active users. The depth value is 54. It is the 17th largest edition of Wikipedia, according to number of articles.

Parameter	en	uk
Number of articles	6,473,638	1,144,596
Total number of pages	55,472,454	3,992,549
Number of active users	127,722	2,702
Depth	1,110	54

Table 1: Parameters of the English and Ukrainian Wikipedias.

4. Method

The flowchart presented in Figure 1 shows the process used for building the lexicon.

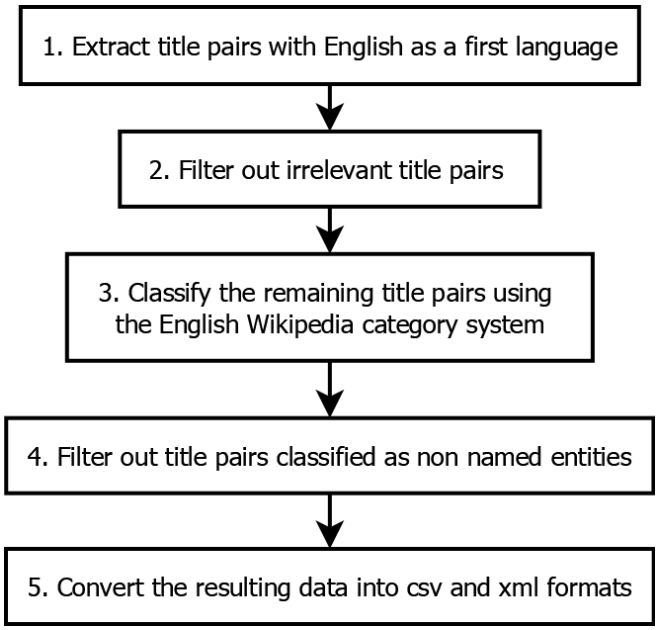


Figure 1: The process flowchart.

1. Extract title pairs with English as a first language

For building multilingual lexicons, two tables from the database are necessary: table of pages and table of interlanguage links. The page table is the "core of the wiki". It contains titles and other essential metadata for different Wikipedia namespaces. The interlanguage links table contains links between pages in different languages. Using these two tables, it is an easy programming task to create huge bilingual dictionaries without having any language expertise.

2. Filter out irrelevant title pairs

The extracted title pairs from the previous step contain a lot of noise. This step deals with it. First, the algorithm removes all the titles that don't belong to the main, template, or category namespaces. Second, there are titles containing some words or word stems that increase the noise and should be filtered out. The page table contains many entries that could not be a part of any lexicon, like user names, nicknames, template names, etc. There are also titles, containing exclusively digits or blanks, which should be removed too.

3. Classify the remaining title pairs using the English Wikipedia category system

In order to classify the extracted named entities, one additional table from the database is required: a table of category links. The task of classifying named entities by means of category links is more complex. Wikipedia articles are generally members of categories. A category may have subcategories, each subcategory its own subcategories, etc. The problem is that the graph could be cyclic, which may cause the algorithm to go into an endless loop.

Various authors propose different classes for named entities. Here, there are five: PERSON, ORGANIZATION,

en	uk	PERS	ORG	LOC	PROD	MISC
Odessa	Одеса	0	0	1	0	0

Figure 2: A lexicon entry in CSV format.

LOCATION, PRODUCT, and MISC. Each named entity belongs to at least one of these classes. The classes comprise:

- ORGANIZATION- political organizations, companies, schools, rock bands, sport teams
- PERSON- humans, gods, saints, fictional characters
- LOCATION- geographical terms, fictional places, cosmic terms
- PRODUCT- industrial products, software products, weapons, artworks, documents, concepts, standards, laws, formats, anthems, algorithms, journals, coats of arms, platforms, websites
- MISC- events, languages, peoples, tribes, alliances, orders, scientific discoveries, theories, titles, currencies, holidays, dynasties, positions, projects, historical periods, battles, competitions, alliances, deceases, programs, set of locations, awards, musical genres, missions, artistic directions, sets of organizations, networks

4. Filter out title pairs classified as non named entities

Most Wikipedia titles are named entities, but not all of them. For example, certain natural terms-like biological species and substances-which are very common on Wikipedia, are not included in the lexicon.

5. Convert the resulting data into CSV and XML formats

The lexicon comes in two formats: CSV and XML.

The first row in the CSV file is a title row and tab is used as a field separator. The columns' titles are: en, uk, PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC. All other rows contain the data: English name, Ukrainian name, and five binary digits. These digits denote the class the named entity belongs to. For example, according to Figure 2, the named entity *Odessa* belongs to the class LOCATION, since the column LOC contains 1. All other classes contain 0's.

The structure of the XML file is similar. An equivalent of the entry from Figure 2 is shown in Figure 3. The columns' names en and uk from the CSV file are now names of elements and *class* denotes the classification.

In realizing the steps 2-3 of Figure 1, which refer to noise reduction and classification of named entities, the experience of creating a parallel lexicon of named entities from English to South Slavic languages (Slovenian, Croatian, Croatian, Bosnian, Ukrainian, Macedonian, and Bulgarian) (Petrovski, 2019) was of great benefit. That lexicon contains 26,155 entries, and the steps 2-3 were done manually.

This methodology has been used to create a multilingual English – Hebrew – Yiddish – Ladino lexicon of named entities. A tool that can be used to search it, can be found on the Internet (Petrovski, 2021).

```
<entry>
        <en> Odessa</en>
        <uk>Одеса</uk>
        <classes>
                <class>LOCATION</class>
        </classes>
</entry>
```

Figure 3: A lexicon entry in XML format.

5. Results

The method presented in previous chapter has been used to build a Python application which extracts title pairs independently on the languages. This application was applied to the Wikipedia database to extract the English - Ukrainian pairs of named entities. The result of the extraction after the first two steps from Figure 1 was 687,799 pairs. After filtering out non named entities, 624,168 pairs remained.

One part of the lexicon is presented in Figure 4.

en	uk	PERSON	ORGANIZATION	LOCATION	PRODUCT	MISC
Kyiv	Київ	0	0	1	0	0
Kyiv Academic Puppet Theatre	Київський академічний театр ляльок	0	0	1	0	0
Kyiv Academic Theatre of Drama and Comedy on the left bank of Dnieper	Київський академічний театр драми і комедії на лівому березі Дніпра	0	0	1	0	0
Kyiv Academic Theatre of Ukrainian Folklore	Київський академічний театр українського фольклору «Берегиня»	0	0	1	0	0
Kyiv Academic Young Theatre	Київський національний академічний Молодий театр	0	0	1	0	0
Kyiv Ballet	Київський балет	0	1	0	0	0
Kyiv Boryspil Express	Kyiv Borys pil Express	0	1	1	0	0
Kyiv Camerata	Національний ансамбль солістів «Київська камерата»	0	1	0	0	0
Kyiv Central Bus Station	Київський центральний автовокзал	0	0	1	0	0
Kyiv Chaika Airfield	Чайка	0	0	1	0	0
Kyiv Chamber Choir	Київ	0	1	0	0	0
Kyiv Chamber Orchestra	Київський камерний оркестр	0	1	0	0	0
Kyiv Christian Academy	Київська християнська академія	0	1	0	0	0
Kyiv City Council	Київська міська рада	0	1	1	0	0
Kyiv City Duma building	Будинок Київської думи	0	0	1	0	0
Kyiv City State Administration	Київська міська державна адміністрація	0	0	1	0	0
Kyiv Conservatory	Національна музична академія України імені Петра Чайковського	0	1	1	0	0
Kyiv Conservatory alumni	Випускники Київської консерваторії	0	1	0	0	0
Kyiv Conservatory faculty	Викладачі Київської консерваторії	0	0	0	0	1
Kyiv Day	День Києва	0	0	1	0	0
Kyiv Day and Night	Київ вдень та вночі	0	0	0	1	0
Kyiv Fortress	Київська фортеця	0	0	1	0	0
Kyiv Funicular	Київський фунікулер	0	0	1	0	0
Kyiv Half Marathon	Київський півмарафон	0	1	0	0	0
Kyiv Higher Party School alumni	Випускники Вищої партійної школи при ЦК КПУ	0	1	0	0	0
Kyiv Hydroelectric Power Plant	Київська ГЕС	0	0	1	0	0
Kyiv Independence Day Parade	Парад на честь дня Незалежності України	0	0	0	0	1
Kyiv Institute of Business and Technology	Київський інститут бізнесу та технологій	0	1	0	0	0
Kyiv International Airport	Міжнародний аеропорт «Київ»	0	0	1	0	0
Kyiv International Film Festival "Molodist"	Молодість	0	1	0	0	0
Kyiv International Institute of Sociology	Київський міжнародний інститут соціології	0	1	0	0	0
Kyiv International School	Київська міжнародна школа	0	1	0	0	0

Figure 4: A part of the lexicon.

The distribution of classes is presented in Table 2.

Class	Number
PERSON	142,850
ORGANIZATION	39,348
LOCATION	237,229
PRODUCT	56,952
MISC	159,952
Total	636,331

Table 2: Distribution of classes.

The total number of classes, 636,331, is slightly higher than the number of entries, since some named entities may belong to more classes. The lexical entry presented in Figure 5 is such an example. *Kherson State University* is classified as both ORGANIZATION (the university as an educational organization) and LOCATION (the building where the organization is located).

```
<entry>
        <en>Kherson State University</en>
        <uk>Херсонський державний університет</uk>
        <classes>
                <class>ORGANIZATION</class>
                <class>LOCATION</class>
        </classes>
</entry>
```

Figure 5: A lexicon entry belonging to two classes.

It is expected that the most of Wikipedia titles are multiwords, i.e. they contain either a space or a hyphen. Table 3 shows the number of multiword NEs per class in the lexicon for both English and Ukrainian.

Class	en	uk
PERSON	132,219	131,354
ORGANIZATION	34,114	30,509
LOCATION	116,974	99,399
PRODUCT	45,781	43,378
MISC	146,498	141,665
Total	475,586	446,305

Table 3: Number of multiword NEs per class.

Table 4 shows the percentage of multiword NEs per class.

It can be seen that the percentage of multiwords is higher in the English than in the Ukrainian Wikipedia. This is most noticeable in the classes ORGANIZATION and LOCATION. Some examples from the lexicon where there is a multiword in English and a single word in Ukrainian are given in Table 5 for the class ORGANIZATION and Table 6 for the class LOCATION.

Contributors to the English Wikipedia add words to the base title, which define it in more detail, or it is simply a matter of adding a definite article, e.g. *Sacramento, California* - Сакраменто, *Malkiya Club* - Малкія, *The Acropolis* - Акрополіс.

Class	en	uk
PERSON	93%	92%
ORGANIZATION	87%	78%
LOCATION	49%	42%
PRODUCT	80%	76%
MISC	92%	89%
All	75%	70%

Table 4: Percentage of multiword NEs per class.

en	uk
Malkiya Club	Малкія
Dnipro Kherson	Дніпро
Sharjah FC	Шарджа
Shin Bet	Шабак
Newtown A.F.C.	Ньютаун
The Day After Tomorrow	Післязавтра

Table 5: Examples of multiwords in English and single words in Ukrainian, class ORGANIZATION.

6. Evaluation of classification

To evaluate classification, two common metrics in information retrieval have been used: precision and recall. Precision refers to the percentage of classes that are correct. On the other hand, recall refers to the percentage of total relevant classes correctly classified by the algorithm.

An alternative to having two measures is the F-measure, which combines precision and recall into a single performance measure. This metric is known as F1-score, which is simply the harmonic mean of precision and recall.

In order to evaluate the classification, a random sample containing 1,000 entries has been extracted from the lexicon. The entries from the sample have been classified manually and then compared to the classification performed by the algorithm. The results are presented in Table 7.

The precision of classification is between 94% for ORGANIZATION and 99% for PERSON. The recall is slightly lower, from 83% for PRODUCT and MISC to 97% for PERSON. The overall results are 97% for precision and 90% for recall.

The higher values of precision show that the classification algorithm was adjusted to classify the named entities correctly, rather than to extract more named entities for the lexicon.

7. Conclusion

Using the methodology presented in this paper, an English - Ukrainian lexicon of named entities has been created. Its size is 624,168 pairs. The named entities have been classified into five classes: PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC (miscellaneous). The quality of classification has been assessed: 97% for precision and 90% for recall.

en	uk
Malmö Airport	Мальме
Shintoku, Hokkaido	Сінтоку
Amarillo, Texas	Амарилло
Sacramento, California	Сакраменто
The Dakota	Дакота
The Acropolis	Акрополіс

Table 6: Examples of multiwords in English and single words in Ukrainian, class LOCATION.

Class	Precision	Recall	F1-score
ORGANIZATION	94%	87%	90%
LOCATION	98%	92%	95%
PRODUCT	96%	83%	89%
MISC	96%	83%	89%
All	97%	90%	93%

Table 7: The results of the classification check.

The lexicon is available on (Petrovski, 2022) under CC-BY-NC-4.0 license (free for non commercial use).

Lexicons, like the one presented in this paper, can be used in machine translation (MT). Most statistical MT systems do not deal explicitly with named entities, simply relying on the model of selecting the correct translation, i.e., mistranslating them as generic nouns. It is also possible that, when not identified, named entities may be left out of the output translation, which also has implications for the readability of the text. Because most NEs are rare in texts, statistical MT systems are not capable of producing quality translations for them. Another problem with MT systems is that failure to recognize NEs often harms the morpho – syntactic and lexical context outside of NEs itself. If named entities are not immediately identified, certain morphological features of adjacent and syntactically related words, as well as word order, may be incorrect. It can be concluded that the identification of named entities in the source text is the first task of machine translators (Hálek et al., 2011). However, developers of commercial MT systems often do not pay enough attention to the correct automatic identification of certain types of NE, e.g. names of organizations. This is partly due to the greater complexity of this problem (the set of proper nouns is open and very dynamic), and partly due to lack of time and other development resources. One solution to this problem is using a parallel lexicon of named entities. If the lexicon contains a translation of the named entity, the translation quality will probably be good.

The European Commission called for language data in Ukrainian to/from all EU languages to train automatic translation systems (European Commission, 2022), (European Union's Horizon 2020 Research and Innovation Programme, 2020) supporting refugees and helpers in the Ukraine crisis. This lexicon was sent to ELRC (European Language Resource Coordination) Secretariat as a response.

8. References

Christian Bøhn and Kjetil Nørvag. 2010. Extracting Named Entities and Synonyms from Wikipedia. In *Proceedings of International Conference on Advanced Information Networking and Applications*, pages 1300–1307.

European Commission. 2022. Digital Europe Programme Language Technologies. https://language-tools.ec.europa.eu/.

European Union's Horizon 2020 Research and Innovation Programme. 2020. Bergamot Translations. https://translatelocally.com/web/.

Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, Toshiro Makino, and Yoshihiro Matsuo. 2012. Creating an Extended Named Entity Dictionary from Wikipedia. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, pages 1163–1178.

Ondrej Hálek, Rudolf Rosa, Aleš Tamchyna, and Ondrej Bojar. 2011. Named entities from wikipedia for machine translation. In *Proceedings of the Conference on Theory and Practice of Information Technologies*, pages 23–30.

Angelina Ivanova. 2012. Evaluation of a Bilingual Dictionary Extracted from Wikipedia. In *Computer Science*.

Johannes Knopp. 2010. *Classification of Named Entities in a Large Multilingual Resource Using the Wikipedia Category System*. University of Heidelberg, Master's thesis, Heidelberg, Germany.

Joel Nothman, James Curran, and Tara Murphy. 2008. Transforming Wikipedia into Named Entity Training Data. In *Proceedings of the Australian Language Technology Workshop*.

Mourad Oussalah and Muhidin Mohamed. 2014. Identifying and Extracting Named Entities from Wikipedia Database Using Entity Infoboxes. In *International Journal of Advanced Computer Science and Applications*, volume 5, pages 164–169.

Aleksandar Petrovski. 2019. EnToSSLNE - a Lexicon of Parallel Named Entities from English to South Slavic Languages. http://catalogue.elra.info/en-us/repository/browse/ELRA-M0051/.

Aleksandar Petrovski. 2021. Jewish Lexicons of Named Entities. https://www.jewishlex.org/.

Aleksandar Petrovski. 2022. A Bilingual English-Ukrainian Lexicon of Named Entities Extracted from Wikipedia. https://catalogue.elra.info/en-us/repository/browse/ELRA-M0104/.

Péter Schönhofen, András Benczúr, Istvan Biro, and Károly Csalogány. 2007. Cross-Language Retrieval with Wikipedia. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007ed Papers*, volume 5152, pages 72–79.

Statista. 2021. Worldwide visits to Wikipedia.org from January to June 2021. https://www.statista.com/statistics/1259907/wikipedia-website-traffic/.

Francis M. Tyers and Jacques A. Pienaar. 2008. Extracting Bilingual Word Pairs from Wikipedia. In *Proceedings of*

the SALTMIL Workshop at the Language Resources and Evaluation Conference, LREC2008.

Wikimedia Foundation. 2022a. List of Wikipedias – Meta. https://meta.wikimedia.org/wiki/List_of_Wikipedias.

Wikimedia Foundation. 2022b. Wikipedia, the Free Encyclopedia. https://en.wikipedia.org/wiki/Main_Page.

Wikimedia Foundation. 2022c. Wikipedia, the Free Encyclopedia. https://uk.wikipedia.org/wiki/Main_Page.

Kun Yu and Jun'ichi Tsujii. 2009. Bilingual Dictionary Extraction from Wikipedia. In *Machine Translation Summit*, volume 12.

Serbian Early Printed Books: Towards Generic Model for Automatic Text Recognition using *Transkribus*

Vladimir Polomac*

* Serbian Language Department, Faculty of Philology and Arts, University of Kragujevac
Jovana Cvijića bb, 34 000 Kragujevac, Serbia
v.polomac@filum.kg.ac.rs

Abstract

The paper describes the process of creating and evaluating a new version of the generic model for automatic text recognition of Serbian Church Slavonic printed books within the *Transkribus* software platform, based on the principles of artificial intelligence and machine learning. The generic model *Dionisio 2.0.* was created on the materials of Serbian Church Slavonic books from various printing houses of the 15th and 16th centuries (Cetinje, Venice, Goražde, Mileševa, Gračanica, Belgrade and Mrkša's Church), and, during the evaluation of its performance, it was noticed that CER was about 2–3%. The *Dionisio 2.0.* model will be publicly available to all users of the *Transkribus* software platform in the near future.

1. Introduction

The research on creating a model for automatic text recognition of the Serbian Church Slavonic printed books from Venice using a software platform *Transkribus*,[1] presented in Polomac (2022), represents the starting point for this paper. This paper describes the process of transcription and creation of a specific model[2] for automatic text recognition of *Prayer Book (Euchologion)* printed between 1538 and 1540 in the printing house of Božidar Vuković,[3] as well as the process of creating a generic model[4] for automatic text recognition of other books printed in Venice in the printing house of Božidar Vuković and his son Vićenco.[5] The most important result of this paper is the creation of the first version of the model *Dionisio 1.0.* (named after an Italian pseudonym for Božidar Vuković – *Dionisio della Vechia*) representing the first publicly available resource for automatic reading of Serbian Church Slavonic manuscripts and printed books within the *Transkribus* software platform (cf. https://readcoop.eu/model/dionisio-1-0/).

The *Dionisio 1.0.* model structure is shown in Table 1, and its performance is displayed in Table 2.

Book	**Word count**
Prayer Book (1538–1540)	39,889
Psalter (1519–1520)	10,132
Miscellany for Travellers (1536)	10,618
Festal Menaion (1538)	10,732
Miscellany for Travellers (1547)	10,006
Hieratikon (Liturgikon) (1554)	10,196
Total	91,573

Table 1: *Dionisio 1.0.* Structure and the Amount of Training Data.

Word count	**Number of epochs**[6]	**CER**[7] **on Train set**	**CER on Validation set**
86,347	100	1.66%	2.09%

Table 2: *Dionisio 1.0* Performance.

[1] *Transkribus* (https://readcoop.eu/transkribus) represents an open-access software platform for automatic text recognition and retrieval developed as part of the READ project at the University of Innsbruck. More details about the technological background and operating system cf. Mühlberger et al. (2019).

[2] The functionality of the *Transkribus* platform is particularly manifested in the potential to train one's own automatic text recognition model, irrespective of the language or script used in the manuscript. The training of the automatic recognition model represents an instance of machine learning based on neural networks in which during the learning process the model compares the manuscript photographs and corresponding letters, words and lines of the text in the diplomatic edition. For more details see Mühlberger et al. (2019) and Rabus (2019a).

[3] Božidar Vuković was a Serbian merchant from Zeta (Podgorica and the area surrounding Lake Skadar). After his arrival at Venice (in 1516 at the latest) he acculturated his Serbian name to the new environment by creating a Latin (*Dionisius a Vetula*) and an Italian pseudonym (*Dionisio della Vecchia*) from his Serbian name and the toponym of Starčeva Gorica (at Lake Skadar), indicating his origin (Lazić, 2018). Books from his printery were aimed at the Serbian Orthodox Church and its flock under Ottoman rule, yet the motives of his printing business were not only patriotic and religious, but also mercantile and financial (Lazić, 2020b).

[4] Unlike a specific model that is trained to recognize a single manuscript or printed book, a generic model contains material from different manuscripts or printed books. More details on the possibilities and pitfalls of training generic models can be found in Rabus (2019b).

[5] After the death of Božidar Vuković, Vićenco Vuković had reprinted several of his father's editions until 1561, and later rented his equipment to other Venetian printers. For more details about his life and work see also Pešikan (1994).

[6] The term *epoch* in machine learning stands for "one complete presentation of the data set to be learned to a learning machine" (Burlacu and Rabus, 2021).

[7] The Character Error Rate (CER) is calculated by comparing the automatically generated text and the manually corrected version. See for more details in Transkribus Glossary https://readcoop.eu/glossary/character-error-rate-cer/.

In the continuation of the research, we aimed at examining the performance of the *Dionisio 1.0.* model on Serbian Church Slavonic books created in other printing houses, firstly in Venetian printing houses created after closing Božidar and Vićenco Vuković's printing house, and then in other old Serbian printing houses of the 15th and 16th centuries (Cetinje, Goražde, Mrkša's Church, Belgrade, Mileševa and Gračanica), thus ultimately offering a generic model for the automatic text recognition of Serbian Church Slavonic printed books as a whole.

2. Applying the *Dionisio 1.0.* Model on Books from Other Venetian Printing Houses

In the first experiment, we tested the performance of the *Dionisio 1.0.* model on several Serbian Church Slavonic books printed in Venice after closing Božidar and Vićenco Vuković's printing house: *Lenten Triodion* was printed in 1561 by Stefan of Scutari in the Camillo Zanetti's printing house, *Prayer Book* (*Miscellany for Travellers)* was printed in 1566 by Jakov of Kamena Reka, *Prayer Book (Euchologion)* was created in 1570 in the printing house of Jerolim Zagurović and *Psalter with Appendices* was printed in 1638 in the printing house of Bartol Ginammi (Pešikan, 1994). The starting hypothesis of the paper in the current experiment was that the model trained on the materials of Serbian Church Slavonic books from the printing house of Božidar and Vićenco Vuković would be useful for automatic text recognition of other Venetian editions printed using their printing equipment.

The statistical results of the experiment are shown in the following table.

Book	**CER**
Lenten Triodion (1561)	9.41%
Miscellany for Travellers (1566)	11.63%
Prayer Book (Euchologion) (1570)	13.67%
Psalter with Appendices (1638)	16.04%

Table 3: Application of the *Dionisio 1.0.* model on publications from other Venetian printing houses.

The unexpectedly high CER does not necessarily indicate poor performance of the *Dionisio 1.0.* model. The largest number of errors in text recognition is the result of the fact that in these books accent marks are used differently than in the books from the printing house of Božidar and Vićenco Vuković, which were used to train the *Dionisio 1.0.* model. This fact is especially evident in *Prayer Book (Euchologion)* from the printing house of Jerolim Zagurović (1570) and *Psalter with Appendices* from the printing house of Bartol Ginammi (1638) in which only *spiritus lenis* with an *oxia* over the initial vowel grapheme was used.

To illustrate this claim, we shall use a comparative presentation of a photograph of a part of sheet 2b *Prayer Book (Euchologion)* (1570) and an automatically read text using the *Dionisio 1.0.* model.

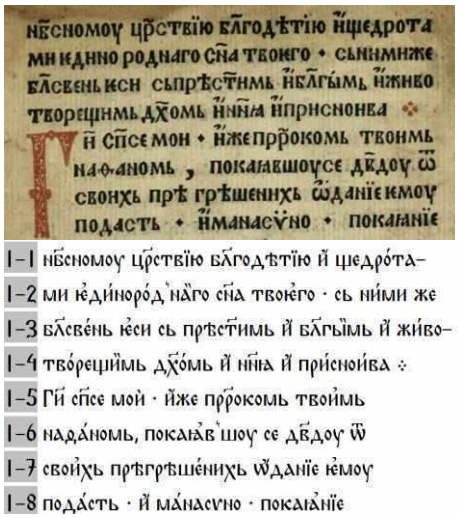


Figure 1: The Automatically Read Text of a Segment of Sheet 2b *Prayer Book (Euchologion)* from 1570.

The greatest number of errors in text recognition refers to cases in which the model outputs accent marks in accordance with the material on which it was trained, although in the text of *Prayer Book (Euchologion)* these marks were not used: so instead of щедротами 1/2, твоѥго 2, ними 2, блсвень ѥси 3, животворещимь 3/4, дхомь 4, присноива 4, мои 5, твоимь 5, наданомь 6, своихь 7, прѣгрѣшенихь 7, ѥмоу 7, подасть 8, манасѵно 8, покаѩнїе 8 the model outputs щедро́тами 1/2, твоѥ́го 2, ни́ми 2, блсве́нь ѥ́си 3, жи́вотво́рещимь 3/4, дхо́мь 4, при́сноива 4, мои̇ 5, твои́мь 5, нада́номь 6, свои́хь 7, прѣгрѣше́нихь 7, ѥ́моу 7, пода́сть 8, ма́насѵно 8, покаѩ́нїе 8. Along with the accent marks, the model inccorectly reads a *pajerak mark* in two examples only: instead of ѥдинороднаго 2, покаѩвшоу 6 there is the incorrect ѥди́норо́д҆на҆го 2, покаѩ́в҆шоу 6. In one example, instead of *oxia* there is an incorrect double *circumflex*: instead of бл҃гы́мь 3 there is the incorrect бл҃гы̏мь 3.

The same problem is exhibited by the comparative presentation of the photograph of a part of sheet 5b *Psalter with Appendices* (1638) and the automatically read text.

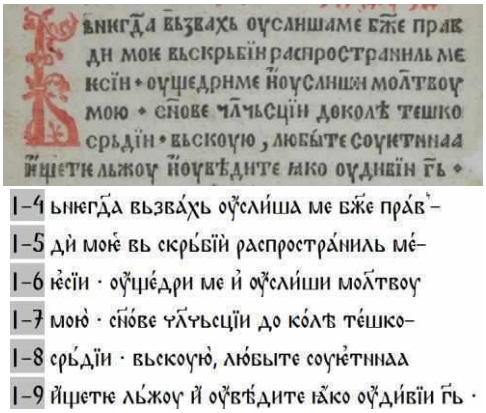


Figure 2: The Automatically Read Text of a Part of Sheet 5b *Psalter with Appendices* from 1638.

Here, too, the largest number of errors refers to cases in which the *Dionisio 1.0.* model outputs accent marks according to the patterns of their use in the Venetian books that served for its training, although in the text of Ginammi's *Psalter with Appendices* these marks were not used. Thus instead of вьзвахъ 4, оуслиша 4, правди 4/5, мое 5, скрьвїи 5, распространиль 5, ме 5, иесїи 6, оущедри 6, оуслиши 6, мою 7, сн҃ове 7, до колѣ 7, тешкосрьдїи 7, вьскоую 8, любыте 8, соуѣтнаа 8, льжоу 9, оувѣдите 9, ꙗко 9, оудивїи 9 the model incorrectly outputs вьзва́хъ 4, оу҃сли́ша 4, пра́в᷄ди 4/5, мое᷈ 5, скрь́вїи 5, распростра́ниль 5, ме́- 5, е᷈сїи 6, оу҃ще́дри 6, оу҃сли́ши 6, мою᷈ 7, сно́ве 7, до ко́лѣ 7, те́шкосрь́дїи 7, вьскоую᷈ 8, лю́быте 8, соуѣ́тнаа 8, ль́жоу 9, оу҃вѣ́дите 9, ꙗ̆ко 9, оу҃ди́вїи 9. Here, as well, the other types of errors are confirmed by isolated examples: *pajerak mark*: instead of правди 1/2 there is the incorrect пра́в᷄ди 1/2; space between words: instead of ме 5 the incorrect ме́- 5; initials: instead of Вьнегда 4 the incorrect ьнегда 4; incorrect accent recognition: instead of и᷈ 6 there is the incorrect и́ 6.

The given examples of the most common errors show that, despite the high percentage of incorrectly recognized characters, after the automatic post-correction of the transcripts which would include accent marks removal using the *Search/Replace chosen chars in transcript* option, the *Dionisio 1.0.* model can also be very efficient in recognizing Serbian Church Slavonic books created in the printing houses of Jerolim Zagurović and Bartol Ginammi during the 16th and 17th centuries.

The greatest number of errors in the automatic recognition of the text *Lenten Triodon* (1561) by Stefan of Scutari and *Prayer Book (Miscellany for Travellers)* (1566) by Jakov of Kamena Reka also refers to the recognition of accent marks. However, what distinguishes these books from the books from the printing houses of Jerolim Zagurović and Bartol Ginammi is that accent marks are actually used, yet in different positions compared to the books from the printing house of Božidar and Vićenco Vuković on which the *Dionisio 1.0.* model was trained. To illustrate this claim, we will first use a comparative presentation of a part of sheet 3a *Lenten Triodon* (1561) by Stefan of Scutari and the automatically read text using the *Dionisio 1.0.* model.

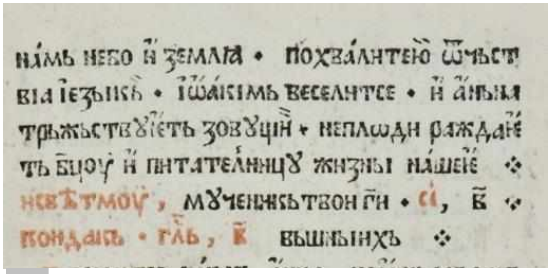


2-1 вале́ща᷄ се соуѣ́тною᷈ · и᷈ да́
2-2 жь съкроу́ше᷈нъ᷄и᷈ дїи᷈ и᷈ ср-
2-3 цоу᷈зовоу́щоу᷈ те́бѣ᷈ · ѡ᷈цѣ́-
2-4 сти᷈ нѣ́дне᷈ чл҃колю́бче е᷈ё, и᷈
2-5 пода́ждь ми о᷈ста́в᷄ленїе ·
2-6 Е᷈ди́нь сы᷈ бл҃гопрѣ᷈мѣ́ни́те-
2-7 ль · пѣ᷈ зі҃р҃а҃ Ро́сода́-
2-8 вицоу᷈ Хвопе́щьсь твори᷈ а҃г
2-9 г҃ль, сѣ᷈тынмъдѣ́темь · хал᷄-

Figure 3: The Automatically Read Text of a Part of Sheet 3a *Lenten Triodon* from 1561.

Errors in accent mark recognition: instead of валеща 1, соуѣтною 2, съкроушенъи 2, сѣцоу 2/3, тебѣ 3, ѡцѣ- 3, о᷈ставленїе 5, пѣщь сътвори᷈ 8, хал- 9/10 the model incorrectly outputs ва́леща᷄ 1, соуѣ́тною᷈ 1, съкроу́ше᷈нъ᷄и᷈ 2, сѣцоу᷈ 2/3, те́бѣ᷈ 3, ѡцѣ- 3, о᷈ста́в᷄ленїе 5, пе́щь сътвори᷈ 8, ха́л᷄- 9/10. Errors in recognizing spaces between words are also of high frequency: instead of да́- 1, цоу᷈ зовоу́щоу᷈ 3, пѣ᷈ зі҃ 7, вво пѣщь сътвори᷈ 8, а҃г- 8, сѣтыимъ дѣ́темь 9 the model incorrectly outputs да́ 1, цоу᷈зовоу́щоу᷈ 3, пѣ᷈зі҃ 7, вво пе́щь сътвори᷈ 8, а҃г 8, сѣтыимъдѣ́темь 9. In a fewer number of examples, errors in recognizing *pajerak mark*, superscript letters and *titlo mark* can be found: instead of съкроуше́нъи᷈ 2, о᷈ставленїе 5, хал- 9/10, сѣ- 2, пѣ᷈ 7 the model incorrectly outputs съкроу́ше᷈нъи᷈ 2, о᷈ста́в᷄ленїе 5, ха́л᷄- 9/10, сѣ- 2, пѣ 7.

A comparative presentation of a part of sheet 7a *Prayer Book (Miscellany for Travellers)* from 1566 and the automatically read text using the *Dionisio 1.0.* model displays similar errors.



1-1 на́мь не бо᷈ и᷈ земла᷄ · похва́литею᷈ ѡ᷈чьст-
1-2 вїа їе́зыкъ᷈ · їѡа́кїмь весе́лит се · и᷈ а᷈нна
1-3 трь́жьствы᷈ѣ́ть зовы᷈щи᷈ · неплѡ́ди раждаі᷈ѐ
1-4 ть е᷈́цоу᷈ и᷈ пи́тател᷄ницъ᷈ жи́зны на́шеѥ ·
1-5 и᷈ свѣ́тмоу᷈, мо᷈у᷈е́никь твои᷈ г҃и · сь, е᷈́ ·
1-6 воі᷈ѐдакь · г҃л᷈ь, е᷈́ вышныихъ ·

Figure 4: The Automatically Read Text of a Part of Sheet 7a *Prayer Book (Miscellany for Travellers)* from 1566.

Errors in recognizing accent: instead of небо 1, земла 1, похва́лите ю᷈ 1, ѡ᷈чьствїа 1/2, їезыкъ᷈ 2, весе́лит 2, трьжьствы᷈ѣ́ть 3, неплѡди 3, раждаі᷈ѐ- 3, питател᷄ницъ᷈ 4, жизны 4, на́шеѥ᷈ 4, и 5, мо᷈у᷈е́никь 5, кондакь 6 the model incorrectly outputs не бо᷈ 1, земла᷄ 1, похва́литею᷈ 1, ѡ᷈чьствїа 1/2, їе́зыкъ᷈ 2, весе́лит 2, трь́жьствы᷈ѣ́ть 3, неплѡ́ди 3, раждаі᷈ѐ 3, пи́тател᷄ницъ᷈ 4, жи́зны 4, на́шеѥ 4, и᷈ 5, мо᷈у᷈е́никь 5, воі᷈ѐдакь 6. A certain number of errors is connected to recognizing spaces between words: instead of небо 1, похва́лите ю᷈ 1, раждаі᷈ѐ- 3, свѣт моу᷈ 5 the model incorrectly outputs не бо᷈ 1, похва́литею᷈ 1, раждаі᷈ѐ 3, свѣтмоу᷈ 5. Several errors in recognizing letters may perhaps be related to poor quality of the photograph: instead of ѕі 5, кондакь 6 the model incorrectly outputs сь 5, воі᷈ѐдакь 6.

The illustrated examples of the most frequent errors in *Lenten Triodon* (1561) and *Prayer Book (Miscellany for Travellers)* (1566) show that the *Dionisio 1.0.* model can be used for obtaining transcripts that can, after appropriate manual correction, be used for creating specific models for automatic text recognition of the aforementioned two books.

3. Applying the *Dionisio 1.0.* Model on Books from Other Serbian Printing Houses of the 15th and 16th Centuries

In the second experiment, the performance of the *Dionisio 1.0.* model was tested on selected books from other printing houses of the 15th and 16th centuries (Cetinje, Goražde, Gračanica, Mileševa, Belgrade and Mrkša'a Church). During the research, we started from the hypothesis that the model trained on the material of books from the Venetian printing house Vuković will be useful for books from other printing houses, since there are not many orthographic variations in Serbian Early Printed Books as there are in medieval manuscripts.

The results of the experiment are shown in the following table.

***Book* (Printed House, Year)**	**CER**
Octoechos, mode 1–4 (Cetinje, 1495)	8.24%
Psalter with Appendices (Goražde, 1519)	6.44%
Octoechos, *mode 5–8* (Gračanica, 1539)	11.11%
Prayer Book (Euchologion) (Mileševa, 1546)	5.43%
Tetraevangelion (Belgrade, 1552)	11,28%
Tetraevangelion (Mrkša's Church, 1562)	12.06%

Table 4: Application of the *Dionisio 1.0.* model on publications from other printing houses in the 15th and 16th centuries.

Based on the previous table, it can be concluded that the *Dionisio 1.0.* model achieved the best results in the automatic recognition of the text of *Prayer Book (Euchologion)* (1546) from the printing house of the Mileševa monastery and *Psalter with Appendices* (1521) from the Goražde printing house. These results can be explained by the fact that *Prayer Book (Euchologion)* (1546) had been printed in Mileševa with the same typographic characters as *Psalter with Appendices* (1521) from Božidar Vuković's printing house, as well as by the fact that *Psalter with Appendices* (1519) was printed in Goražde using the typographic equipment imported from Venice (Lazić, 2020a).[8]

To illustrate the efficiency of the *Dionisio 1.0.* model we may firstly use the comparative presentation of the photograph of a part of sheet 5b *Prayer Book (Euchologion)* (1546) from the printing house of the Mileševa monastery and the automatically read text in Figure 5.

In this book, as well, the greatest number of errors refers to accent marks recognition: instead of и҃мéни 4, йстúныи 4, і҆ѣдúнорóднаго 4/5, сѓтаго 5, и мѣ 7, сподобльшаго 7, w҃ноу҆ 10 the *Dionisio 1.0.* incorrectly outputs и҃мéнѝ 4, йстúный 4, і҆ѣдúнорóднаго 4/5, сѓтáго 5, й мѣ 7, сподóбльшаго 7, w҃ноу 10. Other errors are fewer in number and relate to recogizing initials, spaces between words and *pajerak mark*: instead of w҃ и҃мéни 4, подь 9 и дрѣвнꙋ̀ю 10 the model incorrectly reads и҃мéнѝ 4, по дь 9 и дрѣв҃нꙋ̀ю 10.

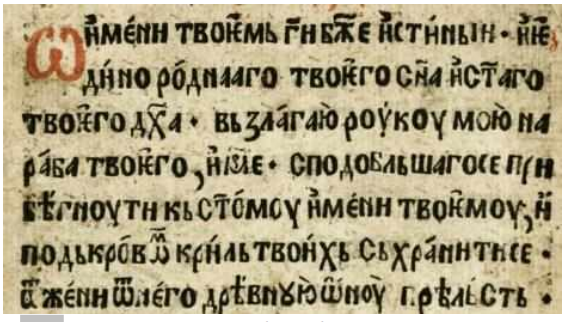


Figure 5: The Automatically Read Text of a Part of Sheet 5b *Prayer Book (Euchologion)* from 1546.

Similar errors are indicated by the comparative illustration of the photograph of a part of sheet 35a *Psalter with Appendices* (1519) from the Goražde printing house and the automatically read text using the *Dionisio 1.0.* model.

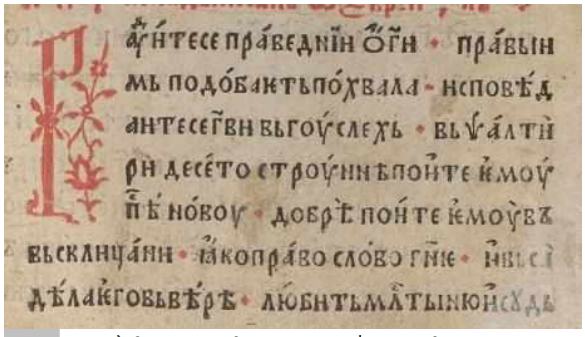


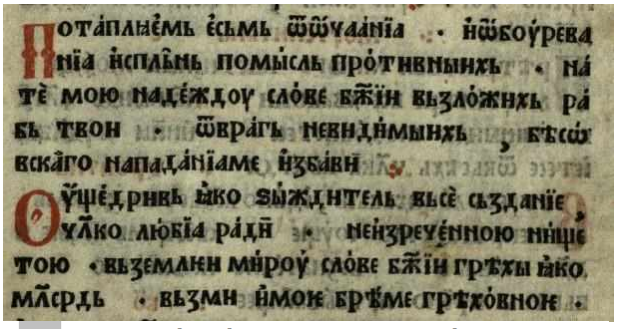
Figure 6: The Automatically Read Text of a Part of Sheet 35a *Psalter with Appendices* from 1519.

[8] Scholars likewise claim that *Psalter with Appendices* (1519) and *Prayer Book (Euchologion)* (1544) from Goražde printing house could have been printed in Venice, as well, which corresponds to the widespread practice of the time to place a counterfeit place of printing on the colophonies of editions (Lazić, 2020a).

The previous illustration demonstrates how the *Dionisio 1.0.* model makes the most frequent errors while recognizing accent marks: instead of пра́ведніи 14, подо́баѥть 15, по́хвала 15, исповѣ́даите се 15/16, ѱа́лтѷри 16/17, ѥмоу́ 17, добрѣ 18, ѥго 19, мл҃тыню 19, сѣдь 19 the model incorrectly outputs пра́веднїи 14, подо́баѥть 15, похва́ла 15, йсповѣ́д҃аите се 15/16, ѱа́л҃тѷрй 16/17, ѥмоу 17, до́брѣ 18, ѥго̀ 19, мл҃тыню 19, сѣдь 19. The other errors pertain to recognizing spaces between words, *pajerak mark* and initials: instead of пра́выи- 14, ѱа́лтѷ- 16, десе́тостроу́ннѣ 17 the model incorrectly reads: пра́выи 14, ѱа́л҃тѷ 16, десе́то строу́ннѣ 17; instead of исповѣ́даите се 15/16, ѱа́лтѷ- 16 there is the incorrect йсповѣ́д҃аите се 15/16, ѱа́л҃тѷ 16; instead of Раꙋ́ите се 14 there is the incorrect ПРаꙋ́ите се 14. There is merely one example of an incorrectly recognized letter: instead of вьсклица́ни 19 the model incorrectly reads вь свлица́ни 19.

The *Dionisio 1.0.* model also shows a similar performance during the automatic recognition of the text of the oldest printed Serbian Church Slavonic book – *Octoechos, mode 1–4* (1495) from the Cetinje printing house. The percentage of unrecognized characters is somewhat higher than in the previous two books due to poor photo quality and issues with recognizing certain letters and punctuation marks.

To illustrate the efficiency of the model, we will use a comparative presentation of a part of sheet 33b and the automatically read text in the following figure.

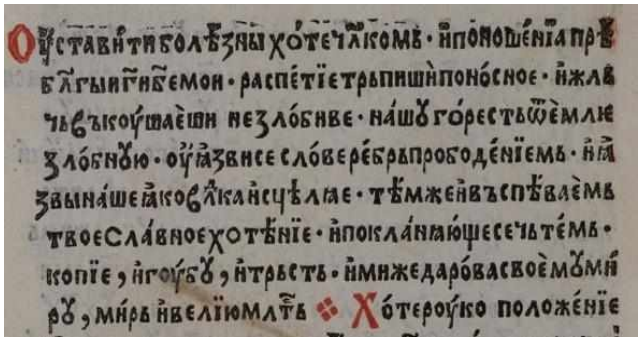


1-8 Пота́плнѣ́мь ѣсьмь ѿ ѿча́нїа · и ѿбоу́рева
1-9 нїа йспльнь помы́сль про́тивныихь · на
1-10 тѐ мою на де́ждоу сло́ве бж̃їи вьзло́жихь ра́-
1-11 бь твои · ѿ вра́гь невиди́мыихь · ⁘ вѣсѡ́
1-12 в҃ска́го напада́нїа ме йзба́ви
1-13 Оу҃ме́дривь ꙗ̀ко бы́ждитель вьсѐ сьзда́нїе ·
1-14 ꙗ̀колюбїа ра́дѝ · неизрече́нною ни҃ще҃
1-15 тою · вь зе́млѥй ми́роу сло́ве бж̃їи грѣхы ꙗ̀ко
1-16 мл҃содь · вьзми и̇̀ моѥ̀ брѣ́ме грѣхо́вное ·

Figure 7: The Automatically Read Text of a Part of Sheet 33b *Octoechos, mode 1–4* from 1495.

In this book, too, the largest number of errors in the automatic text recognition occurs with accent marks: instead of ѣсьмь 8, йспльнь 9, на́ 9, мою 10, твои 11, вѣсѡ́вскаго 11/12, йзба́ви 12, ꙗ̀ко 13, сьзда́нїе 13, неизрече́нною 14, ни҃ще҃тою 14/15, зе́млѥй 15, ꙗ̀ко 15, вьзми 16, и̇̀ 16 the *Dionisio 1.0.* model incorrectly reads: ѣсьмь 8, йспльнь 9, на 9, мою̀ 10, твоѝ 11, вѣсѡ́ в҃ска́го 11/12, йзба́ви 12, ꙗ̀ко 13, сьзда́нїе 13, неизрече́нною 14, ни҃ще҃тою̀ 14/15, зе́млѥй 15, ꙗ̀ко 15, вь҃зми 16, и̇̀ 16. The issues with recognizing spaces between words and *pajerak mark* can be illustrated by the following examples: instead of ѿбоу́рева- 8, наде́ждоу 10, вѣсѡ́- 11 there is the incorrect ѿбоу́рева 8, на де́ждоу 10, вѣсѡ́ 11; instead of вѣсѡ́вска̀го 11/12 there is the incorrect вѣсѡ́ в҃ска̀го 11/12. In this book, as we have already mentioned, the *Dionisio 1.0.* model likewise incorrectly recognizes certain letters and punctuation marks: instead of ѿ 8, зы̀ждитель 13, мл҃сердь 16 there is the incorrect ѡ̀ 8, вы̀ждитель 13, мл҃содь 16; instead of невиди́мыихь, 11, йзба́ви ⁘12 there is the incorrect невиди́мыихь · ⁘ 11 и̇̀зба́ви ⁘12.

In the rest of the books listed in Table 4, (*Octoechos, mode 5–8* (1539) from Gračanica, *Tetraevangelion* (1552) from Belgrade and *Tetraevangelion* (1562) from Mrkša's Church), the CER is slightly higher, around 11–12%. The categories in which the *Dionisio 1.0.* model outputs errors are mostly the same in all three books, so we will only take a comparative presentation of a part of sheet 27b *Octoechos, mode 5–8* (1539) from Gračanica and the automatically read text as an illustration.



1-1 Оу҃ста́вити бо́лѣзны̀ хо́те ꙗ̀комь · и̇̀ попоше́нїа пр҃ѣ-
1-2 бл҃гыи г҃и бѐ моѝ · распе́тїе тр҃пишѝ поно́сноѐ · и̇̀ жл҃в
1-3 ꙋьвькоу́шаѐши незло́бивѣ · на́шѪ го́рость ѡ̀млѥ̀
1-4 зло́бнꙋю · оу҃ꙗ̀зви се сло́ве ре́брь провод҃е́нїемь · и̇̀ ꙗ̀-
1-5 звы на́ше ꙗ̀ко вл҃ка йсцѣлꙗѐ · тѣ́мъ̀же и̇̀ вьспѣва́емь
1-6 твоѐ сла́вноѐхотѣ́нїе · и̇̀ покла́ня́ющѝе се ꙋь те́мь ·
1-7 копїе, и̇̀ гоу҃бѢ, и̇̀ тр҃сть · и̇̀ ми́же дарова́ своѐ мѢ́ми-
1-8 рѢ, ми́рь и̇̀ ве́лїю мл҃ть ⁘ Хо́те роу́ко положе́нїе

Figure 8: The Automatically Read Text of a Part of Sheet 27b *Octoechos, mode 5–8* from 1539.

The greatest number of errors is related to the recognition of accent marks: instead of бо́лѣзны 1, и̇̀ 1, 2, 5, 6, 8, мои 2, тр҃пишѝ 2, поно́сноѐ 2, ꙋь вькоу́шаѐши 3, ѡ̀млѥ 3, провод҃е́нїемь 4, ꙗ̀звы 4/5, ꙗ̀ко 5, йсцѣлꙗе 5, вьспѣва́емь 5, твое 6, сла́вное хотѣ́нїе 6, покла́ня́ющѝе се 6, и̇̀миже 7, своѐмѢ ми- 7, ве́лїю 8 the *Dionisio 1.0.* model incorrectly outputs бо́лѣзны̀ 1, и̇̀ 1, 2, 5, 6, 8, моѝ 2, тр҃пишѝ 2, поно́сное 2, ꙋьвькоу́шаѐши 3, ѡ̀млѥ̀ 3, провод҃е́нїемь 4, ꙗ̀звы 4/5, ꙗ̀ко

5, и҆сцѣ́лꙗѐ 5, вьспѣ́ваѐмь 5, твоѐ 6, сла́вноѐхотѣ́нїе 6, покла́нꙗю҆ще се 6, и҆ ми́же 7, своѐ мꙋ́ми- 7, ве́лїю 8. Recognizing spaces between words represents the problematic issue in a multitude of cases: instead of жль- 2, чь вькоу́шаѐши 3, сла́вное хотѣ́нїе 6, чьте́мь 6, и҆миже 7, своѐмꙋ ми́- 7, роу́коположе́нїе 8 the model incorrectly outputs жл҃в 2, чьвькоу́шаѐши 3, сла́вноѐхотѣ́нїе 6, чь те́мь 6, и҆ ми́же 7, своѐ мꙋми́- 7, роу́ко положе́нїе 8. The other errors pertain to the recognition of superscript letters and *pajerak mark*, as well as regular letters in a few examples: instead of жль- 2, мл҃ть 8 the model outputs жл҃в 2, млт҃ь 8; instead of тѣ́мже 5 there is the incorrect тѣ́м҆же 5; instead of поноше́нїа 1, жль- 2, ѿѐмлѥ 3 the model reads попоше́нїа 1, жл҃в 2, ѡ҆ѐмлѥ̏ 3.

The quantitative and qualitative analysis conducted in this chapter demonstrates that the *Dionisio 1.0.* recognizes the text of the Serbian Church Slavonic books created in other printing houses of the 15th and 16th centuries with varying degrees of success. The quantitative analysis shows that the lowest CER was recorded in books from Mileševa and Goražde printing houses, which is expected considering the fact that these books were printed using the typographic printing equipment from Venice. An acceptable CER was noted during the recognition of *Octoechos, mode 1–4* (1494) from the Cetinje printing house, while this percentage exhibited in books from other printing houses (Belgrade, Gračanica, Mrkša's Church) underscores the need for training a new version of the generic model with improved performance. The qualitative analysis showed that the *Dionisio 1.0.* model usually makes errors when recognizing accent marks, but also when recognizing spaces between words. The errors in recognizing superscript letters, *pajerak mark*, initials and regular letters are far less common.

4. Creation and evaluation of the generic model *Dionisio 2.0.*

When creating a new version of the model, we started from the transcripts of Serbian Church Slavonic books listed in Table 4 obtained using the *Dionisio 1.0.* model. By means of the manual correction of the transcripts, the Ground Truth[9] data was obtained for training the generic model *Dionisio 2.0.* In accordance with our findings on the interdependence of model success and the amount of training data (Polomac, 2022), as well as similar findings for Church Slavonic books from the Berlin State Library (Neumann, 2021), the goal was set to provide a critical mass of at least 10000 words for each printed book in order to train the generic model *Dionisio 2.0.* While training the generic model *Dionisio 2.0.* we used the Ground Truth data prepared for the *Dionisio 1.0.* model (see Table 1 here), as well as the new Ground Truth data from Serbian Church Slavonic books printed in other printing houses of the 15th and 16th centuries listed in the following table.

***Book* (Printed House, Year)**	**Word count**
Octoechos, mode 1–4 (Cetinje, 1495)	15,667
Psalter with Appendices (Goražde, 1519)	16,445
Octoechos, mode 5–8 (Gračanica, 1539)	15,179
Prayer Book (Euchologion) (Mileševa, 1546)	15,003
Tetraevangelion (Belgrade, 1552)	15,333
Tetraevangelion (Mrkša's Church, 1562)	15,733

Table 5: The *Dionisio 2.0.* model – Ground Truth data from other printing houses of the 15th and 16th centuries.

The performance of the generic model *Dionisio 2.0.* is shown in the following table.

Word count	**Number of epochs**	**CER on Train set**	**CER on Validation set**
176,481	200	2.03%	2.44%

Table 6: Performance of the generic model *Dionisio 2.0.*

In order to compare the performance of the two models, we tested them on ten sheets from *Psalter with Appendices* (1495) from the Cetinje printing house and *Hieraticon* (1521) from the Goražde printing house, the latter two representing Serbian Church Slavonic books that did not form the material for training the model. The results of the experiments are shown in the following table.

***Book* (Printed House, Year)**	***Dionisio 1.0.* CER**	***Dionisio 2.0.* CER**
Psalter with Appendices (Cetinje, 1495)	5.71%	1.50%
Hieraticon (Goražde, 1521)	9.38%	4.61%

Table 7: Comparing the Performance of the Two Models on Books from Cetinje and Goražde Printing Houses.

As can clearly be seen from the previous table, the *Dionisio 2.0.* model displays significantly better results compared to the *Dionisio 1.0.* model. To illustrate the exceptional efficiency of the *Dionisio 2.0.* model we provide a comparative presentation of a part of sheet 3b *Psalter with Appendices* (1495) from Cetinje printing house and the automatically read text in the figure 9.

As we can see in the figure, the *Dionisio 2.0.* model erros only in a few examples in which the *spiritus lenis* and *perispomena* are insufficiently clearly differentiated: instead of ю҆нїи 8, пою҆ть 9, и҆сти́ною҆ 10, нака́зоую҆ть 11 the model incorrectly outputs ю҆нїи 8, пою҆ть 9, и҆сти́ною҆ 10, нака́зоую҆ть 11. There is a single example of the model mixing *spiritus lenis* and *oxia*: instead of ѿи҆де 13 there is the incorrect ѿи́де 13. The space between words was also

[9] The term Ground Truth Data in machine learning refers to completely accurate data used to train the model. In our case, these would be exact transcripts of digital photographs of the manuscript. For more details on this term, see Transkribus Glossary at https://readcoop.eu/glossary/ground-truth/.

incorrect in one example solely: instead of мнѣти 9 there is the incorrect мнѣ ти 9. In the other examples on the shown part of sheet 3b the *Dionisio 2.0.* model regularly recognizes letters, spaces between words, *titlo* and accent marks. The exceptional efficiency of the *Dionisio 2.0.* model in recognizing *Psalter with Appendices* (1495) from the Cetinje printing house, especially compared to *Hieraticon* (1521) from the Goražde printing house, has resulted from the fact that there are no superscipt letters in *Psalter with Appendices* (1495), while accent marks are given in expected positions.

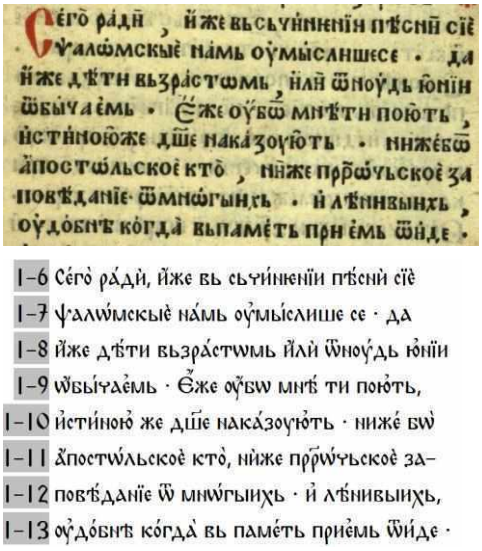


Figure 9: The Automatically Read Text of a Part of Sheet 3b *Psalter with Appendices* (1495).

On the other hand, superscript letters, as well as accent marks, found frequently in unexpected positions, are present in *Hieraticon* (1521) from the Goražde printing house, which definitely affects a somewhat less efficient CER in this book. To illustrate the aforementioned, we shall use the comparative presentation of a part of sheet 9b and the automatically read text in the following figure.

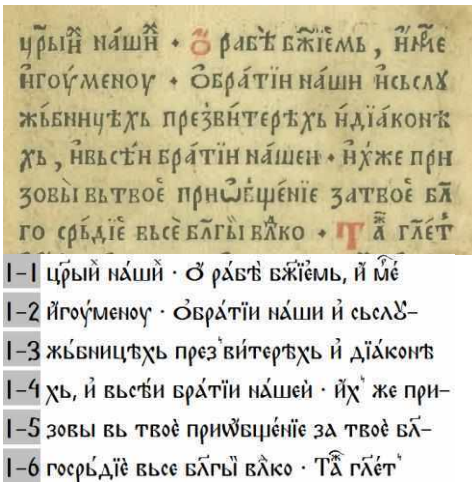


Figure 10: The Automatically Read Text of a Part of Sheet 9b *Hieraticon* (1521).

The previous illustration points to the fact that the *Dionisio 2.0.* model makes errors almost exclusively during accent marks recognition. Thus, instead of рабѣ 1, бжїѣмь 1, мѣ 1, йгоу́меноу 2, и́ 3, на́шеи 4, йх҃ 4, призовы 4/5, твоѥ̀ 5x2, приѡбще́нїе 5, бл҃госрьдїѥ̀ 5/6, вьсѣбл҃гы̀ 6 the model incorrectly reads ра́бѣ 1, бжїѣмь 1, мѣ́ 1, йгоу́меноу 2, и́ 3, на́шеѝ 4, йх҃ 4, призовы 4/5, твоѥ̀ 5x2, приѡбще́нїе 5, бл҃госрьдїѥ̀ 5/6, вьсе бл҃гы̀ 6. Along with the aforementioned errors, there are a few examples of incorrect recognition of spaces between words: instead of о̀ бра́тїи 2, сь слѫ-2, дїа́конѣ- 3 вьсѣбл҃гы̀ 6 the model reads о̀бра́тїи 2, сьслѫ- 2, дїа́конѣ 3 вьсе бл҃гы̀ 6.

5. Concluding Remarks

The research showed how the *Transkribus* software platform, based on the principles of machine learning and artificial intelligence, could be used to create efficient models for automatic text recognition of Serbian Church Slavonic printed books from the end of the 15th to the middle of the 17th century. Having in mind the limitations of the *Dionisio 1.0.* model in the automatic recognition of the text of the Serbian Church Slavonic books printed outside Venice, the paper describes the process of creating a generic model *Dionisio 2.0.*, capable of recognizing Serbian Church Slavonic printed books as a whole. The generic model *Dionisio 2.0.* was trained on the material of the Serbian Church Slavonic books printed in various Serbian printing houses of the 15th and 16th centuries: Cetinje, Venice, Goražde, Gračanica, Mileševa, Belgrade and Mrkša's Church. The quantitative analysis of the performance of this model showed that it could be used to automatically obtain transcripts with a minimum percentage of incorrectly recognized characters (about 2-3%). Most frequently, CER depends on the quality of the photo of the book, the frequency of use of accent marks and superscripts, as well as the correct use of accent marks in the appropriate positions. Using the *Dionisio 2.0.* model transcripts of Serbian Church Slavonic printed books can be obtained automatically, which, after being edited by a competent philologist, can be used for further philological and linguistic research, primarily for creating searchable digital editions of books, as well as electronic corpora, thus creating opportunities for diachronic research of Serbian early modern literacy on a large quantity of data. In the near future, the generic model *Dionisio 2.0.* will become publicly available to all users of the *Transkribus* software platform, which will enable further improvement of its performance, which could ultimately lead to the creation of a generic model for automatic text recognition of Church Slavonic printed books as a whole.

6. Acknowledgment

The research conducted in the paper was financed by the Ministry of Education, Science and Technological Development of the Republic of Serbia, contract no. 451-03-68/2022-14/ 200198, as well as by the German Academic Exchange Service (DAAD) within the project *Automatic Text Recognition of Serbian Medieval*

Manuscripts and Early Printed Books: Problems and Perspectives.

7. References

Constaţa Burlacu and Achim Rabus. 2021. Digitising (Romanian) Cyrillic using Transkribus: new perspectives. *Diacronia*, 14:1–9.

Miroslav Lazić. 2018. Od Božidara Vukovića do Dionizija dela Vekije: identitet i pseudonim u kulturi ranog modernog doba. In: Anatolij A. Turilov et al., eds., *Scala Paradisi*, pages 165–185, SANU, Beograd.

Miroslav Lazić. 2020a. Inkunabule i paleotipi: srpskoslovenske štampane knjige od kraja 15. do sredine 17. veka. In: Vladislav Puzović and Vladan Tatalović, eds., *Osam vekova autokefalije Srpske pravoslavne crkve*, Vol. 2, pages 325–344. Sveti arhijerejski sinod Srpske pravoslavne crkve–Pravoslavni bogoslovski fakultet, Beograd.

Miroslav Lazić. 2020b. Between an Imaginary and Historical Figure: Božidar Vuković´s Professional Identity. *Ricerche Slavistiche*, 43:141–156.

Vladimir Neumann, 2021. Deep Mining of the Collection of Old Prints *Kirchenslavica Digital*. *Scripta & e-Scripta* 21: 207–216.

Vladimir Polomac. 2022. Serbian Early Printed Books from Venice. Creating Models for Automatic Text Recognition using *Transkribus*. *Scripta&e-Scripta*, 22 [in print].

Günther Mühlberger, L. Seaward, M. Terras, S. Oliveira Ares, V. Bosch, M. Bryan, S. Colluto, H. Déjean, M. Diem, S. Fiel, B. Gatos, A. Greinoecker, T. Grüning, G. Hackl, V. Haukkovaara, G. Heyer, L. Hirvonen, T. Hodel, M. Jokinen, P. Kahle, M. Kallio, F. Kaplan, F. Kleber, R. Labahn, M. Lang, S. Laube, G. Leifert, G. Louloudis, R. McNicholl, J. Meunier, J. Michael, E. Mühlbauer, N. Philipp, I. Pratikakis, J. Puigcerver Pérez, H. Putz, G. Retsinas, V. Romero, R. Sablatnig, J. Sánchez, P. Schofield, G. Sfikas, C. Sieber, N. Stamatopoulos, T. Strauss, T. Terbul, A. Toselli, B. Ulreich, M. Villegas, E. Vidal, J. Walcher, M. Wiedermann, H. Wurster, and K. Zagoris. 2019. Transforming scholarship in the archives through handwrittwn text recognition. *Journal of Documentation*, 5 (75):954–976.

Mitar Pešikan. 1994. Leksikon srpskoslovenskog štamparstva. In: Mitar Pešikan et al., eds., *Pet vekova srpskog štamparstva 1494–1994: razdoblje srpskoslovenske štampe XV–XVII*, pages 71–218, Narodna biblioteka Srbije–Matica srpska, Beograd.

Achim Rabus. 2019a. Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach using *Transkribus*. *Scripta & e-Scripta*, 19:9–32.

Achim Rabus. 2019b. Training Generic Models for Handwritten Text Recognition using Transkribus: Opportunities and Pitfalls. In: *Proceeding of the Dark Archives Conference*, Oxford, 2019b, in print.

Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref

Eva Pori,* Jaka Čibej,* Tina Munda,† Luka Terčon,† Špela Arhar Holdt*†

* Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana
eva.pori@ff.uni-lj.si; jaka.cibej@ff.uni-lj.si; spela.arharholdt@ff.uni-lj.si

† Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Večna pot 113, 1000 Ljubljana
tina.munda@fri.uni-lj.si; luka.tercon@fri.uni-lj.si

Povzetek

V prispevku predstavimo proces in rezultate ročnega pregledovanja lem in oblikoskladenjskih oznak MULTEXT-East v6 korpusa SentiCoref, ki bo pod okriljem projekta Razvoj slovenščine v digitalnem okolju (RSDO) vključen v novi učni korpus za slovenščino (trenutni ssj500k). Opišemo delotoke označevalne kampanje, ki je ena najobsežnejših tega tipa v našem prostoru, označevalne dileme, ki so razkrile določene vrzeli v referenčnih označevalnih smernicah, kot tudi rešitve in rezultate, ki smo jih oblikovali med delom in jih bo mogoče uporabiti v prihodnje.

Lemmatization and Morphosyntactic Annotation in the SentiCoref Corpus

The paper presents the process and the results of manual lemma and MULTEXT-East v6 morphosyntactic tag annotation in the SentiCoref corpus, which is planned to be included in the new Slovene training corpus (currently known as ssj500k) as part of the "Development of Slovene in a Digital Environment" project. The paper describes the workflows of the annotation campaign – which was among the most extensive campaigns of this type in Slovenia –, the annotation dilemmas that revealed gaps in previous versions of annotation guidelines, as well as the resulting solutions that will be useful in future annotation campaigns.

1. Uvod

Med leti 2020 in 2023 s podporo Ministrstva za kulturo Republike Slovenije in Evropskega sklada za regionalni razvoj poteka aplikativni projekt Razvoj slovenščine v digitalnem okolju (RSDO).[1] Med cilji projekta je infrastruktura za kontinuirano grajenje slovenskih korpusov: delotoki sprotnega zbiranja besedil, označevalni cevovod in dokumentacija za označevanje na različnih jezikovnih ravneh ter nekatera nova orodja za ročno označevanje ter pregledovanje korpusnih podatkov. Kot temeljna jezikovna vira za razvoj cevovoda za strojno označevanje sodobne slovenščine sta v nadgradnjo vključena tudi leksikon besednih oblik Sloleks (Dobrovoljc et al., 2019) in učni korpus ssj500k (Krek et al., 2020), s katerim se povezuje tudi pričujoči prispevek.

Učni korpus ssj500k v različici 2.3 (Krek et al., 2021) vsebuje 27.829 povedi oz. 500.295 besednih pojavnic, označenih na ravneh od stavčne segmentacije, tokenizacije, lematizacije, oblikoslovja in oblikoskladnje prek odvisnostne skladnje, imenskih entitet in večbesednih leksemov do udeleženskih vlog. Kot je značilno za učne korpuse, so jezikoslovne oznake ročno pregledane, s čimer je dosežena zanesljivost, ki jo potrebujemo za nadzorovano učenje strojnih postopkov. Na rezultate vplivata tudi obseg in zastopanost gradiva, zato je glavni cilj nadgradnje povečanje učnega korpusa na 1.000.000 besednih pojavnic. Na projektu bo za višje, kompleksnejše nivoje označevanja pripravljeno omejeno število novooznačenih povedi, osnovni nivoji pa bodo ročno pregledani za vse novo gradivo.

V prispevku predstavljamo označevalno kampanjo, v kateri smo ročno pregledali in popravili tokenizacijo, segmentacijo, leme in oblikoskladenjske oznake sistema MULTEXT-East (Erjavec, 2012) v korpusu SentiCoref 1.0 (Žitnik, 2019), ki predstavlja približno 76 % predvidene povečave učnega korpusa.[2] SentiCoref vsebuje besedila z novičarskih portalov, v katera so ročno vpisane oznake koreferenc in imenskih entitet, in odgovarja na potrebo, da se v učni korpus vključi gradivo, ki omogoča označevanje jezikovnih značilnosti prek meja povedi (Arhar Holdt in Čibej, 2021).

Namen prispevka je opisati delo, rezultate, zlasti pa označevalne dileme na ravni lem in oblikoskladnje, ki so razkrile določene vrzeli v referenčnih označevalnih smernicah (Holozan et al., 2008), kot tudi rešitve, ki smo jih oblikovali med delom in jih je mogoče uporabiti za prihodnje primerljive naloge. Novi učni korpus bo skupaj z nadgrajenimi označevalnimi smernicami ob zaključku projekta RSDO odprto na voljo na repozitoriju CLARIN.SI.

2. Preteklo in sorodno delo

Učni korpus ssj500k se kot referenčni vir za nadzorovano učenje strojnega jezikoslovnega označevanja sodobnih slovenskih pisnih besedil razvija že več kot desetletje (Krek et al., 2020). Do sedaj so bili na tem korpusu naučeni različni označevalniki, npr. Obeliks (Grčar et al., 2012), ReLDI (Ljubešić in Erjavec, 2016), nevronski označevalnik, ki ga je razvil Belej (2018), in CLASSLA StanfordNLP (Ljubešić in Dobrovoljc, 2019), ki se nadalje razvija tudi na projektu RSDO.

Začetki učnega korpusa segajo v čas projekta MULTEXT-East, ki je spodbudil razvoj sistema za oblikoskladenjsko označevanje (tudi) slovenščine (Dimitrova et al., 1998). Sistem oznak je bil revidiran in nadgrajen pod okriljem projekta Jezikoslovno označevanje slovenščine (JOS), v katerem je nastal korpus jos100k (Erjavec in Krek, 2008). Nato je bilo v projektu

[1] Spletna stran, ki predstavlja projektne cilje in sodelujoče partnerje: https://slovenscina.eu/.

[2] Za preostalih 24 % so v načrtu raznolike besedilne množice, ki bodo zagotovile (a) temelje za semantično označevanje, kot npr. slovenska različica vzporednega korpusa Elexis-WSD (Martelli et al., 2022), (b) izbrane nezastopane besedilne vrste, npr. tvite, ki predstavljajo uporabniško generirane spletne vsebine, (c) v rabi redkejše dvoumne besedne oblike: enakopisne zaimke, dvojinske oblike ipd. (Arhar Holdt in Čibej, 2021: 49–50).

Sporazumevanje v slovenskem jeziku pregledanih dodatnih 400.000 besed, pripravljene pa so bile tudi referenčne smernice za označevanje lem in oblikoskladnje po sistemu JOS oz. MULTEXT-East v4 (Holozan et al., 2008). Trenutna različica korpusa vsebuje oznake sistema MULTEXT-East v6, ki na sistemski ravni vsebuje 1.900 možnih oznak z informacijo besedne vrste in različnih slovarsko-slovničnih značilnosti, kot so npr. spol, sklon, število in lastnoimenskost pri samostalnikih.[3]

SentiCoref 1.0 (Žitnik, 2019) je korpus z 837 besedili oz. približno 433.000 pojavnicami, ki je bil vzorčen iz korpusa SentiNews 1.0 (Bučar, 2017). Čeprav SentiCoref 1.0 neposredno ne vsebuje enakih oznak sentimenta kot SentiNews 1.0, sta korpusa medsebojno povezljiva. SentiCoref 1.0 vsebuje tudi oznake imenskih entitet (oseb, organizacij in lokacij) ter koreferenc na imenske entitete skupaj s koreferenčnimi verigami, ki označujejo sentiment za vsako entiteto. SentiCoref 1.0 je odprto dostopen pod licenco CC BY 4.0 na repozitoriju CLARIN.SI, in sicer v tabelaričnem formatu TSV3, ki ga podpira označevalno orodje INCEpTION (Klie et al., 2018), naslednik orodja WebAnno (Eckart de Castilho et al., 2014).

3. Priprava na označevanje

3.1. Priprava podatkov

SentiCoref 1.0 je sicer tokeniziran, ne vsebuje pa lem in oblikoskladenjskih oznak. Kar zadeva delitev na pojavnice, SentiCoref 1.0 ni bil zasnovan z mislijo na potencialne dodatne jezikoslovne nivoje označevanja, zato v nekaterih primerih odstopa od tokenizacijskih pravil, ki jih pri označevanju korpusov trenutno uporabljamo v slovenskem prostoru (označevalnik *classla*[4] oz. vanj vključeni tokenizator *Obeliks*[5]), npr. pri deljenju kratic ("STA-jev" > "STA", "-", "jev") in številnikov ("2,356" > "2", ",", "356"). Prav tako delitev v SentiCorefu 1.0 ne vsebuje podatkov o presledkih. Pred strojnim oblikoskladenjskim označevanjem in ročnim popravljanjem oblikoskladenjskih oznak je bilo treba najprej popraviti tokenizacijo (vzporedno z njo tudi strojno lematizacijo) ter razdeliti besedilo na povedi (stavčna segmentacija). Za pregledovanje smo korpus pripravili v tabelaričnem formatu v okolju *Google Preglednice* (ang. *Google Sheets*), saj INCEpTION ne podpira spreminjanja tokenizacije. Tokenizacija je bila v celoti popravljena ročno, stavčna segmentacija pa je bila najprej strojno pripisana (na podlagi ločil), nato pa ročno pregledana in potrjena.

Pri pregledovanju segmentacije je bilo 17.095 strojno pripisanih koncev povedi ročno potrjenih kot ustreznih (z ujemanjem treh pregledovalcev in potrditvijo končnega razsojevalca oz. kuratorja). 2.528 koncev povedi so pregledovalci pripisali ročno: pri 2.151 koncih so se strinjali vsi pregledovalci (in kurator), pri 275 po dva, pri 156 pa je konec povedi označil le en pregledovalec. 2.992 koncev povedi je bilo potrjenih kot neustreznih; od tega jih je bilo 1.409 označenih avtomatsko, 940 ročno s popolnim ujemanjem med tremi pregledovalci, 167 ročno z ujemanjem dveh pregledovalcev, 476 ročno z oznako le enega pregledovalca. Pri večini primerov, v katerih je razsojevalec zavrnil odločitve pregledovalcev, gre za popravke tokenizacije in lem, ko so pregledovalci npr. kot konec stavka označili piko, ki je v resnici del okrajšave ("d.o.o", "." > "d.o.o.").

Na popravljenem in ustrezno segmentiranem korpusu smo leme in oblikoskladenjske oznake označili z označevalnikom CLASSLA StanfordNLP v0.0.11.[6]

3.2. Priprava smernic za označevanje

Pri pregledovanju oznak smo sledili smernicam za oblikoskladenjsko označevanje JOS (Holozan et al., 2008), ki vključujejo nabor oblikoskladenjskih oznak (MSD), splošna načela lematizacije ter natančnejše opredelitve posameznih označevalnih kategorij in podkategorij, ponazorjene z označenimi korpusnimi primeri. Smernice smo pripravili v okolju *Google Dokumenti* (ang. *Google Docs*), da smo jih lahko dopolnjevali na podlagi sprotne obravnave ključnih označevalskih dilem ter ponovnega pregleda in evalviranja problematičnih mest. Predvsem na te vidike smernic se bomo osredotočili tudi v nadaljevanju.

4. Pregledovanje oznak

4.1. Obseg in delotoki označevalne kampanje

Ročni pregled strojno označenega gradiva je potekal v okolju *Google Preglednice*. Podatki iz 837 besedil so bili pripravljeni v prav toliko datotekah. Vsaka datoteka je vsebovala metapodatke in za pregledovanje relevantne informacije: obliko pojavnice, lemo, strojno pripisano oblikoskladenjsko oznako (z možnostjo izbire popravka s spustnega seznama vseh obstoječih oznak, kar je olajšalo popravljanje in zmanjšalo možnost zatipka) in celico za morebiten komentar pregledovalca.

Podatke je pregledovalo 24 študentov jezikoslovnih smeri, razdeljenih v 3 skupine. Vsaka izmed teh skupin študentov je pregledovala iste datoteke; namen tega, da vsako pojavnico pregledajo 3 študenti, je bil doseči večjo zanesljivost odločitev. Vsakemu izmed 8 pregledovalcev v skupini je bila dodeljena besedna vrsta oz. več besednih vrst, pri čemer je dodelitev potekala na osnovi preferenc študentov, predhodno ugotovljenih v anketi. Glede na težavnost označevanja ter pogostost vsake besedne vrste v korpusu sta samostalnik pregledovala dva študenta; glagol, pridevnik in zaimek po en študent; za izbiro oznake preprostejše besedne vrste pa smo združili v skupine, pri čemer je en študent pregledoval po eno skupino: prislov in členek; predlog in veznik; številnik, okrajšava, medmet in "neuvščeno". Pred pričetkom pregledovanja so bile pregledovalcem predstavljene smernice (gl. 3.2) in demonstracija postopka v obliki videa. Pregledovanje je potekalo v dveh fazah.

4.1.1. Pregledovanje

Uvodni teden pregledovanja je bil namenjen poglobljeni seznanitvi s smernicami in razreševanju potencialnih nejasnosti, zato je bilo vsakemu pregledovalcu dodeljenih le 5 datotek. Število datotek se je postopoma zviševalo do 20 tedensko, hkrati pa smo okretnejšim ali bolj časovno razpoložljivim pregledovalcem omogočili večji obseg dela (individualizirani pristop). Analiza (ne)ujemanja med tremi vzporednimi pregledovalci je predstavljala izhodišče za 2. fazo – kuracijo.

[3] Označevalni sistem je opisan na spletni strani: http://nl.ijs.si/ME/V6/msd/.
[4] https://github.com/clarinsi/classla
[5] https://github.com/clarinsi/obeliks
[6] https://pypi.org/project/classla/0.0.11/

4.1.2. Kuracija

Posamezne odločitve pregledovalcev smo uredili v enotno tabelo, da so bile ob pojavnicah prikazane vse 3 odločitve, pri čemer so bile posebej označene tiste oznake, pri katerih je med pregledovalci prišlo do razhajanja. Naloga kuratorjev je bila pregledati prav te pojavnice in jim pripisati končno oznako. 7 kuratorjev je bilo izbranih iz vrst pregledovalcev, po eden za vsako besedno vrsto oz. skupino besednih vrst. Označevalna kampanja se je zaključila v 12 tednih, od katerih so bili štirje namenjeni samo kuraciji.

4.2. Označevalne dileme

Ob kuraciji smo identificirali dve vrsti označevalnih težav: (a) primeri, pri katerih so bile označevalne smernice jasne, a pri delu niso bile dosledno upoštevane in (b) primeri, ki so se pokazali kot zahtevnejši: slabše predstavljeni v smernicah in mestoma tudi nedosledno obravnavani v obstoječem ssj500k 2.3.[7]

Težave prvega tipa smo analizirali, odpravili nekonsistentnosti in jih označili v skladu z označevalnimi smernicami. Nekaj več informacij o tipičnih tovrstnih težavah povzemamo v poglavju 4.3. Posebno pozornost pa smo posvetili drugi skupini težav, ki smo jih identificirali kot bolj kompleksne in zahtevne, saj so njihove rešitve zahtevale premislek o odprtih vprašanjih na ravni lematizacije in oblikoskladnje (tudi) v korpusu ssj500k in posledično nadgradnjo označevalnih smernic. V nadaljevanju predstavimo te težave, v poglavju 5 pa predlagane spremembe smernic.

4.2.1. Občnoimenska prekrivnost v stvarnih lastnih imenih

Pregledovalcem je težave povzročalo pravilo, da je v stvarnih lastnih imenih, kjer je lastnoimenski samostalnik prekriven z občnoimenskim samostalnikom, tako lema kot oblikoskladenjska oznaka občnoimenska. Iz tega sledi, da je lematizacija slovenskih imen podjetij, časopisov, revij, knjig, tudi televizijskih oddaj, serij ali filmov ipd. z malo začetnico: npr. podjetje ***Iskra*** [iskra, Sozei]; časnik ***Delo*** [delo, Sosei]. Na iskanje prekrivnosti, ki zaradi pomenske oddaljenosti občnoimenske "ustreznice" pogosto ni enoznačno (gl. tudi 4.2.3), je bilo treba večkrat opozoriti, saj je bilo pregledovalcem bolj intuitivno ohraniti zapis leme z veliko začetnico. Opozarjati jih je bilo treba tudi, da načelo prekrivnosti dogovorno velja samo pri samostalnikih (stranka ***Zares*** [Zares, Slmei]). Manj težav smo zaznali pri pregledovanju tistih primerov stvarnih lastnih imen, ki niso imela prekrivne leme z občnim samostalnikom in smo jih lematizirali z veliko začetnico (podjetje ***Mercator*** [Mercator, Slmei]).

4.2.2. Izlastnoimenski svojilni pridevniki

Del pravila, da pri svojilnih pridevnikih, ki izvirajo iz osebnih ali zemljepisnih lastnih imen, ohranjamo lemo z veliko začetnico (***Aškerčeva*** *ulica* > lema: Aškerčev), je bil jasen, več dilem je bilo pri pregledovanju tistih izlastnoimenskih svojilnih pridevnikov, ki se v rabi pišejo z malo ali pa prehajajo v zapis z malo, ker niso v pomenu prave svojine (***Parkinsonova*** *bolezen* > lema: parkinsonov).

Pri lematizaciji izlastnoimenskih pridevnikov v stvarnih lastnih imenih je pregledovalce zmedla različna obravnava primerov v korpusu ssj500k (***Delova*** *dopisnica* > lema: Delov vs. ***Magov*** *novinar* > lema: magov), zato je bilo treba ta del pravila, ki v izhodiščnih smernicah ni bil pojasnjen, posebej razložiti.

4.2.3. Tuja stvarna lastna imena

Ker načelo prekrivnosti z občnoimenskimi samostalniki (gl. 4.2.1) velja primarno za slovenske samostalnike, se je pogosto pojavljalo vprašanje, katere besede obravnavati kot slovenske (prevzete besede, ki se pregibajo s slovensko morfologijo, vedno umeščamo med slovenske, če potrditve za pregibanje v rabi ni, pa se je treba odločiti na podlagi drugih kriterijev). Dileme so se nanašale predvsem na: (a) prevzete besede, ki pogosto nastopajo kot deli tujejezičnih imen sicer slovenskih podjetij (tip *leasing, holding*) ter (b) ostale občnoimenske besede v tujejezičnih zvezah, ki so prekrivne s slovenskimi občnoimenskimi samostalniki, pri čemer pa pogosto ne izpolnjujejo kriterija pomenske prekrivnosti (tip *trans, global*).

Podrobneje smo obravnavali tudi skupino stvarnih lastnih imen tipa *Zagrebačka banka*, *Večernji list*. Ker gre za imena v hrvaščini, ki zaradi sorodnosti s slovenščino mestoma prinašajo besedje, enako slovenskemu, so bili pregledovalci v dilemi, ali tako pridevnik kot samostalnik označiti kot slovensko besedo in pri tem pridevniku pripisati v slovenščini neobstoječo lemo, ali (vsaj) pridevnik umestiti med tujejezično besedišče.

4.2.4. Ločevanje pridevnikov od prislovov

Odločitve pregledovalcev so se pogosto razhajale pri primerih, ki so izkazovali tipično povedkovnodoločilno rabo pridevnikov oz. obravnavo pridevniških oblik, ki so se prekrivale z osnovno prislovno obliko. Smernice so že vsebovale splošno navodilo o označevanju pridevnikov, ki lahko nastopajo v prilastkovi ali povedkovi rabi (*Sledil je* ***prelomni*** *korak* > pridevnik kot levi prilastek; *uradno še ni* ***rehabilitiran*** > pridevnik kot povedkovo določilo), pa tudi pravilo za ločevanje pridevnikov od prislovov v primeru pridevniškega niza (***uradno*** *prečiščeno besedilo* > prislov). Niso pa naslovile razlike med pridevniško in prislovno lemo pri posameznih zahtevnejših primerih (npr. *smotrno, potrebno, mogoče, možno* v primerih kot npr. *bi bilo* ***smotrno****, da bi [...]*), ki so se tudi v korpusu ssj500k pokazali kot nekonsistentno označeni: pogosto smo zasledili prislovno lemo namesto dogovorno ustrezne pridevniške leme. Neskladja so predstavljala izhodišče za nadaljnje analize, ki so vključevale ponovni pregled vseh primerov oz. zgledov (v korpusu SentiCoref) s prekrivnimi pridevniškimi in prislovnimi oblikami ter oblikovanje dopolnjenega pravila za pripisovanje pridevniških in prislovnih lem.

4.2.5. Nesklonljivi prilastki (tip *bruto, solo*)

Pregledovalci so imeli težave z razumevanjem navodila v izhodiščnih smernicah, da tiste primere tipa bruto, solo (npr. ***solo*** *uspeh*, *rast* ***bruto*** *zadolževanja*, ***info*** *točka*), ki so sklonljivi, obravnavamo kot samostalnike, tiste, ki niso, pa kot pridevnike. Predvsem v navodilu ni jasno, kako preverjati (ne)sklonljivost in kaj je vodilo za odločitev (sistemska možnost, gradivo).

[7] Smernice Holozan et al. (2008) predstavljajo v slovenskem prostoru sprejet in široko apliciran označevalni standard, zato smo jim sledili v največji možni meri. Tudi dopolnitev smernic, ki smo jo pripravili na projektu RSDO, ostaja v zastavljenih konceptualnih okvirih. Morebitne korenitejše spremembe označevalnega sistema, kjer izstopa predvsem vprašanje lematiziranja (pravopisno, ne pa tudi oblikoslovno) različnih samostalnikov in tudi drugih besednih vrst z veliko ali malo začetnico, zahtevajo širši premislek, ki ga nakažemo v pogl. 6.

4.2.6. Prislovne zveze (tip *na novo*)

Težave so bile tudi z obravnavo t. i. prislovnih zvez oz. označevanjem nepredložnega dela teh zvez. Smernice posredno nakazujejo, naj označevanje teži k pridevniškim lemam (*na **drobno*** > lema: droben), se je pa recimo pri primeru *v živo* pokazalo, da so bili v korpusu ssj500k vsi takšni primeri označeni kot prislovni (*v **živo*** > lema: živo). Na osnovi tega neskladja smo naredili podrobnejšo analizo in odkrili več primerov neenotnega označevanja enakovrstnih primerov.

4.3. Pregledani podatki

Analiza popravkov po koncu pregledovanja in kuriranja kaže, da se delež vnesenih popravkov sklada s pričakovanim deležem napak pri avtomatskem označevanju slovenskih besedil z označevalnikom CLASSLA StanfordNLP (Ljubešič in Dobrovoljc 2019: 31–32). Na ravni lematizacije je bilo skupaj popravljenih 5.588 lem, kar je približno 1,3 % vseh pojavnic v korpusu, kar se sklada s približno 98-odstotno natančnostjo lematizacije. Na ravni oblikoskladenjskih oznak je bilo skupaj 12.586 popravkov, kar pomeni 2,9 % vseh oznak v korpusu (ob skoraj 97-odstotni natančnosti oblikoskladenjskega označevanja).

Pri popravkih lem so bili med najpogostejšimi lastnoimenskimi samostalniki, ki so prekrivni z občnoimenskimi (npr. ***Luka** Koper* > lema: luka), okrajšave, sestavljene iz ene ali dveh črk (npr. *dr.* > lema: dr.), pa tudi besede s prekrivnimi oblikami v oblikoskladenjski paradigmi (npr. *delo* in *del*). Pri popravkih oblikoskladenjskih oznak je šlo povečini za ločevanje med občnimi in lastnoimenskimi samostalniki (tip *Leasing – leasing*; 1538 popravkov oz. 12 %; v obratni smeri iz občnoimenskega v lastno je bilo popravkov manj: 235 oz. 1,8 %), med moškim in ženskim spolom (825 popravkov oz. 6,6 %; pri tem gre npr. za imena določenih strank, kot je *Desus*) ter med prekrivnimi oblikami v imenovalniku, tožilniku in rodilniku (skupaj 1.617 popravkov oz. 12,8 % pri samostalnikih; npr. neživi samostalniki moškega spola: *odbor*, *posel* v imenovalniku in tožilniku). Na ravni besednih vrst je šlo največkrat za težje ločevanje med prekrivnimi prislovi in prirednimi vezniki (npr. *tako*; 130 popravkov oz. 1,1 %), med lastnoimenskimi samostalniki in neuvrščenimi tujejezičnimi izrazi (npr. *Amnesty International*; 118 popravkov oz. 1,0 %) ter med členki in prirednimi vezniki (npr. *sicer*, *niti*, *ne*; 97 popravkov oz. 0,7 %). Ker je bila količina popravkov relativno majhna, bi se bilo v prihodnjih označevalnih kampanjah morda smiselno osredotočiti le na najpogostejše pričakovane napake. Kot vodilo lahko pri tem služijo v tem poglavju naštete najpogostejše dileme in težave.

5. Nadgradnja označevalnih smernic

Na podlagi analize najpogostejših označevalskih dilem in pregleda označevalnih odločitev v korpusu ssj500k smo pripravili rešitve glede (nadaljnjega) pregledovanja in dopolnitve smernic za problematične kategorije, naštete v poglavju 4.2. Nadgrajene smernice bodo objavljene ob koncu projekta RSDO.

I. **Občnoimenska prekrivnost v stvarnih lastnih imenih:** splošno načelo, da stvarna imena, prekrivna z občnim samostalnikom, označujemo kot občni samostalnik in lematiziramo z malo začetnico, ostala, ki prekrivnosti ne izkazujejo, pa z veliko začetnico, smo dopolnili s konkretnimi zgledi rabe. Izbrali smo kategorije, ki so povzročale največ težav (podjetja in časnike), npr. *O tem, da so bile v **Iskri*** [iskra, Somem] *potrebne spremembe, so čivkali že vrabci na veji.*; *Večino hrane kupimo v **Mercatorju*** [Mercator, Slmem] *ali **Intersparu*** [Interspar, Slmem].; *Kot smo poročali v prejšnji številki **Mladine*** [mladina, Sozer].

II. **Izlastnoimenski svojilni pridevniki:** v smernice smo dodali pravila za rabo velike in male začetnice s primeri:

(a) Pridevniki iz osebnih in zemljepisnih lastnih imen: načeloma ohranjamo lemo z veliko začetnico, tiste primere, ki se v rabi pišejo z malo ali so na prehodu v zapis z malo, ker niso v pomenu prave svojine, pa lematiziramo z malo, npr. *Celjska občina je prejšnji teden objavila razpis za najem vile v **Aškerčevi*** [Aškerčev, Psnzem] *7 v Celju.*; *Gre za zdravilo za zdravljenje **parkinsonove*** [parkinsonov, Psnzer] *bolezni.*

(b) Pridevniki iz stvarnih lastnih imen: dodatno smo opredelili načelo lematizacije primerov tipa ***Delova** dopisnica* > lema: Delov in ***Magov** novinar* > lema: magov. Pri primerih, kjer je bila prekrivnost sistemsko sicer možna, vendar v dejanski rabi neizkazana, smo ohranili veliko začetnico, npr. *S tega stališča je polemika z **Mladininim*** [Mladinin, Psnmeo] *doktorjem sociologije že skorajda na robu smiselnega* (občni samostalnik *mladina* sicer obstaja, vendar je svojilni pridevnik v rabi izredno redek, tj. ima eno samo pojavitev v referenčnem korpusu Gigafida 2.0). Nasprotno je v primerih, ki izkazujejo pogostejšo rabo svojilnega pridevnika, npr. *vsi pa občudujejo njegovi operi Jevgenij Onjegin in **Pikova*** [pikov, Psnzei] *dama.*

(c) Pridevniki na -ski, -ški kot del zemljepisnih lastnih imen: lematiziramo jih z malo začetnico, pri čemer je treba posebej izpostaviti razliko v odnosu do primerov tipa *Kranjska*, *Štajerska* ipd. Pri imenih regij gre za samostalnike in jih lematiziramo z veliko: *V Vinski kleti **Goriška*** [goriški, Ppnsmi] *Brda zadovoljni s poslovanjem v minulem letu*; *Črnivec je poleg prelaza Volovjek najsevernejši cestni prehod, ki povezuje **Kranjsko*** [Kranjska, Slzet] *in **Štajersko*** [Štajerska, Slzet].

(č) Splošni pridevniki kot del zemljepisnih lastnih imen: lematiziramo jih z malo (tip *nov*, *spodnji*), če v splošni rabi ne obstajajo, pa ohranimo veliko začetnico, npr. *Britanija, Avstralija in **Nova*** [nov, Ppnzei] *Zelandija*; *Mlekarna Celeia iz **Arje*** [Arji, Ppnzer] *vasi je namreč edina domača mlekarna v večinski lasti zadrug.*

III. **Tuja stvarna lastna imena:** po posvetu s širšo projektno ekipo smo se odločili, da bomo oblikovno prekrivne občnoimenske samostalnike "iskali" tudi v tujejezičnih večbesednih stvarnih lastnih imenih. Pri tem je treba upoštevali predvsem dve merili: pregibanje v rabi in prevzetost (prisotnost v referenčnih priročnikih, npr. *Hypo **Leasing*** [leasing, Somei], *Infond **Holding*** [holding, Somei]), ne pa nujno tudi merilo pomenske prekrivnosti – v nekaterih primerih lahko ima tuja beseda v stvarnem lastnem imenu podoben pomen, kot ga ima (prekrivna) slovenska beseda, v nekaterih pa ne. Primere, ki so bili oblikovno prekrivni, pomensko pa ne, smo zbrali na posebnem seznamu in po analizi sprejeli odločitev, da jih vse obravnavamo kot občne samostalnike, npr. *Trade **Trans*** [trans, Somei] *Invest*, *Prevent **Global*** [global, Somei].

V smernice smo dodali odločitev glede označevanja slovenščini sorodnih tujih primerov (tip *Večernji list*): pri

samostalnikih upoštevamo načelo prekrivnosti s slovenskim občnoimenskim besediščem, pridevnike obravnavamo kot tuje besedišče, pri katerem ostane lema enaka besedni obliki, npr. *Jutarnji* [Jutarnji, Nj] *list* [list, Somei], *Zagrebačka* [Zagrebačka, Nj] *banka* [banka, Sozei].

IV. **Ločevanje pridevnikov od prislovov:** pri opredelitvi razlike med pridevnikom in prislovom v vlogi povedkovega določila smo v smernicah izpostavili skladenjski vidik. Opredelitev razlike, da je beseda v vlogi povedkovega določila prislov, če je iz stavka izpustljiva, pridevnik pa, če je nepogrešljiva (obvezna), smo podkrepili s primeri, npr. *O tem ni ***(mogoče)*** [mogoč, Ppnsei] *sklepati.*; ***(Mogoče)** [mogoče, Rsn] ste ga vznemirili.*

V. **Nesklonljivi prilastki (tip *bruto, solo*):** pri obravnavi nesklonljivih prilastkov se je smiselno opreti na preverjanje njihove sklonljivosti v dejanski rabi. Oblikovali smo pravilo, da če najdemo potrditev v referenčnem korpusu, da se določen primer lahko pregiba kot samostalnik, potem to opcijo upoštevamo, če pa potrditve ne najdemo, primer dosledno obravnavamo kot pridevnik: *so se do konca leta povprečne **neto*** [neto, Ppnmein] *plače realno povečale za okoli 33 odstotkov.*

VI. **Prislovne zveze (tip *na novo*):** v smernice smo dodali eksplicitno pravilo, da primere tega tipa obravnavamo kot zveze predloga in pridevnika. Na primerih, ki so pregledovalcem predstavljali največ težav, smo ponazorili, da obravnavamo nepredložni del zveze torej kot pridevnik in ne kot prislov, npr. *Če bi se **na*** [na, Dt] ***hitro*** [hiter, Ppnset] *ozrl, bi videl, da ga zasledujejo.*

6. Zaključek in nadaljnje delo

Pregledovanje osnovnih označevalnih nivojev korpusa SentiCoref predstavlja eno najobsežnejših kampanj te vrste v našem prostoru in – ob kampanji, ki se je osredotočala na gradivo računalniško posredovane komunikacije Janes (Čibej et al., 2018) – tudi eno prvih priložnosti za ponovitev dela z uporabo metodologije, ki se je vzpostavila pri pripravi izhodiščne različice učnega korpusa.

Po opravljeni kuraciji, končni kontroli kvalitete označenega in statističnem pregledu dilem in popravkov je mogoče potegniti nekaj zaključkov. Pomembno je, da so se pomanjkljivosti označevalnih smernic kazale zlasti pri temah, povezanih z označevanjem lastnih imen (samostalnikov, izlastnoimenskih pridevnikov), še zlasti pri odločitvah, ki so povezane s presojanjem, ali je določena beseda slovenska ali tujejezična. Ker korpus SentiCoref vsebuje atipično visoko število raznovrstnih lastnih imen (tako je bil namreč zgrajen), smo pogosto srečevali težave, ki so bile pri pripravi ssj500k redkejše in za smernice manj relevantne.

Obstoj kategorije lastnoimenskosti na ravni oblikoskladnje in posledično lematiziranje ob iskanju prekrivnosti med občno- in lastnoimenskimi entitetami odpira konceptualne težave, ki bi jih kazalo v ponovno premisliti. Prva je, da je označevalno kategorijo najti samo pri samostalnikih, prekrivnost (po nekako drugačni logiki) iščemo tudi pri pridevnikih, ne pa pri ostalih besednih vrstah. Težava je tudi, da pri odločitvah glede zapisa leme z veliko ali malo začetnico na raven oblikoskladenjskega označevanja prenašamo vprašanja, ki se dotikajo pravopisa (oz. pravopisov, če upoštevamo, da se vse dileme preslikavajo in potencirajo pri srečevanju s tujejezičnimi elementi), pri čemer sistem sledi predpostavki, da avtorji besedil pravopisu vedno sledijo. Pri razlikovanju obravnave zemljepisnih/osebnih imen ter stvarnih imen se v sistem še dodatno vpletajo načela, ki so bolj kot na oblikoskladnjo vezane na (s semantiko in trenutnim pravopisom povezano) metajezikovno klasifikacijo referentov. Zdi se, da na ravni označevanja lem in oblikoskladnje sprejemamo odločitve, ki bi sodile na raven jezikovnega opisa in predpisa, ob čemer se opiramo na jezikovne vire, kjer prav te odločitve pogosto še niso sprejete.

Ker težave pripisovanja občno- oz. lastnoimenskosti samostalnikov prednjačijo v veliki sliki vseh opravljenih popravkov, obenem pa identifikacijo lastnoimenskih zvez v zadnjih letih uspešno opravljamo pri označevanju imenskih entitet, bi kazalo ponovno razmisliti o dodani vrednosti te kategorije na ravni oblikoskladnje. Če se izkaže, da je kljub vsemu koristna, bi se določene težave dalo odpraviti z radikalnejšim posegom v smernice, npr. z odpovedjo iskanja prekrivnih občnih in lastnih samostalnikov in sledenju rabi, kakršna se v besedilih pojavlja. Enako velja za obravnavo tujega besedišča, ki ga po trenutnem sistemu med slovenske samostalnike umeščamo precej popustljivo in obenem nedosledno. S širjenjem označevanja na besedilne vrste, kjer je tujejezičnih elementov več in v slovenščino prehajajo po manj predvidljivih vzorcih, bi bilo smiselno opredeliti jasen namen ločevanja po jezikih in oblikovati dosledne in pripisljive kriterije zanj. Problem bi bilo dobro nasloviti celovito in podati rešitve za vse relevantne označevalne ravni, ne le lematizacijo in oblikoskladnjo.

Druga večja skupina označevalnih težav je bila vezana na enakopisne oblike, pogosto pridevnike in prislove, pa tudi nekatere slovnične besedne vrste. Tudi tu je opaziti, da se v smernicah pojavljajo semantični (ne le oblikoslovni in skladenjski) kriteriji za presojanje, kar pa se je izkazalo za manj pereče od (po novem vsaj deloma naslovljenih, ne pa povsem odpravljenih) dilem glede uporabe referenčnih jezikovnih virov, npr. za opredeljevanje sklonljivosti. Pri tej skupini težav je ključna ugotovitev, da označevanje tudi v ssj500k ni potekalo povsem usklajeno, zato smo ob delu pripravili seznam težav, ki bi jih bilo v prihodnosti smiselno preveriti in ustrezno urediti za nazaj.

Pri vsem pa je treba upoštevati, da je strojno pripisovanje lem in oblikoskladenjskih oznak za slovenščino že doseglo raven, ko bi bilo celovite ročne preglede smotrno nadomestiti z delnimi, za katere pa bi bilo treba razviti (referenčne in dokumentirane) postopke za avtomatsko ali polavtomatsko identifikacijo problematičnih mest. Spoznanja, ki jih navajamo v prispevku, so lahko izhodišče za takšno nadaljevanje.

Pregledani in popravljeni SentiCoref bo v nadaljevanju projekta RSDO umeščen ob ostale besedilne množice, ki bodo sestavljale povečani učni korpus za slovenščino. V prihodnje bomo v celotnem učnem korpusu izvedli še serijo polavtomatskih popravkov (npr. ali so enobesedni vezniki, kot je "zato", vedno ustrezno označeni kot vezniki), s čimer bomo poskrbeli, da bodo enake dileme v celotnem učnem korpusu razrešene konsistentno. Na podoben način bomo učni korpus primerjali tudi s Slovenskim oblikoslovnim leksikonom Sloleks (Dobrovoljc et al., 2019), da npr. preverimo, ali se glagolski vid glagolov v učnem korpusu ujema s Sloleksom. V okviru projekta RSDO je istočasno z nadgradnjo učnega korpusa potekala tudi nadgradnja Sloleksa, zato smo nalogo prestavili na poznejši termin.

Učni korpus bo skupaj z nadgrajenimi označevalnimi smernicami in ostalo dokumentacijo ob zaključku projekta javnosti odprto na voljo na repozitoriju CLARIN.SI.

7. Zahvala

Projekt *Razvoj slovenščine v digitalnem okolju* (RSDO) sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru *Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020*. Raziskovalna programa št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik) in št. P6-0215 (*Slovenski jezik - bazične, kontrastivne in aplikativne raziskave*) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Avtorice in avtorji se sodelujočim v označevalni kampanji iskreno zahvaljujemo za vse delo, prav tako pa tudi recenzentoma za relevantne in konstruktivne komentarje.

8. Literatura

Špela Arhar Holdt in Jaka Čibej. 2021. Analize za nadgradnjo učnega korpusa ssj500k. V: Š. A. Holdt, ur., *Nova slovnica sodobne standardne slovenščine: viri in metode*, str. 15–53. Znanstvena založba Filozofske fakultete, Ljubljana. Zbirka Sporazumevanje. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/325/477/7313-1.

Primož Belej. 2018. *Oblikoskladenjsko označevanje slovenskega jezika z globokimi nevronskimi mrežami*. Magistrsko delo, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

Jože Bučar. 2017. *Manually sentiment annotated Slovenian news corpus SentiNews 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1110.

Jaka Čibej, Špela Arhar Holdt, Tomaž Erjavec in Darja Fišer. 2018. Ročno označeni korpusi Janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave. V: D. Fišer, ur., *Viri, orodja in metode za analizo spletne slovenščine*, str. 44–73. Znanstvena založba Filozofske fakultete, Ljubljana. Zbirka Prevodoslovje in uporabno jezikoslovje. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/111/203/2416-1.

Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevič in Dan Tufis. 1998. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern European languages. V: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, zvezek 1, str. 315–319, Montreal, Quebec, Kanada. Association for Computational Linguistics. https://aclanthology.org/P98-1050.pdf.

Kaja Dobrovoljc, Simon Krek in Tomaž Erjavec. 2015. Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 80−105. Znanstvena založba Filozofske fakultete, Ljubljana. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/15/47/489-1.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Jaka Čibej, Luka Krsnik in Marko Robnik-Šikonja. 2019. *Morphological lexicon Sloleks 2.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1230.

Richard Eckart de Castilho, Chris Biemann, Irina Gurevych in Seid Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. V: *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Nizozemska. https://www.clarin.eu/sites/default/files/cac2014_submission_6_0.pdf.

Tomaž Erjavec in Simon Krek. 2008. The JOS morphosyntactically tagged corpus of Slovene. V: *Proceedings. 6th International Conference on Language Resources and Evaluation (LREC 2008)*, str. 322–327, Marakeš, Maroko. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/89_paper.pdf.

Tomaž Erjavec. 2012. MULTEXT-East: Morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik (Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene). V: *Proceedings of the 8th Language Technologies Conference*, zvezek C, str. 89–94, Ljubljana, Slovenija. IJS. http://nl.ijs.si/isjt12/proceedings/isjt2012_17.pdf.

Peter Holozan, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman in Aleš Velušček. 2008. *Specifikacije za učni korpus*. Projekt "Sporazumevanje v slovenskem jeziku". http://projekt.slovenscina.eu/Vsebine/Sl/Kazalniki/K2.aspx.

Jan-Christoph Klie, Michael Bugert, Beto Bullosa, Richard Eckart de Castilho in Irina Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. V: *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, ZDA. https://aclanthology.org/C18-2002.pdf.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek in Anja Zajc. 2021. *Training corpus ssj500k 2.3*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1434.

Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Polona Gantar, Špela Arhar Holdt, Jaka Čibej in Janez Brank. 2020. The ssj500k Training Corpus for Slovene Language Processing. V: D. Fišer in T. Erjavec, ur., *Jezikovne tehnologije in digitalna humanistika: zbornik konference*, str. 24–33, Ljubljana, Slovenija. Inštitut za novejšo zgodovino. http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene--Language-Processing.pdf.

Nikola Ljubešić in Kaja Dobrovoljc. 2019. 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. Firence, Italija. The Association for Computational Linguistics, Stroudsburg. https://www.aclweb.org/anthology/W19-3704.

Nikola Ljubešić in Tomaž Erjavec. 2016. Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. V: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, str. 1527–1532, Pariz,

Francija. European Language Resources Association (ELRA). https://aclanthology.org/L16-1242.pdf.

Federico Martelli, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, in Tina Munda. 2022. *Parallel sense-annotated corpus ELEXIS-WSD 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1674.

Slavko Žitnik. 2019. *Slovene corpus for aspect-based sentiment analysis - SentiCoref 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1285.

Document Enrichment as a Tool for Automated Interview Coding

Ajda Pretnar Žagar,* Nikola Đukić,* Rajko Muršič‡

*Laboratory for Bioinformatics
Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, SI-1000 Ljubljana
ajda.pretnar@fri.uni-lj.si, nd1776@student.uni-lj.si

†Department of Ethnology and Cultural Anthropology
Faculty of Arts
University of Ljubljana
Zavetiška ulica 5, SI-1000 Ljubljana
rajko.mursic@ff.uni-lj.si

Abstract

While widely used in social sciences and the humanities, qualitative data coding remains a predominantly manual task. With the proliferation of semantic analysis techniques, such as keyword extraction and ontology enrichment, researchers could use existing taxonomies and systematics to automatically label text passages with semantic labels. We propose and test an analytical pipeline for automated interview coding in anthropology, using two existing taxonomies, Outline of Cultural Materials and ETSEO systematics. We show it is possible to quickly, efficiently and automatically annotate text passages with meaningful labels using current state-of-the-art semantic analysis techniques.

1. Introduction

Qualitative data coding is a well-established procedure in social sciences, particularly in sociology, cultural studies, oral history, and biographic studies. The technique is gaining ground in anthropology, where interview transcriptions abound. Ethnographic text coding can become a serious research technique, using existing ethnographic systematics, categories, vocabularies, and codes. Data coding facilitates the analysis of themes and close reading of the interview segments on each theme, which is one of the main analytical techniques of ethnographies in anthropology, be they computer-assisted or manual.

Computer-assisted qualitative data analysis (CAQDAS) is used to determine topics of interview segments, where the topics are not discrete but can overlap. The coder would normally define a codebook with the topics, then go over the text and label passages with corresponding tags. In the end, the coder can review selected topical passages, define topic co-occurrence, and extract a subset of documents on a specific topic.

Manual labelling can take a long time and requires a somewhat experienced coder to handle the tagging. However, we can construct an automatic pipeline for segment tagging due to the rapid development of natural language processing tools and language resources. The pipeline is built on the recent developments in ontology enrichment, which uses pre-defined ontologies (or taxonomies). Documents are preprocessed, and then the resulting tokens, typically words, are compared by similarity to tokens from the ontology. A simple approach is based on TF-IDF[1] transform. In contrast, the current state of the art uses graph models (i.e., YAKE) and word embeddings (Godec et al., 2021) for determining concept similarity.

Qualitative data coding is often based on grounded theory (Strauss and Corbin, 1997). The theory, which is more of an analytical approach, focuses on codes to emerge from the data (Holmes and Castañeda, 2014) rather than imposing them. Coding can also stem from a linguistic paradigm, especially semantic approaches, where text would be labelled based on the occurrence of words in it. The first approach still requires human input, while the second is based on unsupervised machine learning. Thus, having a general ethnographic taxonomy or classification scheme enables researchers to inductively elicit prevalent topics from the data rather than devising elaborate codebooks in advance. Our contribution is applying semantic annotation and ontology mapping to interview transcripts.

Semantic enrichment of documents means assigning conceptually relevant terms to documents or document segments. The procedure can include automatic keyword extraction, which identifies relevant keywords in the text (Bougouin et al., 2013; Campos et al., 2020) or relating existing lists of terms to texts (Massri et al., 2019). The latter can be either unsupervised or supervised. Unsupervised refers to the terms being scored by their similarity to the text and (multiple) terms assigned to each document if their similarity to the document is above a certain threshold. Supervised means the terms are used for document classification, where a document is assigned the most probable term.

In continuation, we propose a technique using unsuper-

[1]TF-IDF is a document vectorisation technique which uses word counts to describe documents, weighing them based on overall word frequency.

vised ontology enrichment to automatically label text segments with the corresponding topic labels. Automatic segment labelling uses existing (anthropological) taxonomies to label interview segments and thus assist researchers in navigating interview transcripts. The proposed technique doesn't apply only to anthropology – it could be used in any text analysis research. We use anthropology as a use case since the use of computer-assisted techniques is still somewhat rare in this discipline.

Finally, a short note on terminology. The term "ontology" is used in computer science to describe a structured hierarchical list of terms (Gruber, 1995), while in social sciences and the humanities, it means a branch of philosophy studying concepts of existence. In this paper, we use the term ontology in the former sense, sometimes referring to it as a taxonomy for clarity.

2. Interview transcripts

Interview transcripts are specific since they contain questions from the interviewer and answers from the interviewee. The transcripts are usually structured, with names or abbreviations denoting the speaker. If the interview is (semi-)structured, questions between different interviews will be very similar, if not identical. Moreover, interviewing a person often requires the interviewer to ask for clarification, affirm the interpretation of the answer or simply confirm (s)he understood what the interviewee said. Hence including questions in the analysis is often not a good approach.

Delineating between questions and answers depends on the structure of the digital document. A dedicated parser would consider new lines as segment delineations and names, pseudonyms, or initials as speaker identifiers. Ideally, the parser would consider the continuation of a reply, even when it was interrupted by the interviewer. But without a proper co-reference resolution for the given language (Žitnik and Bajec, 2018), it is difficult to determine such conceptual segments.

3. Related work

Back in 1983, Podolefsky and McCarty (1983) had an interesting idea - how about using computers to help us navigate numerous ethnographic notes and transcripts? Those were the days when most anthropologists stored their data on physical paper. Navigating such texts apparently required duplicating pages to store them under various categories. Nowadays, this is no longer necessary. Ethnographic data is often multimodal and predominantly stored digitally. It includes images, videos, and audio recordings along with the text. When navigating digital text data, one can easily use the "find" function to look for different text segments, while similar techniques exist for navigating other data types.

Nevertheless, organising interview data is not an easy task, and there are ways computers can help. Podolefsky and McCarty (1983) proposed developing coding categories for marking text passages. This is the precursor to modern qualitative data analysis software, such as NVivo, Atlas.ti, or MaxQDA. These, too, require a predetermined set of categories used for labelling the data.

Modern computer-assisted qualitative data analysis (CAQDAS) approaches don't require using punched cards with per-page summaries to navigate the text, as was the case in earlier times. They can quickly retrieve segments tagged with the specified tag. MacMillan and Phillip (2010) use a semi-anthropological approach to better gauge the connection between venison price and cull effort. They conduct in-depth interviews with stalkers, people employed by the British estates that hunt wild game, and analyse the interviews with NVivo. They use the qualitative data from the interviews to corroborate the quantitative findings – deer hunting is deeply rooted in tradition and seen as a sport rather than economic activity.

Researchers studying sensorially-charged biographic experiences in Turku, Brighton, and Ljubljana defined the main categories with a larger list of subcategories. Coding only the translated transcripts and using Atlas.ti, they extracted similarly charged testimonies related to different sensations, for example, sounds (Venäläinen et al., 2020) or smells (Bajič, 2020).

Most commonly, CAQDAS is used in discourse analysis. Hitchner et al. (2016) analyse discourses on bio-energy to elicit key metaphors used to create common imaginaries. Using this approach, they were able to identify three discursive units that guide the bio-energy narrative. Cuerrier et al. (2015) identified 134 categories referring to climate change in 46 interviews conducted with the Inuit population in Nunavik. Next, they created ordinal and binary matrices describing the change in quantity and the presence or absence of topics. They used various statistical approaches to determine whether different communities of Nunavik differ in terms of knowledge of climate change. Both papers retrieve popular taxonomies created by people under study.

Discourse analysis is also prominent in Schafer (2011), who uses Atlas.ti to analyse over 30 in-depth interviews with secular funeral officiants called "funeral celebrants" in New Zealand. The author identified key conceptual categories in funeral celebrant ethnographies, specifically the narratives on connection, identity, and personalisation of funeral practices.

CAQDAS can also be used to retrieve relevant text passages. Yilmaz et al. (2019) conducted 30 interviews with highly educated Turkish-Belgian women to determine the factors affecting their marriage choices. They stem from grounded theory and use predetermined codes for the first round of coding, then refine and enhance their codebook later. With iterative codebook improvements, they determined women's decisions and the driving factors behind them, for example, the structural and general constraints in marriage choices.

Conversely, Wehi et al. (2018) do not use CAQDAS software but instead observe raw word frequencies in Māori oral tradition. They collected ancestral sayings called whakataukī and identified references to animal extinctions in the data.

It is interesting to note that many contributions using quali-quantitative text analysis were published in the Human Ecology journal, which testifies to the (still) marginal use of these methods in anthropology. Ideally, we will see many more journals willing to publish such research and

more researchers ready to use these tools in practice.

Longan (2015) expresses the sentiment to perfection: "There is room for innovation in the creation of technological aids to facilitate mesoscale qualitative online research that lies between massive data sets and small qualitative studies. Though the major qualitative software suites have improved over time, much of the process is still tedious and requires hours of snorkelling and coding by hand." First, he explicitly points to the nor-big-nor-small issue of many contemporary anthropological studies. Even organising just thirty interview transcripts can be complicated, let alone a hundred records. Yet one hundred records can hardly be described as "big data" requiring "big tools". There's a need for a mid-level tool to help organise the data in a time-efficient way. Second, he points to the issue of coding by hand, which takes time and effort from the researcher. Third, he identifies an opportunity for technological innovation for qualitative data analysis that surpasses modern qualitative analysis software.

Previously, ontology enrichment for labelling text passages was used predominantly in biology and medicine (Bifis et al., 2021). In social sciences and the humanities, automated segment labelling was expressed as more of a wish rather than a reality (Hoxtell, 2019). In contrast to CAQDAS, ontology enrichment provides a way to automatically label large amounts of text in a short period of time. At the same time it enables relating interview transcripts to existing domain-specific ontologies. Our contribution showcases automated interview segment labelling with existing ontologies, thus providing a practical example of how machine learning can support ethnographic analysis.

We propose an approach using ontology enrichment from computer science to help organise and structure interview transcripts, fieldwork notes, and archive data. The three-fold example described below is a prototype for machine-assisted data coding, which uses standard anthropological taxonomies, such as the Outline of Cultural Materials (Bernard, 1994, p. 519-528), or more local and specific ethnographic taxonomies, related to the European ethnology studies of the so-called folk or traditional culture (Kremenšek et al., 1976), to label text passages.

4. Ontologies as codebooks

Instead of pre-defining codebooks for manual coding, we propose to use existing anthropological taxonomies to automatically label the data. One such well-established taxonomy, which we call "ontology" in text mining, is the Outline of Cultural Materials. Human Relations Area Files is a non-profit research organisation whose aim is to foster cross-cultural research (Melvin, 1997). One of its key achievements is the establishment of several databases that contain previous cross-cultural research. The database entries, such as ethnographic reports, are indexed using the Outline of Cultural Materials (OCM), an ethnographic subject classification system developed by Murdock and colleagues (Murdock et al., 1969; Ford, 1971).

The taxonomy is designed in a decimal classification system, similar to the librarian Universal Decimal Classification. Its main categories start with Orientation (10), Bibliography (11) and Methodology (12), and end with Socialisation (86), Education (87) and Adolescence, Adulthood, and Old Age (88). The categories are still very general, so more specific categories must be coded additionally.

Ethnographic systematics (ETSEO) is derived from continental ethnographic practices, mostly interested in traditional culture of the European peasantry. Its taxonomy is hierarchically extensive, starting with the essentially defined material, spiritual and social culture categories. Since the taxonomy was designed for museum archives, the most detailed is the field of material culture, subdivided on as many levels as necessary, and taxonomy in general fits folk taxonomy and practices. Spiritual culture is further divided into general categories comprising folklore, ritual practices, and art-related activities. Less detailed is the so-called "social culture" field containing festivities in a calendar year, celebrations of live events, and communal activities, practices, and rules. This system is much more detailed but, at the same time, only partly decimally classified and only somewhat comparable to the OCM taxonomy. It was designed for classical archive work and is now only partially accepted as a digitised taxonomy.

OCM's main aim was to facilitate searching the large database of ethnographic entries and organise basic information on ethnic and social groups. Hence it is easy to extend the idea of an ethnographic classification system to a codebook – each entry represents a concept relevant to describing a culture. One could use the well-defined system with descriptions of categories to automatically tag text passages with relevant ethnographic concepts. For example, if the passage describes using outdoor toilets, the corresponding codes should be "744 Public Health and Sanitation", "515 Personal Hygiene", "336 Plumbing", and "312 Water Supply". Besides already existing taxonomies for ethnographic materials (OCM and ETSEO), it is useful to produce native or folk taxonomies as "a description of how people divide up domains of culture, and how pieces of a domain are connected" (Bernard, 1994, p. 386). Automated accurate tagging would enable quickly retrieving relevant parts of the text on the one hand and observing dominant topics and their inter-relatedness on the other.

5. Document enrichment

Analysis of interview transcripts would normally include labelling documents or interview segments with corresponding codes, identifying topics/codes, observing their frequencies in the corpus, and retrieving interview segments for a given topic/code. We show how to perform these tasks in a visual-programming data mining tool Orange (Demšar et al., 2013). Workflow (as seen in Figure 1) for replicating the analysis is available online (Pretnar Žagar, 2022b) along with a Slovenian translation of OCM ontology (Pretnar Žagar, 2022a). The corresponding data are not publicly available due to privacy issues.

5.1. Data and preprocessing

To demonstrate how contemporary ontology enrichment and semantic analysis approaches can be used in anthropology, we are using interview transcripts from twenty interviews on smart buildings (Pretnar and Podjed, 2019). The interviews are in colloquial Slovenian and describe

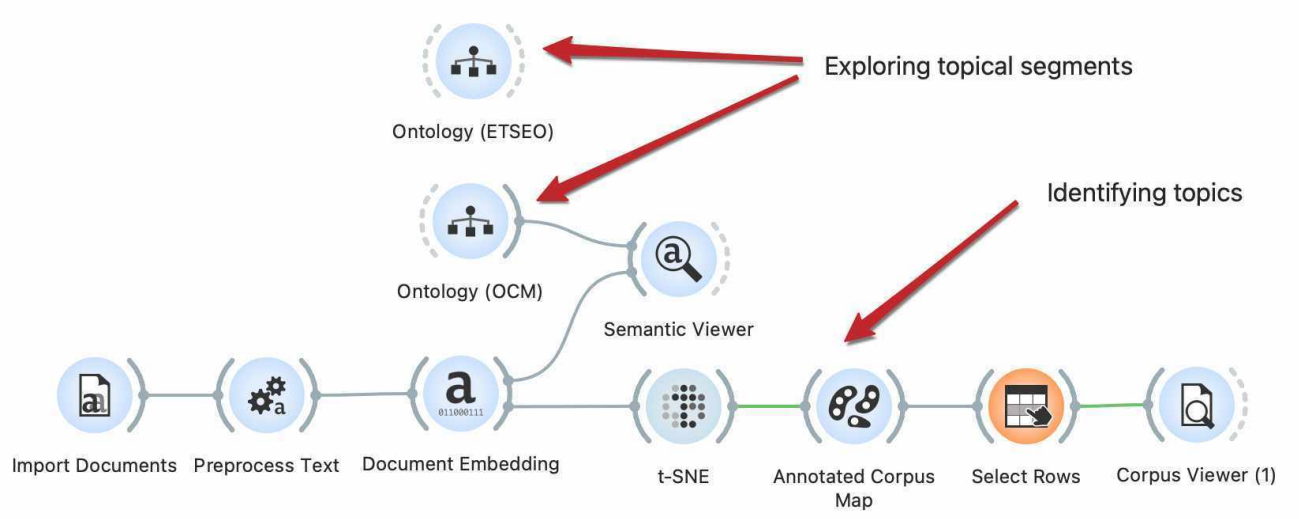


Figure 1: Workflow for ontology enrichment and extracting interview topics from annotated visualization.

the experiences and struggles of faculty staff with a smart building. The interview is segmented into questions and answers. Each answer represents the utterance and constitutes a single document in the final corpus resulting in 1126 data instances. The metadata includes the question, the interviewee, and the interview date.

Tokens are constructed by passing the text through the CLASSLA pipeline for non-standard Slovenian. Then, lemmas and POS tags are retrieved, and only nouns and verbs are kept for the analysis. Tokens are used to compute document embeddings, a mean-aggregation of word embeddings based on fastText models (Bojanowski et al., 2017). We tried simple lowercasing, Lemmagen lemmatization (Juršič et al., 2010) and stopword removal for preprocessing, but the results were not as informative (they mostly contained generic verbs, such as to have and to go, discourse particles and fillers). Moreover, while SBERT embeddings generally perform better due to their context-parsing abilities, they produced worse results in the t-SNE visualisation. Specifically, fastText identified a group of segments with short, unspecific replies (i.e., "Yes.", "Uh-huh."), while SBERT did not.

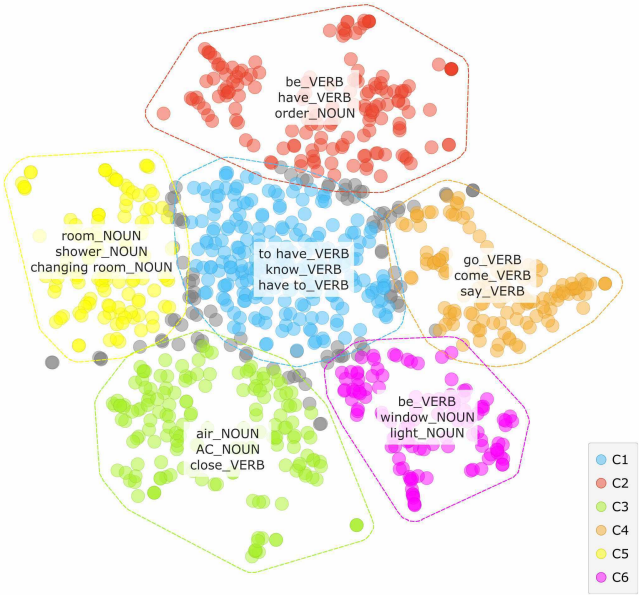


Figure 2: t-SNE document map with annotated semantic groups.

5.2. Identifying topics

Generally, the researchers will know which topics the corpus covers because often, they will be its creators. In the case of interviews, the researcher is likely also the interviewer who guided the interview based on research questions. However, ethnographic narratives often take unexpected turns or focus on unforeseen details, which the researcher can uncover by coding the data and iteratively refining the codebook. Alternatively, one can use document maps, where segments with semantically similar content will lie close together.

To semantically represent the content of interview segments, we will pass them to document embedding. The procedure will take the words (tokens) identified in preprocessing and find their vector representation. The representation models the meaning of the words in a way that relates "king" to "prince" and "queen" to "princess". Once the embedding of each word is retrieved, words from the document are aggregated into the mean document vector.

This numeric representation will be used to plot a t-SNE document map, where segments with similar content will lie close to each other[2]. But a bare map is not very informative on its own. Hence, we added Gaussian mixture models to identify groups of segments and retrieve their characteristic words (Figure 2). The procedure identified segments referring to air quality (green cluster), lighting (magenta cluster), room descriptions (yellow cluster), and so on.

[2]In t-SNE, we selected a larger group of segments for annotation. There was a smaller group of 121 segments representing short replies, such as "yes", "no", and "I don't know".

5.3. Exploring topical segments

Ontologies can be used to enrich interview segments by measuring how similar given ontology terms are to each segment. Automatic identification of segments helps researchers quickly identify relevant parts of the interview.

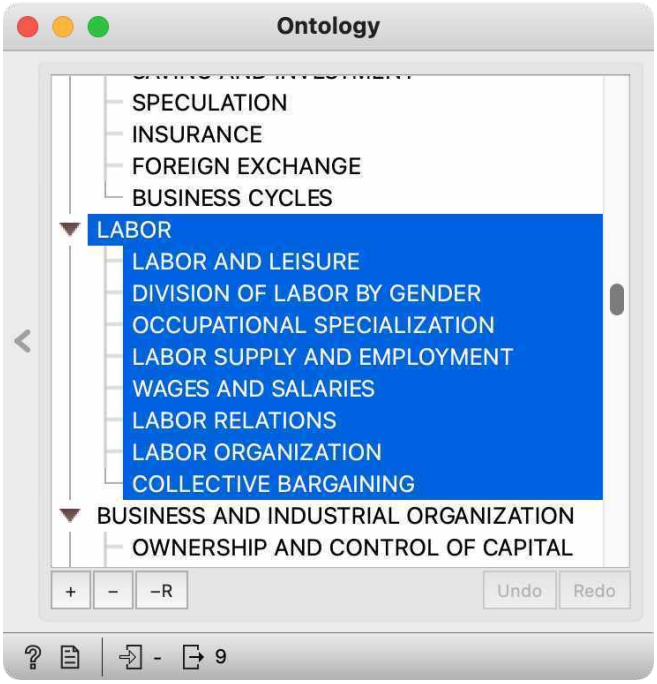


Figure 3: Selecting a part of the OCM ontology referring to work ("delo") and work-related terms.

For example, we can look for "delo" (orig. 350 equipment and maintenance of buildings) and its child terms from the OCM ontology in the corpus (Figure 3). Selected terms from the ontology are used for semantic annotation of interview segments.

Semantic annotation scores each segment by how similar its sentences are to the input terms, using SBERT embeddings (Reimers and Gurevych, 2019). SBERT was used because it specialises in sentence embeddings and considers word context. Ideally, this procedure identifies passages talking about work-related topics, including breaks, employment, paychecks, and work relations. One can sort the results by either the overall segment score, an aggregate of all sentence scores, or by matches, which counts how many input words appear in the segment.

Here, we show the latter option, namely displaying the segments with the most matches. We have selected all the segments matching any of the input terms and highlighted them (Figure 4). Ontology enrichment successfully identified segments discussing the office environment, research work, work routine, schedules, weekend work, etc.

5.4. Assigning terms to segments

The final goal of any automated coding system would be to return a corpus with assigned codes. We prototyped a procedure that uses the above technique of semantic scoring to identify the code with the highest score for each segment. We decided on a 0.6 cosine similarity threshold for the code

At work, not so much. More so in the afternoon.

It mostly works. But one sometimes has to fix things.

Mostly, yes.

Yeah, sure. I mean PhD students for sure.

From the office?

Both. If we are in the meeting at the same time, then absolutely, if we're on a break, the same.

For research, for teaching preparations, for meetings with the students, with PhD students, with anyone who comes to me. Purely for work purposes.

That works.

Figure 4: Annotating text segments with a part of ontology referring to work ("delo").

to be assigned, which resulted in segments that did not have a corresponding code. After loading the corpus, we remove all the interview segments without any codes. We retain 252 segments with codes and observe their frequencies. The results are somewhat promising but with some obvious errors (Figure 5).

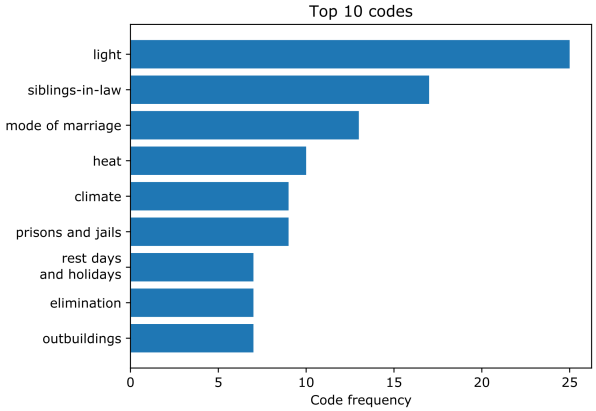


Figure 5: Top 10 codes identified in the corpus. While some are plain wrong, most are quite accurate and useful.

The most frequent code is "luč" (light), which is indeed a very prominent topic in the corpus. Then the results get a little strange. The two next topics are "svaki in svakinje" (brothers and sisters in law) and "tipi porok" (marriage types), which are not among the interview topics. The errors are likely caused by the multilingual SBERT model used for word embedding, which sometimes cannot distinguish between South Slavic languages. For example, it considers the Slovenian slang term "ratal" (succeeded) as "war" based on its similarity to the Serbian "rat" (war).

However, there are some quite relevant topics among the top ten codes, for example, "toplota" (warmth), "podnebje" (climate), "dnevi počitka in dela prosti dnevi" (rest

days and holidays), “stranska poslopja” (outbuildings), and “bivališča” (dwellings). Clicking on a label, for example, “toplota” (warmth), outputs text segments discussing the interviewees’ attitude to temperature regulation. With a few steps, the researcher can identify and extract interview segments discussing a specific topic and read them to better understand the context of these segments and which subtopics the respondents deem relevant. For example, the texts on temperature regulation mostly refer to difficulties with adjusting office temperature.

The system could be improved with specifically developed language resources for non-standard Slovenian. Nevertheless, even in its current imperfect form, it can be a useful tool for semi-automated coding, where the researcher can manually adjust the suspicious/incorrect codes.

5.5. Comparison to ETSEO taxonomy

While the OCM taxonomy is widely recognised in the anthropological community, the ETSEO taxonomy is strictly regional. The project Ethnological Topography of Slovenian Ethnic Territory (ETSEO) began in 1971 by a large group of Slovenian ethnologists led by Slavko Kremenšek. The project entailed the development of the questions based on ethnological systematics, ethnographies of Slovenian towns and cities (18 in total), and detailed ethnographies on a specific topic. The taxonomy is a result of the first part of the project, namely the questions and detailed ethnological systematics. The ETSEO questions were published between 1976 and 1977 in twelve books, including the introductory volume with reports of ethnological institutions (Kremenšek et al., 1976) and eleven volumes of topical presentations and suggested questionnaires. The series served as a theoretical and practical guide for ethnographic fieldwork (Ravnik, 1996).

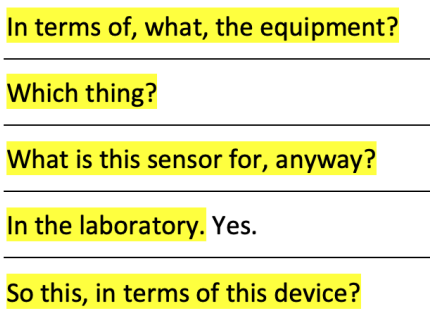


Figure 6: Matches for ETSEO entry “technical knowledge”.

ETSEO taxonomy contains 53 areas of ethnographic interest. Still, it lacks explicit hierarchy, although it follows the classical division of ethnographic material for the so-called folk culture: material (volumes I to V), social (volumes VI to VIII) and spiritual (volumes IX to XI). A rough hierarchy could be formed from the eleven books in which these questions were published, but the books lack hypernyms. Hence, we will use this as a flat taxonomy. There are fewer relevant areas to choose from than in the OCM. However, looking for “tehnično znanje” (technical knowledge) returns relevant interview passages (Figure 6).

The ETSEO taxonomy is less useful than the OCM taxonomy. This is due to the somewhat outdated nature of the questions, which were based on the main foci of Slovenian ethnology and were less relevant for anthropology. They are missing some key contemporary areas of anthropology, namely media, urban areas, internet communities, and migration. Nevertheless, the taxonomy could be extremely useful for older ethnographic texts and, with some updates, even for contemporary materials.

6. Conclusion

Anthropology can greatly benefit from the recent developments in text analysis. Ontology enrichment, along with other data exploration and visualisation methods, is a useful tool providing an overview of the collected data.

In the time when anthropologists are using larger corpora (Culhane and Elliott, 2016), when data is created online for many different purposes (Wang, 2012), and when anthropologists use online platforms to store raw ethnographic multimedia data (Przybylski, 2021), it is of utmost importance to store and later archive data meaningfully, using relevant classification and coding systems. It is even more important in archival work, which is no longer just an additional part of anthropological research, supplementing ethnographic fieldwork, but is becoming highly relevant for digital aspects of our lives.

Updating taxonomic systems is an urgent task for anthropologists. However, using existing taxonomies to explore and visualise data already benefits the analytic process, especially in re-studies and comparative research. Classical anthropological coding of ethnographic material is no longer possible, so automated coding is the first step to expanding the range of anthropological data analysis. However, in the absence of specialised word embedding models for Slovenian (SBERT is currently multilingual and conflates South Slavic languages), the approach does not yet achieve the accuracy of a human annotator.

While automated coding, particularly for languages with fewer language resources, still has a long way to come to be comparable to human input, it facilitates data exploration and extracting general topics from the text. Ontology enrichment tools support the iterative analytical process of ethnography. They provide a starting point for forming new research questions, enhancing existing ones and can be easily repeated on new data.

Many improvements could be made to improve automated coding for the Slovenian language:

- Developing a Slovenian-only sentence transformer used in semantic search.
- Re-writing transcripts in standard Slovenian or further improving CLASSLA to handle slang terms and non-standard Slovenian.
- Implementing co-reference resolution for Slovenian to resolve issues with indirect references in text, further clarifying the exact content of the document.

While these improvements would greatly enhance coding capabilities for Slovenian, they are, for the most part,

available for larger languages, thus already enabling similar research.

7. Acknowledgements

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.

8. References

Blaž Bajič. 2020. Nose-talgia, or, olfactory remembering of the past and the present in a city in change. *Ethnologia Balkanica*, 22:61–75.

H Russell Bernard. 1994. *Research Methods in Anthropology: Qualitative and Quantitative Approaches*. Sage Publications, Thousand Oaks, London, New Delhi.

Aristeidis Bifis, Maria Trigka, Sofia Dedegkika, Panagiota Goula, Constantinos Constantinopoulos, and Dimitrios Kosmopoulos. 2021. A hierarchical ontology for dialogue acts in psychiatric interviews. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, PETRA 2021, page 330–337, New York, NY, USA. Association for Computing Machinery.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 543–551.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Alain Cuerrier, Nicolas D Brunet, José Gérin-Lajoie, Ashleigh Downing, and Esther Lévesque. 2015. The study of inuit knowledge of climate change in nunavik, quebec: a mixed methods approach. *Human Ecology*, 43(3):379–394.

Dara Culhane and Denielle Elliott. 2016. *A Different Kind of Ethnography: Imaginative Practices and Creative Methodologies*. University of Toronto Press, North York, Ontario, Canada.

Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, et al. 2013. Orange: data mining toolbox in python. *The Journal of Machine Learning Research*, 14(1):2349–2353.

Clellan S Ford. 1971. The development of the outline of cultural materials. *Behavior Science Notes*, 6(3):173–185.

Primož Godec, Nikola Đukić, Ajda Pretnar, Vesna Tanko, Lan Žagar, and Blaž Zupan. 2021. Explainable point-based document visualizations. *arXiv preprint arXiv:2110.00462*.

Thomas R Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5-6):907–928.

Sarah Hitchner, John Schelhas, and J Peter Brosius. 2016. Snake oil, silver buckshot, and people who hate us: metaphors and conventional discourses of wood-based bioenergy in the rural southeastern united states. Human Organization, 75(3):204–217.

Seth M Holmes and Heide Castañeda. 2014. Ethnographic research in migration and health. *Migration and Health: A Research Methods Handbook*, pages 265–277.

Annette Hoxtell. 2019. Automation of qualitative content analysis: A proposal. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, volume 20.

Matjaž Juršič, Igor Mozetič, Tomaž Erjavec, and Nada Lavrač. 2010. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9):1190–1214.

Slavko Kremenšek, Vilko Novak, and Valens Vodušek. 1976. *Etnološka topografija slovenskega etničnega ozemlja. Uvod. Poročila.* Raziskovalna skupnost slovenskih etnologov, Ljubljana.

Michael W Longan. 2015. Cybergeography irl. *Cultural Geographies Special Issue - New Methods in Cultural Geography*, 22(2):217–229.

Douglas Craig MacMillan and Sharon Phillip. 2010. Can economic incentives resolve conservation conflict: the case of wild deer management and habitat conservation in the scottish highlands. *Human Ecology*, 38(4):485–493.

M Besher Massri, Sara Brezec, Erik Novak, and Klemen Kenda. 2019. Semantic enrichment and analysis of legal domain documents. *Artificial Intelligence*, page 2.

George Peter Murdock, Clellan S. Ford, Alfred E. Hudson, Raymond Kennedy, Leo W. Simmons, and John W. M. Whiting. 1969. *Outline of Cultural Materials*. Human Relations Area Files, New Haven.

Aaron Podolefsky and Christopher McCarty. 1983. Topical sorting: A technique for computer assisted qualitative data analysis. *American Anthropologist*, 85(4):886–890.

Ajda Pretnar and Dan Podjed. 2019. Data mining workspace sensors: A new approach to anthropology. *Prispevki za novejšo zgodovino*, 59(1):179–196.

Ajda Pretnar Žagar. 2022a. *OCM ontology - Slovenian*. Figshare. https://doi.org/10.6084/m9.figshare.19844107.v1.

Ajda Pretnar Žagar. 2022b. *OCM ontology enrichment*. Figshare. https://doi.org/10.6084/m9.figshare.19787065.v1.

Liz Przybylski. 2021. *Hybrid Ethnography: Online, Offline, and in Between*. Sage Publications, Los Angeles; London; New Delhi; Singapore; Washington DC; Melbourne.

Mojca Ravnik. 1996. Način življenja slovencev v 20. stoletju. *Traditiones*, 25:403–406.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Cyril Schafer. 2011. Celebrant ceremonies: life-centered funerals in aotearoa/new zealand. *Journal of ritual studies*, 25(1):1–13.

Anselm Strauss and Juliet M Corbin. 1997. *Grounded Theory in Practice*. Sage.

Juhana Venäläinen, Sonja Pöllänen, and Rajko Mursic. 2020. The street. The Bloomsbury Handbook of the Anthropology of Sound.

Tricia Wang. 2012. The tools we use: Gah- hhh, where is the killer qualitative analysis app? http://ethnographymatters.net/blog/2012/09/04/the-tools-we-use-gahhhh-where-is-the-killer-qualitative-analysis-app/.

Priscilla M Wehi, Murray P Cox, Tom Roa, and Hēmi Whaanga. 2018. Human perceptions of megafaunal extinction events revealed by linguistic analysis of indigenous oral traditions. *Human Ecology*, 46(4):461–470.

Sinem Yilmaz, Bart Van de Putte, and Peter AJ Stevens. 2019. The paradox of choice: Partner choices among highly educated turkish belgian women. *DiGeSt. Journal of Diversity and Gender Studies*, 6(1):5–24.

Slavko Žitnik and Marko Bajec. 2018. Odkrivanje koreferenčnosti v slovenskem jeziku na označenih besedilih iz coref149. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, 6(1):37–67.

Parliamentary Discourse Research in History: Literature Review

Jure Skubic♦, Darja Fišer*♦
♦Institute of Contemporary History, Ljubljana, Slovenia
*Faculty of Arts, University of Ljubljana, Slovenia
♦Privoz 11, 1000 Ljubljana, Slovenia
*Aškerčeva cesta 2, 1000 Ljubljana, Slovenia
jure.skubic@inz.si, darja.fiser@ff.uni-lj.si

Abstract

Historical research of parliamentary discourse focuses not only on the origins but especially on the development of parliamentary discourse. It is predominantly based on textual data analysis, employing various methodological frameworks. In this literature review we provide an overview of these methods and present commonalities and differences of approaches established in history with corpus-driven approaches. This allows for a better understanding of historical analysis of parliamentary discourse and highlights the importance of ParlaMint project and the integration of parliamentary corpora into historical research.

1. Introduction

Parliamentary discourse is a salient research topic in both humanities and social science disciplines, such as sociology, political science, sociolinguistics, and history. Especially historical research is highly interested in studying not only the origins but also the development of parliamentary discourse. History is often focused on researching parliamentary debates and as Ihalainen (2021) observes, in historical research, parliamentary debates can be approached analytically as nexuses of past political discourses which means that they can be viewed as "meeting places" where in a certain time and space various political discourses have intersected.

This literature review is one in the series of literature reviews conducted in the context of the ParlaMint project (Erjavec et al., 2022). A similar literature review has been compiled for sociological research (Skubic and Fišer, 2022). The ParlaMint project develops comparable corpora of parliamentary proceedings from more than 20 European countries, accompanied by literature overviews, showcases and tutorials which will hopefully help maximize the use of these corpora in different disciplinary communities interested in analyzing parliamentary debates. This literature review summarizes historical research of parliamentary debates and the most popular research methods employed. It needs to be explicitly noted, however that despite the obvious usefulness of ParlaMint corpora, the researchers ought to consider also other qualitative and quantitative data and information in order to come to objective and unbiased conclusions. Also, in this review we focus mostly on written parliamentary records since the main interest of ParlaMint project is on written parliamentary sources. However, the importance of other sources such as surveys, records of election results, territorial records, etc. must be recognized as well since they present an important part of historical research.

The review is structured as follows. In the first part, we describe the selection procedure of the relevant articles and briefly enumerate the methods they employ. This allows for a better understanding of the methods most frequently employed in historical analysis of parliamentary discourse. In the second part, we summarize the articles we identified in terms of 1) the main aim and topic of the research, 2) methods used, 3) data collection methods and 4) a short discussion about the possible improvements and/or problems of the research. We conclude the review with a discussion of how historical research could benefit from corpus data and corpus research methods.

2. Literature Selection and Methods

As Torou et al. (2009) show, the main objective of history is to recreate the past by researching and analyzing existing records and their interconnectedness. It is through this process that historians employ their academic knowledge, rely on experience, and decide on the relevant information and appropriate sources which this information is extracted from. Especially in political history, it is uncommon for historians to rely on only one type of source, but rather focus on various so called primary and secondary sources. The former are most commonly gathered from historical archives since they include document or artefacts created by the participants in an event or the witnesses, whereas latter include oral sources, newspapers, memoirs, visual representations, practices, etc. This means that an important factor in historical research is to understand the nature of information as well as the research methodologies and models historians use while conducting research (ibid.).

Although the variety of issues and approaches in political history is large, the emerging and quite narrow focus of political history is on analyzing the history of parliamentary discourse and political debates. Ihalainen and Saarinen (2019) show that political history frequently builds its research on textual data (documents, diaries, texts) although sometimes the exact textual methods used are not explicated. Ihalainen and Saarinen (2019) note that when conducting textual analysis, historians often draw on selected methodological tools from methods which are otherwise common in humanities and social sciences and especially qualitative sociological research, such as (critical) discourse analysis as well as content analysis. In addition to those and to other fields which include the study

of history (memory studies, conceptual history, etc.) researchers sometimes opt for mixed methods approach, corpus assisted discourse studies or text mining.

2.1 Selection of Articles

The reviewed articles were carefully selected among hundreds of sources which focus on parliamentary debates by considering some important research criteria. We identified the following scholarly search engines to look for the articles:

- Taylor and Francis Online (https://www.tandfonline.com),
- SAGE Journals (https://journals.sagepub.com),
- Wiley Online Library (https://onlinelibrary.wiley.com),
- Semantic Scholar (https://www.semanticscholar.org),
- MUSE Project (https://muse.jhu.edu),
- JSTOR (https://www.jstor.org),
- Elsevier (https://www.elsevier.com), and
- Google Scholar (https://scholar.google.com).

We applied the following filters in order to identify the relevant articles:

- Publication period: 2012 – 2022,
- Discipline: History
- Article ranking: 'most relevant' and 'most cited'
- Relevant journals: sometimes we needed to apply additional filter where we selected relevant historical journals.

By using those filters, most prominent historical journals were identified, such as Parliamentary History, Historical Research, Memory Studies, Contributions to the History of Concepts and Historical Social Research, although articles included in this review were also published elsewhere. All articles, the title of which was considered potentially relevant were skimmed; we specifically analyzed the abstract, methodology and analysis sections to confirm the relevance of the articles. A high number of articles was discarded either because of the lack of methodological explanation or because the analysis did not focus on parliamentary data. In this review we wanted to include only those articles which dealt specifically with parliamentary records and/or legislative documents and the majority of the selected research conformed to this criteria. Some of the articles, however, also included other sources which emphasizes the fact that historians use a variety of sources when researching parliamentary discourse. This is also to show that although parliamentary records could present one of the primary sources for historical research (and projects such as ParlaMint would be helpful in providing relevant data), historians still often opt for a broader research perspective and combine parliamentary records with other, complementary sources of data in their research.

2.2 Overview of Methods

A total of 27 articles were initially determined as relevant for our literature review and are listed in a Google spreadsheet.[1] We then retained only those that clearly described the method and the data used, taking into account only the papers which primarily used parliamentary records as a source. This resulted in 11 articles which were then submitted to a more detailed analysis. Since the research questions were so heterogeneous, we did not group the articles thematically.

We reviewed predominantly articles which focused on historical research of parliamentary discourse and political communication. Out of 11 reviewed articles, 3 employed the methodological framework of Discourse Analysis, 2 articles employed Content Analysis, 1 opted for the method of Memory Studies, 2 articles used Mixed methods approach, 2 articles employed the framework of Conceptual History (Begriffsgeschichte), and 1 article employed the method of Topic Modelling.

3. Reviewed Research and Employed Methods

In this part of the literature review, we give a detailed account of the historical research that analyzes parliamentary discourse and political communication as well as the methods they employ. We provide a short description of the methodological framework and show why it is important for historical research. Then, we give an overview of the studies which employed this method.

3.1 Conceptual History

Conceptual History (Begriffsgeschichte) is a strand of historical studies which deals with historical semantics and the evolution of paradigmatic ideas and value systems over time. It was first defined by Koselleck in 1997 who shows (as cited by Litte, 2016) that the major aim of conceptual history is to uncover the logic and semantics of the concepts that have been used to describe historical events and processes in addition to being interested in historical evolution of some concepts over time. Ihalainen and Saarinen (2019) note that Conceptual History, when combined with Political History, mostly focuses on past human interaction and communication, and understands discourses as central interlinked elements of political processes, events, and action.

Interest in the field of Conceptual History was quite high in the 20th Century Germany, especially when conducting historical research of the World War II. Later, the field became prominent in political history for the analysis of political communication and events. As shown by Litte (2016), conceptual history has three main tasks: firstly, to identify the concepts that are possible in characterization of history, then to locate those concepts in the context of political or social discourses and finally to critically evaluate those concepts for their usefulness for historical analysis

3.1.1 Debates on Democracy in Sweden

Research problem: Friberg's (2012) article aims to explore the concepts of democracy that were used in Sweden and especially focuses on how the concepts were

[1]https://docs.google.com/spreadsheets/d/13mF_X3OB9CKtdfsUFDLPJZJ44VcxZ1uv9OzAE2E_E-I/edit#gid=1690588464

used by the Social Democrats (SDP) during the interwar years when the party was establishing itself and its political agenda. It examines the Swedish parliamentary rhetoric about democracy after the full suffrage reform.

Research method: The author employed German Begriffsgeschichte (conceptual history approach) as introduced by Koselleck, and the theory of ideologies (Freeden, 1998 as cited by Friberg, 2012). According to Friberg, these two methods complement each other since conceptual history emphasizes how socio-political context influences the changing meaning of the concept whereas theory of ideologies finds the meaning of this concept dependent on morphological structure.

Data collection: The main source of the data for this article were the debates in Swedish Parliament during the interwar years. In addition, the author used other governmental materials, such as reports from different committees. Both sources were only available as hardcopies, but they provided coherent source materials. The debates which were analyzed were chosen according to two important criteria. First, the debates needed to be explicit discussions in Parliament and needed to focus on the concept of democracy in the interwar years. Second, the debates had to be related to a topic that a political party (in this case Social Democrats) claimed was connected to democracy in a certain way. The debates which conformed to the first criterion were identified through the subject index of the governmental records, whereas the debates which needed to observe the second criterion were recovered through an extended analysis of materials such as party manifests, newspaper articles and records from the congress. This was necessary to get the feeling of what the SDP claimed to be connected to the democracy and then compare those records with parliamentary records. In addition, the author analyzed the articles from the Social Democratic journal titled Tiden, which throughout the 20th century was one the most important Social Democratic newspapers for conducting internal debates. The analysis of these articles added to the reliability of the conceptual analysis.

Discussion: One of the problems with data collection was that all the records were accessible only as hardcopies and not electronically. Although the author gives no information about that, we could assume that the documents needed to be thoroughly read and notes taken. Also, parliamentary records do not exactly depict the actual debates since the process from an actual debate to a printed one used to be rather complicated and long. This results in sometimes significant differences between the actual speech and the written text. This long process of editing, changing, and adapting the actual text to be suitable for a printed version results in the data not objectively depicting what was said during the debate.

3.1.2 Debates on Immunity in France and Romania

Research problem: In this article Negoita (2015) analyzes the concept of parliamentary immunity. His main goal is to identify not only historical premises but also linguistic, political, and legal instruments that played part in conceptualization of parliamentary immunity in two countries – France and Romania. This article, therefore, although historical in nature, employs interdisciplinary perspective when studying parliamentary discourse and investigates the concept of the word "immunity" as used in parliamentary discourse.

Research method: The author employs methodological framework of Conceptual History and makes comparative analysis of the two aforementioned countries. We could therefore understand this method as comparative conceptual analysis.

Data collection: The data was collected from a variety of sources which were mostly not parliamentary ones. For French, dictionaries (Le Grand Robert de la langue francaise, Dictionnaire de l'Academie francaise, etc.), scientific works which focused on the history of French parliamentarism (Histoire de France or Les caracteres ou les moueurs de ce siècle), as well as various political documents and French Constitution were used. Romanian data was also gathered from dictionaries (Dictionar al institutiilor feudale din Tarile Romane, for example), as well as various historical documents and different versions of democratic Constitutions. What all documents had in common was that although they were not strictly records of parliamentary debates, they did focus on the parliamentary and political language and discourse.

Discussion: This research is slightly different from the others in this review since it does not draw directly from the parliamentary records. This analysis successfully shows how historical analyses frequently draw on sources other than explicitly parliamentary data.

3.2 (Collective) Memory Analysis

Memory analysis combines intellectual strands from various domains such as history, sociology, anthropology, education, etc. Since this is an emerging field of research, its qualitative and quantitative methodological tools are not yet fully developed. Instead, researchers who conduct memory analysis usually borrow methodological tools from other social sciences and adapt them for their own purpose. These methods frequently contain content and (critical) discourse analysis.

The main aim of memory analysis is the study of forms and functions of representing the past. Data collection includes a careful examination of primary historical sources and archival studies, as well as secondary sources such as case studies, interviews, surveys, and eyewitness reports. Once the data is collected, the aforementioned methodological tools are employed to thoroughly analyze the data. Memory analysis frequently also includes the research of collective memories and narratives. Collective memory as defined by Hogwood (2013) is a concept, which is used across disciplines to refer to the ways the past is "perceived, shaped, and constructed" and its main aim is to extract useful data from collective conversations, sharing ideas and media. This then leads to a synthesis of voices and formation of a common information thread among peers.

One of the major methodological problems that occurs inside memory analysis, is that when researchers conduct research, they usually use whatever evidence is readily available, without digging deeper into the event and research it more thoroughly. This points to the fact that even though memory analysis is a useful field of historical analysis, researchers must be attentive to employ other

approaches with which they confirm and legitimate the findings of memory analysis.

3.2.1 The Nation in Parliamentary Discourse on Immigration

Research problem: De Saint-Laurent (2014) focuses on exploring the meaning that is attributed to the national group. The aim of her article is to analyze collective memories (she names them narratives) and show what meaning they give to the nation, how this meaning is produced and how the stories told by different groups relate to one another.

Research method: She employs a qualitative analysis of collective narratives of the past. In connection with memory analysis, she employs dialogism as a methodological tool since the analysis of dialogic overtones helps reconstruct the social processes through which the discourse is done.

Data collection: It needs to be noted that this article is an analysis of the meaning which is given to the concept of nation in French parliamentary debates over a bill on Immigration and Integration. The data used consisted of official transcripts of fifteen parliamentary sessions which happened between May 2 and May 17, 2006. In addition to that, the author also included the vote session which happened on June 30, 2006. All documents are made available to the public through the official parliamentary website. In addition to using general reactions of the Assembly, the author used transcripts of the participants interventions and interruptions from the entire sessions. Once the author determined the datasets, she began with relevant data selection, which happened in three stages. In the first stage, the author identified those excerpts which were relevant for the study of the role of collective memory. She did that with the help of Nvivo[2] software (QSR International Pty Ltd., 2020). In this stage the author also extracted relevant references by carefully reading through all the debates and employing a keyword search, which contributed to pinpointing the indirect references. The second stage was the coding stage, where firstly thematic coding happened to map out relevant historical periods and secondly the groups which the speakers belonged to were coded into two categories – political party and outside the political spectrum. In the third stage the fragmented excerpts and data were used to reconstruct past narratives, which were then thoroughly analyzed.

Discussion: This paper is not only historical since the author herself notes that it also adopts "socio-cultural psychological perspective on memory" (ibid.). She also notes that because of the reconstructive aspect of her analysis, she checked the narratives against certain complementary sources (research in French newspapers, blogs, websites, etc.). This made the research much more reliable.

3.3 Discourse Studies

Van Dijk (2018) uses the term discourse studies to refer to a field of research, which includes various qualitative and quantitative methods and different genres, such as news reports or parliamentary debates. This field emerged in the 1960s and is very prominent inside humanities and especially social sciences. The field of Discourse Studies includes various methods, such as Discourse Analysis (DA), Critical Discourse Analysis (CDA) and Political Discourse Analysis (PDA). All three were detected as salient in this literature review.

Discourse Analysis (DA) is one of the most frequently used methods in those social science disciplines, where the focus is frequently on the study of language and text. In historical research, Discourse Analysis is sometimes referred to as Discourse Historical Approach (DHA) and its main defining feature is in acknowledging the historical context and attempting to integrate this knowledge together with background of social and political fields into research. DHA focuses on studying the display of power through language and conceptualizes history through a theorized lens of critique. This method shares various common features with the Critical Discourse Analysis (CDA) and provides a clear description of how to integrate historical context to critical discourse analysis, highlighting the importance of historicity to understand the continuities of discourses (Achugar, 2017). Sometimes DA, when used to analyze political discourse, is referred to as Political Discourse Analysis (PDA) (Dunmire, 2012).

Critical Discourse Analysis (CDA) examines the means by which political power is manifested or abused through discourse structures and practices (Dunmire 2012). Achugar (2017) shows that since the past has become an area of focus for CDA, this method has become a salient one in historical research. One of its major aims is to provide an explanation of the power differences in contemporary society by researching the past events and their context.

3.3.1 British Parliament and Foreign Policy in the 20th Century

Research problem: Ihalainen and Matikainen (2016) investigated the parliamentarization of the foreign policy in British Parliament throughout the 20th century. They argue that throughout the 20th century, parliaments in general gained more power in discussing foreign policy and especially in British Parliament this parliamentarization of foreign policy debates was highly noticeable.

Research method: They combine analysis of the policy documents with the more discourse-oriented analysis of parliamentary debates. Their research method is more discourse-oriented than the traditional diplomatic history since they do not focus only on policy documents but also consider the discourse of parliamentary debates in that time.

Data collection: The authors utilized a wide variety of primary sources with the Hansard constituting the starting point of their analysis. They also used parliamentary papers such as committee reports as well as the relevant sources created by other political actors – the foreign office, other relevant government departments, voluntary associations, the media, etc. Their data therefore consists of parliamentary debates on the one hand and archival

[2] https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

documents, public debates, and interviews on the other. They argue that the use of such a wide range of data was necessary to grasp the multi-sided nature of the policy discourse and to ensure that data was vast enough to provide the complete picture of how the parliamentarization of foreign policy debates occurred. Parliamentary records database was electronically available which resulted in authors utilizing full-text searches to locate sources for contextual analysis of parliamentary debates. They wanted to locate potentially interesting debates and analyze them by using aforementioned historical methods.

Discussion: Authors do not provide a detailed account of how the data was collected and give no information about how the documents, other than the electronically accessible Hansard, were obtained. They do, however, clearly show that in order to conduct thorough historical research, a variety of sources needs to be studied and that focusing only on parliamentary debates is not enough.

3.3.2 British Lobbying and Parliamentary Discourse

Research problem: McGrath (2018) focuses his research on lobbying which he sees as a significant component of the modern politics in Britain. In his article, he provides a detailed explanation of the scale and significance of lobbying and studies how lobbying in Britain was discussed not only by parliamentarians but also by journalists.

Research method: The author utilized keyword search on several digitized archives which helped him gather extracts from parliamentary debates and newspaper articles. He blended both qualitative and quantitative readings of the texts which leads us to assume that some kind of discourse analysis method was employed.

Data collection: The author draws on parliamentary debates as well as three other databases which together consist of 51 newspaper titles between 1800 and 1950. The data was available in electronic archives and already in written form, so no transcription was needed. The unit of the analysis is an individual newspaper article or parliamentary speech. The database consisted of four online archives: 1) Hansard (1803-1950), 2) British Library (1800-1900), 3) The Times (1800-1950), and 4) The Guardian (1800-1950). To gather the source material, the author employed a three-step process; firstly, each archive was searched using a range of keywords which are associated with lobbying which produced roughly 1.691 items. Secondly, each item was printed, carefully read, and sorted according to the descriptor he was searching for. Some of the data has already been discarded here since it did not correspond to the search parameters (e.g., did not relate to governmental bodies, material covered lobbying in countries other than Britain, etc.). In the third stage, the items which were not removed were put into chronological order and the author removed all the duplicates. This resulted in 689 items being determined as suitable for analysis. Once all the unique items were collected the individual items were examined and coded. To acquire the appropriate data, McGrath employed a five-stage process to transform qualitative material into quantitative data, although not all stages needed to be applied. He sourced the material but did not need to transcribe it as it has already been made available in textual form. Then the data was unitized, then categorized on the basis of the actual data and relevant theory and finally each unit was separately coded.

Discussion: The author never explicitly mentioned discourse analysis as his research method. But since he talks about conducting a qualitative analysis of discourse from parliamentary records and newspaper articles, we could assume that he employed discourse analysis approach.

3.3.3 Nationalism and Political Discourse in Scotland

Research problem: The research conducted by Whigham (2019) critically examines the narratives which emerged from party political discourse after Scottish independence referendum in 2014. The aim of the research is to analyze the past discourse on nationalism in Scotland and to critically reflect on narratives about Scottish nation's past.

Research method: The author employs the methodological approach called political discourse analysis (PDA), which was introduced and thoroughly explained by Fairclough and Fairclough (2012). According to Whigham, this method was used since it provides an "original methodological contribution to the study of Scottish nationalism".

Data collection: The author focused on parliamentary discourse of the largest political parties in Scotland, namely the pro-independence Scottish National Party (SNP) on the one hand and Scottish Labor Party as well as Scottish Conservative and Unionist Party on the other. The database consisted of election manifestos and policy documents which were related specifically to the independence referendum. Because of a wide range of potentially useful data, Whigham focused primarily on political manifestos and constitutional policy documents. This also allowed for a more detailed analysis of only crucial information about each party's position on the Scottish constitutional debates. The author used the Nvivo qualitative data analysis software package (QSR International Pty Ltd., 2020) which helped him code content of each of the data sources according to the themes that emerged. This was then followed by a coding process which categorized low-level codes into higher-level discursive forms. This sample allowed for a reflection and thorough analysis of political discourse.

Discussion: It needs to be noted that the application of the Nvivo software is an exemplary one and is not frequently observed in historical research. Also, at times the article reads as a sociological one and we believe that it could just as well be classified as such since the author is also a sociologist. However, a more thorough description of the methodological framework would be appreciated.

3.4 Content Analysis

Content Analysis (CA) primarily focuses on studying and analyzing society and social life by examining the content of the visual and textual media – texts, images, and other media products. Mihailescu (2019) understands it as a research technique for making replicable and valid inferences from data to their contexts which is particularly

useful in humanities and social sciences. It is a methodological approach which can help in development of the deductive and inductive capacities, which are extremely important in historical research. In addition, it is highly useful in historical research where researchers are analyzing data with large amounts of text and where meaningful information need to be extracted from the historical documents.

Since CA frequently intertwines both qualitative and quantitative approaches, it sometimes comes close to a mixed methods. CA can sometimes be mistaken for Discourse Analysis since the two methods are very similar. Although both are interested in providing the context of an event, the main difference between the two is that CA focuses on the content of the text, whereas DA focuses on the language that is used in text and context.

3.4.1 Constructing the Child in Need of State Protection

Research problem: In this article, Smith (2016) explores the development of the discourse surrounding children in need of a state protection in Ireland. She mostly focuses on the discourse produced by legislators and government ministers who are ultimately responsible for child services.

Research method: The author employs content analysis of various bills as well as parliamentary debates. She defines it as a textual analysis, but we regard it as a content analysis since she focuses mainly on the content on bills and debates.

Data collection: The author focuses on a specific timeframe in Irish history, namely between 1922 (the formation of the Irish Free State) and 1991 (the adoption of current legislative framework for children welfare). The data consists of debates from both houses of Irish parliament – the House of Deputies and the Senate. In addition, Smith also focused on the official reports which informed these debates. In one part of her research, she focused on parliamentary debates on the Children Bills of 1928 and 1940 and Cussen Report from 1936. In the second part, she conducts the analysis of the Kennedy Report (1970), the Final Report of the Task Force on Child Care Services (1981) as well as the parliamentary debates on the Child Care Bill of 1988.

Discussion: The author dedicates only one paragraph to explicating where the data was taken from in addition to only briefly mentioning the method she used. We consider this to be one of the major shortcomings in this article since it would be useful to know how the textual analysis was performed, what the author focused on and why, as well as what was her motivation for focusing on those specific bills and debates.

3.4.2 Clientelism in Irish Politics

Research problem: The main aim of this article is to research the emergence and development of discourse which revolves around the concept of clientelism in Irish politics. Kusche (2017) focuses on the analysis of the relationship between Irish deputies and voters, which has been perceived as particularly clientelist.

Research method: Kusche identifies the main method of her research as qualitative content analysis. She shows that although this is a historical research, it does have certain methodological features of the sociological research, since sociology also frequently employs content analysis as the main methodological framework.

Data collection: The author draws on parliamentary speeches as well as newspaper articles in order to research the emergence and development of the Irish political clientelism and its critique. This empirical material was deliberately chosen since it is made continuously available throughout the decades and is easy to access. She gathered the parliamentary data from the official website of the Irish parliament and media data form online archives of the respective newspapers. She opted for data from two of the most frequently read Irish quality papers, namely the Irish Independent and the Irish Times. The first step of her data collection consisted of keyword search of the words "clientelism" and "brokerage" in both parliamentary and newspaper records. After realizing that the two words had not been used until the 1980s, she identified other potentially relevant terms based on their emergence in items referent to the two keywords. This produced several other keywords such as "stroke politics", "gombeen politics", etc., which the author used to find relevant data. The period of her analysis runs up to 2012 and starts in the early 1940s. She notes that in the case of parliamentary records, the unit of her analysis is the contribution of the member of the House of Deputies or the Senate; this can either be a speech or a short intervention. In the case of newspaper articles, the unit of her analysis is an article itself. The respective units were then coded according to their focus and since some units included several views of the matter, they were coded in more than one category. Then those articles and debates which specifically focused on the link between politicians and voters were selected for a more detailed interpretation.

Discussion: This article falls under the category of historical social research and employs methodological approaches which are frequent in both historical and sociological research. She gives a detailed account of the method and data she used and where this data was taken from, which is not always the case in historical research. As seen in some of the previously reviewed articles, the author combined parliamentary and newspaper data so as to address the concept of clientelism in as much detail as possible.

3.5 Mixed Methods

Shorten and Smith (2017) understand the mixed methods approach as drawing on the strengths of both qualitative and quantitative methods, which results in showing a more complete picture of a research problem. It is a highly complementary approach, which means that the results produced by one of the methods, can be elaborated and clarified with the findings from the other method. This means that triangulation of one set of results influences and enhances the validity of inferences. In addition, the combination of different methodologies, approaches, and various fields of research adds to the validity of the research and eliminates the possibility of research bias. Thies (2002) shows that as in many other disciplines (sociology for example), investigator bias as well as unwarranted selectivity of the use of historical source materials are the

main problems of qualitative historical research which emphasizes the importance of the selection of the appropriate methodological approach.

Corpus-Assisted Discourse Studies (CADS) combine qualitative Discourse Analysis with the predominantly quantitative corpus-based approach. The main aim of the CADS is to facilitate understanding from the linguistic perspective as well as from that of humanities and social science. As Partington (2012) shows, this approach uses corpus techniques to investigate a particular political or institutional discourse type and to uncover and analyze obvious patterns of language or aspects of linguistic interaction.

3.5.1 Scottish Political Rhetoric in Invasion of Iraq

Research problem: Elcheroth and Reicher (2014) conduct a systematic analysis of the Scottish debate over the invasion of Iraq in 2003. The aim of their article is to show, on the one hand the development of the debates in Scottish Parliament and conduct the analysis of parliamentary discourse of anti-war Scottish separatist parties, and on the other to examine how the conflict was construed as either for or against national interest.

Research method: The authors employ a mixed-methods approach and used thematic coding. This on the one hand produced structured inventories of arguments which served as the grid for qualitative analysis, and on the other, it produced a database which was then used for content analysis.

Data collection: The data for the analysis consist of all the contributions to four Scottish parliamentary debates referring to the Gulf War. A total of 106 interventions which occurred between January 2003 and June 2004 was used as a dataset. It needs to be noted that during the time of 2003 Gulf War, there was also the campaign for election to the Scottish Parliament which meant that the election debate was definitely influenced by the war debate. Each individual intervention was separately coded to extract the information such as which debate the speech was taken from, what was the party membership of the speaker, what was the overall moral argument and so on. Special emphasis was put on the two parliamentary debates which occurred right before the invasion of Iraq (January and March 2003) as well as on first two substantial parliamentary debates that took place after the invasion (November 2003 and June 2004). The transcripts of these debates were all published in the official records of the parliament, and they constituted the "corpus" data for their further analysis. When determining relevant data, all the transcriptions were read several times and coded for those interventions that included arguments that were thematically fitting for the analysis. The two pre-invasion debates produced 68 relevant interventions whereas the two post-invasion debates produced the remaining 38.

Discussion: This article consists of two separate studies. The first study is the analysis of parliamentary speeches, whereas in the second part, the authors turn from elite discourse to popular understanding of the war. The second part draws on the data from Scottish Social Attitude (SSA) survey and since it does not focus on the parliamentary discourse, only the first study was of interest for us.

3.5.2 Political Discourse of Israeli PMs between 2001 and 2009

Research problem: The aim of this article by Gavriely-Nuri (2013) is to look critically at the uses of collective memories in Israeli politics. Collective memories are of great significance to the case of Israel due to their historical background and this article analyzes how collective memories are used within the corpus of speeches of Israeli Prime Ministers.

Research method: The author employed a methodological approach which incorporated both, Critical Discourse Analysis (CDA) and corpus linguistics. Because of the combination of corpus linguistics and discourse analysis, we regarded this article as using the approach called Corpus-Assisted Discourse Studies (CADS).

Data collection: The data used for this study consisted of speeches of Israeli Prime Ministers, over a period of 9 years (between 2001 and 2009), which were delivered in the Israeli Parliament (Knesset). The author conducted a computerized search in the speech archive that includes addresses of the PMs and constructed a corpus, which was then used as a database. The corpus included speeches by the two selected Prime Ministers, namely Ariel Sharon (2001 – 2005) and Ehud Olmert (2006 – 2009). Her computerized search revealed 274 instances of the word "memory", which was determined as the keyword to identify relevant speeches. All those references were then carefully studied and read in order to determine the context. This resulted in identifying 103 references of the phrase "collective memory" which were distributed among 64 speeches. In this count, the author also included synonyms such as "national memory", "public's memory", "people's memories", etc. Once the data was broadly selected, the author performed a two-stage analysis to determine the actual topics of the speeches. In the first stage, the context in which national events evoked the mention of collective memory was analyzed. In the second stage, specific content included in the mentions was studied.

Discussion: Although the author mentions the Cultural approach to the CDA, she gives no detailed account on how this approach differs from traditional CDA or what its benefits are. One of the possible justifications for employing cultural approach is the study of cultural context of the PMs' speeches and their cultural significance. We also found that in the article there is no explicit elaboration as to why this particular methodological framework was selected and how it contributes to the overall analysis.

3.6 Digital History and Topic Modelling

We have shown that historians gather their data mostly from historical archives and feel "much more confident when using traditional sources in printed format, since they believe to have better access to the historical data required for their research" (Torou, 2009). Guldi (2019) believes that digital methods can help researchers land material for historical synthesis that "builds upon the insights of foregoing historians while potentially illuminating new directions for further research". Some authors (Piersma et al., 2014) regard these methods as the Digital Approach or Digital History, the main function of which is to enable historians use advanced search engines in order to explore large quantities of data.

Topic modelling is capable of scanning a large set of documents within which it detects word and phrase patterns and automatically clusters them into groups according to their meaning. As Guldi (2019) shows, topic modelling has been effectively used in history to identify patterns of historical interest in academic sources and has in combination with discourse studies proven to be useful for historical analysis.

Today, several software packages exist which can be used in a pre-existing database of digitalized texts. This, and the fact that digital methods, such as text mining and topic modelling are becoming increasingly used in historical research of parliamentary discourse, underlines the importance of digitizing historical parliamentary records, and not only enable but also encourage historians to start using them as one of the primary sources of data for their research.

3.6.1 Topic Modelling and Historical Change

Research problem: The aim of Guldi's (2019) article is to research the parliamentary discourse on 19th century British empire infrastructure projects, such as the drainage of the River Shannon in 1860, as well as parliamentary argument of the telegraph connection between England and India.

Research method: The author uses dynamic topic modelling which allowed her to generalize about the discourse on a diachronic dataset, observing trends in different time periods.

Data collection: The data for her research consisted of parliamentary debates in the British parliament in 19th century, gathered from Hansard, the official database of all UK parliamentary debates. The author focused on several topics connected to the infrastructure and employed approximately the same data collection and analysis in all of them. The entire Hansard database was subjected to topic modeling, resulting in a set of words used by MPs most indicative of their discussions a certain topic. The author experimented with using on the one hand debate as a document and then also a speech as a document. In addition, she also experimented with degrees of granularity for analysis, asking the computer to return either 4, 10, 100, 500 or 1000 topics. She obtained most informative results with 500 topics as the search returned fairly specific words which were interesting for further analysis.

Discussion: Guldi shows how topic modelling can be implemented into research and analysis of historical data. Topic modelling is becoming increasingly popular in historical research and is frequently used not only on national but also international level (e.g., when researching debates in the European parliament). It is important to note that topic modelling must always be complemented by an objective analysis and critical skills of the researcher when interpreting the results of topic modelling.

4. Discussion and Conclusion

This literature review shows the most common methods and approaches that (political) historians use in their research of parliamentary discourse as well as tries to understand what kind of data and information historians are looking for and which sources they use. We can confirm observations from Torou et al. (2009) that either printed or digitized sources of primary and secondary documents are used in historical research, with digitized sources (transcriptions, text documents, and corpora) becoming increasingly more popular. This makes historical research community a potentially important user group of the ParlaMint corpora.

This literature review shows how the majority of researchers of political history collect data on their own, using techniques and methods which are often time-consuming and demand a lot of manual work. Such work could be made much more research-friendly and efficient if historical parliamentary corpora were developed, annotated, and documented. They would present a database of collected parliamentary records of the past and would be a useful source of historical parliamentary records which would be an invaluable extension of the ParlaMint project.

Our first aim should therefore be to provide historians with tutorials, workshops and showcases on how to use corpora, corpora data, and the main corpus-analytical techniques. Rich and user-friendly documentation on how the ParlaMint data is gathered, processed, and annotated is to be made available to the historians in addition to offering quick user manuals which would show the basic use of concordancers for historians to learn how to effectively use corpora.

Then, we should encourage them to develop and use their own corpora and datasets for the historical periods they are interested in using the same encoding standards. In this endeavor, we agree with Kytö (2010) that compilers of the data should document their compilation decisions in clear terms in user guides, corpus manuals, and training materials which need to accompany the release versions of the corpora, since it would be impossible for end-users to find information about the background of the texts which are included in the historical corpora without them.

The ParlaMint community should also focus on the implementation of the tagging of the digital repository contents with complete and structured metadata. Some historians (Torou et al. 2009) note that the information which is typically used in research queries by historians (such as the author, topic of the item, date of creation, the period to which the content refers, etc.) should be available as metadata. The availability and reliability of metadata is extremely important since historians often rely on the additional data and information about a certain historical source.

Marjanen (personal communication, 2022) points out that historians researching parliamentary discourse are highly interested in the use of rhetoric, the uses of voice and practices of negotiation and debate. One of the key interests for them is identifying who talked, which makes the availability of any metadata about the MPs of vital importance. According to Marjanen, there are also some historians who together with traditional sources, use audio and video recordings from parliament to study non-verbal elements in parliamentary discourse. He points out that with digitized sources, keyword search has made material much more accessible though many historians are often interested in something broader than keyword search. They focus on the entirety of speeches or discourse related to a certain topic since keyword search often does not produce enough relevant results. Historians are used to finding these

"discourses" on their own but if the process of searching for relevant sources was made easier for them, it would definitely be welcomed.

An increasing availability of the digitized sources appears to be setting an interesting trend. In addition to more and more sources and documents becoming digitized and made available through electronic libraries, various digital research tools and approaches are becoming available, making historical research often very digital. Therefore, political historians nowadays already employ digital approaches and tools to analyze parliamentary data and these approaches allow them to gather and analyze data in a faster, more efficient, and less time-consuming manner. However, the development of parliamentary historical corpora could potentially reshape the entire process of historical research and offer new understanding of the parliamentary data. As Blaxill (2013) shows, the combined approaches of close manual analysis and selective quantification simplify the research as well as facilitate numerical comparison and contextualization.

The argument we want to put forward with this literature review is not that current qualitative historical research of political debates and parliamentary discourse should be completely replaced by more quantitative corpus-assisted approaches, but rather that corpora could be effectively used alongside the traditional qualitative historical analysis. We treat corpora as potentially powerful tools which would not only simplify data collection and generate relevant results much more effortlessly, but also effectively reduce and minimize potential research bias that might be present in the analysis of historical data.

This review also shows the need for more systematic, transparent, and replicable quantitative and qualitative analysis, which makes corpus-assisted approaches ideally suited for historical research of parliamentary discourse. The immediate usefulness of the ParlaMint corpora is also clearly confirmed by this review and it emphasizes the need for further enrichment and the addition of the historical data to the current ParlaMint database.

5. Acknowledgements

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), the Social Sciences & Humanities Open Cloud (SSHOC) project (https://www.sshopencloud.eu/), the CLARIN ERIC ParlaMint project (https://www.clarin.eu/parlamint) and the DARIAH-SI research infrastructure.

6. References

Mariana Achugar. 2017. Critical discourse analysis and history. In J. Flowerdew and J.E. Richardson (Ed.) *The Routledge Handbook of Critical Discourse Studies*, Vol. 1, pages 298-311. Routledge, London.

Luke Blaxill. 2013. Quantifying the language of British politics, 1880–1910. *Historical Research*, 86(232): 313-341. https://doi.org/10.1111/1468-2281.12011

Constance De Saint Laurent. 2014. "I would rather be hanged than agree with you!": Collective Memory and the Definition of the Nation in Parliamentary Debates on Immigration. *Critical Practice Studies*, 15(3): 22-53. http://dx.doi.org/10.7146/ocps.v15i3.19860

Patricia L. Dunmire. 2012. Political discourse analysis: Exploring the language of Politics and the Politics of language. *Language and Linguistics Compass*, 6: 735-751.

Guy Elcheroth and Steve Reicher. 2014. 'Not our war, not our country': Contents and contexts of Scottish political rhetoric and popular understandings during the invasion of Iraq. *British Journal of Social Psychology*, 53: 112-133. https://doi.org/10.1111/bjso.12020

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, et al. 2022. The ParlaMint corpora of parliamentary proceedings. Lang Resources & Evaluation. https://doi.org/10.1007/s10579-021-09574-0

Anna Friberg. 2012. Democracy in the Plural? The Concepts of Democracy in Swedish Parliamentary Debates during the Interwar Years. *Contributions to the History of Concept,* 7(1): 12-35. http://dx.doi.org/10.3167/choc.2012.070102

Dalia Gavriely-Nuri. 2013. Collective memory as a metaphor: The case of speeches by Israeli prime ministers 2001–2009. *Memory Studies*, 7(1): 46-60. https://doi.org/10.1177%2F1750698013497953

Jo Guldi. 2019. Parliament's Debates about Infrastructure: An Exercise in Using Dynamic Topic Models to Synthesize Historical Change. *Technology and Culture*, 60(1): 1-33. https://doi.org/10.1353/tech.2019.0000

Patricia Hogwood. 2013. Selective memory: challenging the past in post-GDR society. In: Saunders, A. and Pinfold, D. (Ed.) *Remembering and Rethinking the GDR*, pages 34-48. Palgrave Macmiallan, London.

Pasi Ihalainen and Satu Matikainen. 2016. The British Parliament and Foreign Policy in the 20th Century: Towards Increasing Parliamentarisation?. *Parliamentary History*, 35(1): 1-14. https://doi.org/10.1111/1750-0206.12180

Pasi Ihalainen, and Taina Saarinen. 2019. Integrating a Nexus: The History of Political Discourse and Language Policy Research. *Rethinking History*: 1-20. https://doi.org/10.1080/13642529.2019.1638587

Pasi Ihalainen. 2021. Parliaments as Meeting Places for Political Concepts. Centre for Intellectual History, University of Oxford. https://intellectualhistory.web.ox.ac.uk/article/parliaments-as-meeting-places-for-political-concepts

Isabel Kusche. 2017. The Accusation of Clientelism: On the Interplay between Social Science, Mass Media and Politics in the Critique of Irish Democracy. *Historical Social Research*, 42(3): 172-195. https://www.jstor.org/stable/44425367

Marja Kytö. 2010. Corpora and historical linguistics. *Revista Brasileira de Linguística Aplicada*, 11(2): 417-457. http://dx.doi.org/10.1590/S1984-63982011000200007

Daniel Litte. 2016. What is "conceptual history"?. Understanding society. https://understandingsociety.blogspot.com/2016/10/what-is-conceptual-history.html

Jani Marjanen. Personal communication. By Jure Skubic, 27 May 2022.

Conor McGrath. 2018. British Lobbying in Newspaper and Parliamentary Discourse, 1800–1950. *Parliamentary History*, 37(2): 226-249. https://doi.org/10.1111/1750-0206.12363

Mimi Mihailescu. 2019. Content analysis: a digital method. http://dx.doi.org/10.13140/RG.2.2.21296.61441

Ciprian Negoita. 2015. Immunity: A Conceptual Analysis for France and Romania. *Contributions to the History of Concepts*, 10(1): 89-109. http://dx.doi.org/10.3167/choc.2015.100105

Alan Partington. 2012. Corpus Analysis of Political Language. In: C.A. Chapelle (Ed.) *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd.

Hinke Piersma, Ismee Tames, Lars Buitinck, Johan van Doornik and Marteen Marx. 2014. War in Parliament: What a Digital Approach Can Add to the Study of Parliamentary History. *Digital Humanities Quarterly*, 8(1): 1-18. http://www.digitalhumanities.org/dhq/vol/8/1/000176/000176.html

Allison Shorten and Joanna Smith. 2017. Mixed methods research: expanding the evidence base. Evidence-based nursing 20: 74-75. https://ebn.bmj.com/content/20/3/74.info

Karen Smith. 2016. Constructing the Child in Need of State Protection: Continuity and Change in Irish Political Discourse, 1922–1991. The Journal of the History of Childhood and Youth, 9(2): 309-323. https://doi.org/10.1353/hcy.2016.0042

Jure Skubic and Darja Fišer. 2022. Parliamentary Discourse Research in Sociology: Literature Review. Accepted for publication in *Proceedings of the ParlaCLARIN III workshop at LREC2022,* pages 91-100, Marseille, France.

Cameron G. Thies. 2002. A Pragmatic Guide to Qualitative Historical Analysis in the Study of International Relations. *International Studies Perspectives*, 3(4): 351-372. https://www.jstor.org/stable/44218229

Elena Torou, Akrivi Katifori, Costas Vassilakis, Georgios Lepouras and Constantin Halatsis. 2009. Capturing the historical research methodology: an experimental approach. In Proceedings of International Conference of Education, Research and Innovation, Madrid, 2009. Madrid, Spain.

Teun Van Dijk. 2018. Discourse and Migration. In Qualitative Research in European Migration Studies, edited by Ricard Zapata-Barrero and Evren Yalaz, 227-247. Springer Open. https://link.springer.com/book/10.1007/978-3-319-76861-8

Stuart Whigham. 2019. Nationalism, party political discourse and Scottish independence: comparing discursive visions of Scotland's constitutional status. *Nations and Nationalism*, 25(4): 1212-1237. https://doi.org/10.1111/nana.12535

Annotation of Named Entities in the May68 Corpus: NEs in modernist literary texts

Mojca Šorli,[*] Andrejka Žejn[†]

[*] ZRC SAZU, Institute of Slovenian Literature and Literary Studies
Novi trg 2, SI-1000 Ljubljana
mojca.sorli@zrc-sazu.si

[†] ZRC SAZU, Institute of Slovenian Literature and Literary Studies
Novi trg 2, SI-1000 Ljubljana
andrejka.zejn@zrc-sazu.si

Abstract

In this paper we present the process of manual semantic annotation of a corpus of modernist literary texts. An extended set of annotations is proposed with respect to the established NER-systems and practices of related projects, i.e. several categories of proper names, foreign language elements and bibliographic citations. We focus on the annotation challenges concerning the names of literary characters seen in transition from common nouns to proper names, as well as giving examples of the results of preliminary analyses of the corpus.

1. Introduction

The starting point of the digital humanist literary project presented here is a corpus of literary texts that was created according to special criteria defined for the purposes of this research. In view of the significance for DH of controlling a large number of texts and their vertical reading, where patterns become visible that cannot be detected with the naked eye or traditional close reading, the corpus size is often seen as a key factor. At the same time, large text volumes require automation of corpus processing for quantitative analysis, involving different levels of (linguistic) annotation in the first phase, and allowing additional levels of semantic annotation in later phases that enrich the text with metadata. In the presented approach, however, the annotation task is performed on a small, specialized corpus that is easier to control and allows for manual annotation. The identified and manually annotated Named Entities are distinguished based on semantic criteria, so we consider this an example of semantic annotation.

Linguistically annotated corpora have long been a standard tool for linguistic research. Named Entity Recognition (hereafter NER) and analysis has also long been relevant in the social sciences and sociology (Ketschik, 2020), from where the method, like several others, has been transferred via linguistics to literary studies, where named entities are most closely associated with literary character research. A more comprehensive picture of the way characters are named in literature, beyond the automatic recognition of Named Entities (hereafter NEs), can be obtained by manually annotating these entities in literary texts, by analyzing the annotation process, and finally by analyzing the data obtained from the annotated corpus itself.

2. The Goal of the paper

In this paper we report on an attempt to identify and annotate three groups of NEs in the "Corpus of 1968 Slovenian literature Maj68 2.0" (short name May68 Corpus) – corpus of Slovenian modernist literary texts from the late 1960s to the early 1970s,[1] discussing these groups from the point of view of three different sources of representation problems that are independent but interrelated: ambiguity, variation, uncertainty. As pointed out in Beck et al. (2020), representational problems in linguistic annotation arise from five different sources (ibid., 61): (i) Ambiguity is an inherent property of the data. (ii) Variation is also part of the data and can occur, for example, in different documents. (iii) Uncertainty is caused by lack of knowledge or information by the annotator. (iv) Errors may be found in the annotations. (v) Bias is a property of the entire annotation system. We list a number of relevant annotated categories, their specific character, and representational problems associated with them. Our choices are discussed when any of the first three listed sources of representation problems apply.

Together with the theoretical concept, the selection of annotation material, and the definition of guidelines for the annotation process (Pagel et al., 2020), the annotation scheme presented here is a model of extended annotation of NEs in modernist periodicals that can be applied in certain segments to other corpora of literary texts. We focus both on the identified inaccuracies and on the benefits of manual annotation of selected groups of NEs in our specialized corpus of literary texts. In the concluding part, we present the preliminary results of an analysis performed on the annotated corpus.

Following the automatic preprocessing (i.e., POS tagging and lemmatization) of the May68 Corpus, further manual annotation was performed to capture more complex linguistic (semantic) phenomena and to provide a more sophisticated annotation model for proper names given the recurring representational problems: At this first stage, a model for identifying and annotating the selected NEs was put in place, with a second stage of the project envisaged, in which the texts will be annotated for the use of metaphor. Here we will focus on some open challenges in the annotation of NEs, in particular problems related to the functional aspects of the annotated elements. We discuss the practical treatment of proper names for the purposes of corpus linguistic and stylistic research, in the hope of

[1] http://hdl.handle.net/11356/1491

improving the reliability of research results and also of NLP models.

3. Automated and manual annotation of corpora

In the context of language technologies, universal concepts and tools for automatic corpus annotation have been developed to some extent, especially for individual language groups, while language-specific concepts and tools are also needed. Established levels of automatic tagging for Slovenian, initially based on lexicographic and linguistic projects, include tokenization and related segmentation into sentences, normalization, morphosyntactic tagging, lemmatization, and syntactic parsing (Erjavec et al., 2015). NEs pose a challenge for automatic extraction of information due to their semantic an functional complexities. For Slovenian, the main tool used is StanfordNER, which assigns lexical units to predefined categories (Ljubešić et al., 2012): personal names, geographical names and common proper nouns. The state-of-the-art of the existing NER tools for Slovenian has not been the focus of this research, but a preliminary review of the tools, as well as of the function of NEs in the texts, has shown their limited applicability to a specialized literary corpus that we set out to investigate.

3.1. NER-systems for corpora of literary texts

For literary texts, narratology in particular has developed various typologies of protagonists, heroes, or major and minor characters in texts, ways of characterizing them, and strategies for recognizing them. Since the advent of digital tools researchers have had to find a way to translate the definitions formed by literary scholars into computer-readable data (Krautter et al., 2018).

While there are no specific NER-systems for annotating literary texts, even though literary texts have a high variation of NEs compared to normal non-fiction texts (Stanković et al., 2019), "universal" systems are often used. However, automatic annotation tends to overlook certain segments of NEs in literary texts (Vala et al., 2015). Attempts are made to overcome these limitations by additional automatic tagging, or to expand the set of annotated entities by manual tagging, often of referential expressions, i.e., linguistic expressions that refer to a specific entity in the text world, where the entities and their references must be interconnected (entity grounding). References and connections themselves can only be inferred from the knowledge of the context (Ketschik, 2020; Papay and Padó, 2020), so in the early stages of research, manual annotation of the corpus is usually required to improve the automatic process.

3.2. Background and related work

Compiling lists of NEs, especially for categories of proper names, represents only the basis for the identification of character names and is as yet insufficient for relevant literary analyses, so these lists must be dealt with by multidimensional approaches that shed additional light on proper names in light of the special features of literary text. Empirical analyses of protagonists in the literature can, at the most basic level, for example, study the characteristics of names, their typicality, archaic character, or "unusualness" for a particular society (cf. Calvo Tello, 2021), compare usage and functions of proper names, exploring to what extent they are genre-related (e.g. children's literature, cf. van Dalen-Oskam, 2022).

Empirical analysis of the ratio between female and male characters in a corpus of English literature up to the mid-20th century (cf. Nagaraj and Kejriwal, 2022), for example, showed the quantitative predominance of male characters over female characters. More complex research also deals with characterization analysis, identifying relationships between main and secondary characters, examining the relationship between active and passive character presence, and distinguishing between "actively present" characters and characters from other fictional worlds (Krautter et al., 2018; Brooke et al., 2016; Ketschik, 2020). One of the more established approaches is the application of social network analysis, a method from empirical sociology that builds on the relationship between NEs. The analysis of social networks in the literature (cf. de Does, 2017) is closely related to quantitative approaches to the study of direct and reported speech or narrator speech and character speech in storytelling and drama, where NEs are an essential component of a broader context (cf. Burrows, 2004; Moretti, 2011; Elson et al., 2010; Papay and Padó, 2020). Digitally supported analysis of the broader picture of characters also draws on concepts derived from Bakhtin's concept of chronotope, such as The Text World Theory – a cognitive-linguistic concept of a unity of characters, time and space, or the concept of situation (Krautter et al., 2018; Mikhalkova et al., 2019).

4. Model annotation schemes

In designing the model for manual annotation of the May68 Corpus, we relied on familiarity with the texts contained in the corpus and on several other well-known models of manual annotation for similar projects, three of which are presented below.

4.1. COST Action ("Distant reading" project)

Distant Reading project for the annotation of the multilingual ELTeC corpus (https://www.distant-reading.net/eltec/)[2] based on European novel provides the following distinct categories: "demonyms (DEMO), professions and titles (ROLE), works of art (WORK) person names (PERS), places (LOC), events (EVENT), organizations (ORG)" (for a brief description of the categories cf. Frontini, 2020). The selection of these categories was partly motivated by the existing possibilities of automated NER, which brings with it certain limitations (Stanković et al., 2019). The project also points out the importance of "cultural references, role models and cosmopolitanism", and these can only be answered "if references to works of art, authors, folklore and periodical publications are detected", which is why in our corpus of

[2] The Distant Reading for European Literary History (COST Action CA16204) started in 2017 with the goal of using computational methods of analysis for large collections of literary texts. It is based on the compilation and analysis of a multilingual open source collection, named European Literary Text Collection (ELTeC).

modernist literary texts we introduced a BIBLIO group to incorporate references to authors, but covered other listed types of references with the "other" group (NAME / XXX). In May68 Corpus, however, we focus for now on proper names.

4.2. CLARIN.SI

The annotation scheme adopted largely follows the guidelines provided for Slovenian in the past (e.g. Štajner et al., 2013), perhaps closest in its granularity to the Janes-NER guidelines (CLARIN.SI) as described by Zupan et al. (2017), except for the derived adjectives (DERIV-PER) type, which is given here an independent status unlike in May68 Corpus, where this is subsumed under the PER-LIT and PER-REAL subtypes.[3]

In addition, we decided in the case of May68 Corpus to conceptualize combinations of nouns denoting professions, functions or titles, and personal names as units, therefore labelling the entire strings as literary personal name (PER-LIT) or real personal name (PER-REAL).

4.3. Annotation schemes for Czech language

Annotation of NEs in Czech corpora is implemented according to more complex models as described in Sevščíková et al. (2007). Our three-level NE taxonomy is, nonetheless, somewhat less fine-grained. Furthermore, unlike the Czech model, ours does not include numbers, such as in addresses, zip codes, or phone numbers, specific number usages and quantitative expressions – entities typically included in NER.

5. May68 Corpus of Slovenian modernist literary texts – corpus description

The Maj68 Corpus is a result of a project on the literature of the avant-garde and modernism in the period of the worldwide student movement, whose activities are also reflected in the transformation of literature. The student journals *Tribuna* and *Problemi*, from which the texts for the corpus were selected, played an important role in the theoretical and literary-artistic innovations of the Slovenian student movement. The Maj68 Corpus 1.0 contains 1,521 texts by 198 known authors published between 1964 and 1972 in the Slovenian periodicals *Tribuna*, *Problemi* and *Problemi.Literatura*. The Maj68 Corpus 2.0 version, which has been further edited and corrected (metadata), contains 647 additional texts from *Tribuna* and *Problemi*.

The compilation of the corpus began with an extensive bibliographic inventory of texts in selected publications that have been digitized and are publicly available on dLib. On the basis of these lists, the original texts of Slovenian authors were converted from .pdf format to .docx format and, in a second phase, linked to metadata in Excel spreadsheets. Finally, the corpus was automatically tagged (see Juvan et. al 2021 for more details on the procedure).The texts contain complete bibliographic data, are classified by text and language type, degree of presence of non-standard Slovenian, foreign languages, modernism, and visual elements. Author details, i.e., gender and year of birth, are included with the texts. The presence of visual elements is also marked in the corpus; 48 texts consist only of visual elements, i.e. they do not contain standard text.

Automatic linguistic annotation includes lemmas, morpho-syntactic descriptions from MULTEXT-East, and morphological features and syntactic annotations from Universal Dependencies. As shown here, manually tagged NEs for persons, geographical locations, organizations, and various names, (foreign) linguistic variations and registers, and cited authors (sources) are additionally marked.

The following sections and subsections introduce the types and categories of NEs, including the dilemmas encountered in the process of annotation and the practical reasons for annotation. From here on, and with a somewhat narrower notion of NER, we speak of categories of "proper names (personal and place names)" rather than "named entities" for the purposes of this paper.

5.1. Annotation procedure and categories

The annotation was implemented using the WebAnno tool (Eckart de Castilho et al., 2016). To simplify the technical aspect, the whole corpus was divided into 1529 sections of five sentences each, on average 380 chunks per section. WebAnno allows annotation of one sentence at a time, which was a disadvantage for longer instances of text marked by the use of foreign language(s). Each annotation round was curated by two curators.[4] However, reiterative annotation was not foreseen, since the primary goal at this stage was not to improve automatic annotation, but to manually annotate the specialized corpus for optimal corpus analysis and stylistic studies.

There is no universally accepted taxonomy for NEs, except for some coarse-grained categories (people, places, organizations). Since we are interested in a semantically oriented annotation and prefer more informative (fine-grained) categories, we opted for a three-level NE classification as shown in Table 1 (cf. Sevščíková et al., 2007). The first level in our annotation model corresponds to the three basic groups: 1. Proper names, 2. Foreign language and register variations, and 3. Cited authors. These groups are labelled as 1. NAME, 2. FOREIGN, 3. BIBLIO respectively, with the first two further subdivided. The second and third levels provide a more detailed semantic classification.

The NAME group includes the following types and subtypes:

- Person (PER), including the person-derived adjective, is subdivided into fictional literary characters (PER-LIT), characters referring to real, i.e., existing and historical or mythological, persons or beings (PER-REAL), literary characters bearing a descriptive name (PER-DES), and members of national and social groups (PER-GROUP).
- Geographical location (GEO) is divided into locations in Slovenia (GEO-SI), in former Yugoslavia (GEO-YU), in Europe (GEO-EU), and in others (GEO-ZZ).
- Organizations and institutions (ORG).
- Miscellaneous (XXX).

A group labelled FOREIGN is used to annotate the foreign language: Serbo-Croatian (SBH), English (EN),

[3] Overall and in the same fashion, in May68 Corpus we also favour larger lexical units.

[4] The texts were annotated by A. Jarc, L. Mandić, and K. Žvanut in accordance with the annotation scheme designed by the authors of this paper, who also curated all of the annotations.

French (FR), Italian (IT), Latin (LA), and German (GE), or register variation (DIALECT, INFORMAL, SLANG) in the corpus.

Once the annotation process was completed, the labels in WebAnno were converted to TEI encoding.[5] Following the conversion thus all proper names (personal names, place names, names of organizations, and real names) are labelled with <name/>, then divided into types with @person, @geo, @misc, @personGrp, and @org attributes, three subtypes for literary characters (@literary, @descriptive, @real), and for geographical names (@SI, @EU, @ZZ and @YU). Units of text with foreign languages and non-standard Slovenian were labelled as <foreign/> and corresponding attributes according to TEI coding.

5.1.1. Person

PERSON (PER) type is divided into PER-LIT, PER-REAL, PER-DES and PER-GRP. While the first three are categorized as subtypes of the same type, PER-GRP is defined as an independent type. The most important subdivision of the type (within the NAME group) is that between real, e.g., historical or real-life, persons appearing in the text, and fictional characters, each of which, however, is further specified according to semantic criteria. Subcategories include names of people and pets, nicknames, pseudonyms, members of national and social groups.

Group	Type	Subtype	Description
NAME	PERSON (PER)	PER-REAL	Real: Characters referring to real, i.e. existing and historical or mythological persons or beings (web sources, Wikipedia, etc.), e.g. *Greta Garbo*.
		PER-LIT	Literary: Fictional literary characters, e.g. *Ančika*, *Zobec*.
		PER-DES	Descriptive: Literary characters that carry a descriptive name (e.g., *dolgolasec*, Eng. the long-haired guy)
		PER-GRP	Group: Members of national and social groups, e.g. *Kranjci*, *Slovenec, Američan*.
	GEO	GEO-SI	Slovenia, e.g. *Ljubljana*
		GEO-YU	Former Yugoslavia (except for Slovenia), e.g. *Zagreb*
		GEO-EU	Europe, e.g. *Frankfurt*
		GEO-ZZ	Other, e.g. *Peking*
	ORG	–	Names of organizations, institutions (*Klub nepismenih*, *Slovenska matica*, *Državna varnost*)
	XXX	–	Common proper nouns, including titles of books and other art works, artefacts, etc., e.g. *Rdeča kapica*, *Empire State Building*.
FOREIGN	HBS	–	Serbo-Croatian
	EN	–	English
	DE	–	German
	FR	–	French
	IT	–	Italian
	LA	–	Latin
	XX	–	Other
	DIALECT	–	Dialect
	VERNACULAR	–	Vernacular
	SLANG	–	Slang
BIBLIO	–	–	Quoted authors (Sources)

Table 1: The main categories of the May68 annotation scheme (WebAnno).

PER-REAL denotes both real, i.e. existing, persons and historical or mythological figures that are basically identifiable in encyclopaedic sources such as online lexicons of proper names, Wikipedia and the like. URL is an additional attribute of the NAME group and is given as a relevant source of information, e.g., a website, for a group of people appearing in the literary text. The assignment of a URL depends on context or extra-linguistic knowledge; if a person can be assumed to be part of common (cultural) knowledge (Descartes, Nietzsche), we do not enrich the corpus with encyclopaedic data.

All standard personal proper names are labelled as NAME and assigned to one of the closed subtypes.

The label PER-GRP with no subtype is assigned to members of a particular social group, most often nationality (Slovenec), regional identity (Kranjci, Štajerci; Novakovi), but also smaller social groups defined on the basis of occupational or other criteria.

Of the categories introduced specifically for the purposes of the May68 Corpus, NAME / PER-DES proved, as expected, to be the most challenging subcategory (see 6.1.1.).

Given their statistical importance in the context of NER, the same annotation rules apply here as for characters in plays when they do not require special treatment with respect to their function. The labelling of proper names in plays depends on the status and/or function of the proper name. Names of individual characters that merely announce an individual character's speech, his/her lines of dialogue, have not been annotated, while names in descriptions of their physical actions or behaviour are treated as ordinary proper names on the model of "sb does sth" etc. (*Pandolfo se ogleduje v zrcalu* / Pandolfo looks at himself in the mirror). Below is an example of a dialogue showing the distinction between the two and a third subtype (the names in bold are labelled as PER-LIT, PER-DES and PER-REAL respectively):

[5] The annotation task was carried out in collaboration with T. Erjavec (technical aspects and data conversion).

BARRÈRE: *Potemtakem moramo danes z njim obračunati.* (***Tallien*** *odide*)
(Davidu): *Si pripravljen s* ***Krepostnim*** *umreti*?
DAVID: *V smrt?*
BARRÈRE: *Se nisi maločas naglas pridušil?*
DAVID: *Čudovit črtež sem zamislil. Kako dviga* ***Sokrates*** *čašo strupa k ustom. Naš dobri prijatelj je tako presunljivo govoril.*

Adjectives derived from personal proper nouns are annotated as the corresponding proper nouns. Their derived character is revealed by morpho-syntactic tagging.

5.1.2. Geographical location (GEO)

Place names are labelled as NAME and the following closed-set subtypes: SI, YU, EU, ZZ, depending on whether the location is in Slovenia, in the former Yugoslav republics, in the rest of Europe, or outside all of these areas.

As with personal names, a distinction is made between real and fictitious geographical names (*Indija* vs. *Eldorado*). Commentators decide whether a place is real or fictitious (such as street names in a fictitious city) based on context and common knowledge. Places typically include continents, countries, regions, cities, towns, and natural geographical objects, as well as streets, squares, and neighbourhoods, and functional infrastructure such as churches, airports, and local cultural and natural sites. Place names used metaphorically, e.g. *Eden*, are categorized as "other" and assigned the label NAME / GEO-ZZ – the same label is used for place names outside the European territory. At this stage, we have not paid special attention to the treatment of proper names (personification) used metaphorically, such as

Jadra so pogorela, ***Delfi*** *molčijo …* [The sails have burnt down, and *Delfi* stays silent …]

This type of analysis is planned for the later stages of annotation (which will include the annotation of metaphors).

Adjectives derived from place names, e.g. African, European, were included in the annotation by analogy with geographical names and divided into the same subtypes (SLO, YU, EU, ZZ).

5.1.3. Organizations and common proper nouns

As with geographical names, there are no subgroups for the two groups of so-called common proper names and names for organizations. Capitalization is an obvious but not a necessary condition for this classification. Thus, no distinction is made here between real and fictitious; what matters is that the name be recognized as "common proper" in the literary context of the text.

Organizations and institutions subsume names of museums and other cultural institutions, as well as political and civic organizations. Organizations are labelled as ORG and usually include businesses, institutes, media, cultural, and educational institutions. However, we have treated restaurants, music groups, and other "entertainment" establishments as "miscellaneous" rather than organizations.

Miscellaneous is a category reserved mainly for common proper nouns, as explained above, such as titles of books and other works of art, artefacts, films, documents, brand names, commercial products, events, including place names, such as mythological places, place names used metaphorically, etc. These NEs are labelled as XXX.

For many common nouns, one can observe a transition to the category of proper names, which seems to exist as a continuum. For example, the word *krčma* (Eng. inn, pub) assumes the function of a proper noun referring exclusively to a particular unit/object, in this case "inn". The word is therefore referred to as NAME / XXX.

5.1.4. BIBLIO

BIBLIO is typically used for literary works cited or mentioned in the literary texts. It contains text passages that refer to literary works or other bibliographic units, and is annotated for authors, not titles or citations, e.g.

The patamus can never reach The mango on the mango tree ***(T. S. Eliot****: The Hippopotamus)*

5.1.5. Language and register

In the case of language and register variation, we use the FOREIGN group that subsumes (foreign) language and register variation (see Table 1). This group is not directly relevant to this paper.

6. Dilemmas of annotation in the framework of representational problems

A number of dilemmas are discussed here in terms of the three categories – ambiguity, variation, and uncertainty – as detailed, for example, in Beck et al. (2020), who outline the main representational problems in linguistic annotation (we disregard the two additional categories addressed in the model: error and bias).

The interpretation of the listed categories is tailored to the nature of our data, and the problems are assigned to the listed categories accordingly. The annotation process is consistently guided by the identified function of the annotated elements. The three dilemmas are described below.

6.1. Ambiguity

In principle, ambiguity occurs whenever a unit admits several interpretations. Ambiguities between form and meaning occur in natural language at the phonological, morpho-syntactic, lexical, or pragmatic levels and are a major source of representational problems (Beck et al., 2020).

6.1.1. Transition from personal proper names to "common proper nouns"

The most striking example of ambiguity is the transition from common nouns to those that function as personal names. This is a pervasive and rather complex representational problem. The dilemma concerns the category NAME / PER-DES, i.e., descriptive names of literary characters, especially in relation to the category NAME / PER-LIT, which refers to standard proper names that are recognizable as such because of their form and conventional properties (e.g., capitalization). This group includes examples where common nouns optionally combine with proper names to refer to individual characters like "inšpektor (Kos)" [inspector (Kos)], or "veteran" [the veteran], including capitalized adjectival derivatives, such as "Brezposelni" [The jobless one], functioning as personal names, etc.

However, capitalization is not a necessary condition for the NAME / PER-DES designation, especially in a corpus of modernist texts that frequently employ modernist and/or

idiosyncratic conventions, with orthographic rules applied to proper names or descriptive linguistic units that typically eschew capitalization (e.g.,"fant" [the boy], "starka" [the old woman]). A key feature of proper names, as it turns out, is "descriptive continuity," which shows that there is no clear boundary between what can be considered a standard proper name (which is traditionally subsumed under onomastics) and what can be understood as an instance of a text that performs the function of a proper name, but does not, strictly speaking, qualify as such.

The assignment of a noun to NAME / PER-DES is decided primarily on the basis of context. Often, a lexical unit (word or phrase) is used to describe a particular property of the character to which the proper noun initially refers, and which is then gradually but clearly transformed into a (descriptive) unit that functions as a proper name (whether capitalized or not), such as "Rdečelasi" [The red-haired one]. The descriptive name is used only when the transition is complete, which must be evident from the broader context. The quantitative criterion (in longer texts) is a minimum of three occurrences of the same designation, such as below:

> Videl je same znane obraze — *inšpektorja Kosa, vratarja Žorža, kurirja Enorokega, Žana*, nekoliko v ozadju pa je stal bledi ***Novinec*** [the (pale) new guy], …

Other examples include *dolgolasec* [the long-haired guy], *mladenič* [young man], *mojster* [the master], *debelušček* [the fatty] and typically correspond to phrases introduced with a definite article in English. In principle, PER-DES is not limited to a maximum number of components, but the likelihood that a lengthy description, such as Zagledal je na tleh sedečega *fanta upadlih lic in kuštravih las* [He saw *a boy with skinny cheeks and messy hair* sitting on the floor], should appear three times at least in the text(s) is minimal. Even if descriptive units tend to recur they normally vary in at least one of their elements.

Capitalization itself does not preclude a lexical unit from being labelled PER-DES, as with *Mož brez imena* [the Nameless Man].

Appellatives, nicknames, and pseudonyms are labelled as ordinary personal proper names (NAME / PER-LIT), except for those expressing description, such as *Dolgi Džon* [John the Longish].

6.1.2. Nesting

Another example of ambiguity concerns nesting, which often creates additional annotation problems. Instead of a potential two- (or three-level) nesting model, a single-level nesting is used throughout, taking as the basic annotated unit the largest possible lexical unit, typically a geographical name or the name of an organization composed of one or more proper names: in the case of *Državna založba Slovenije* [National Publishing House of Slovenia], the entire unit is labelled as an organization (ORG) and the proper name *Slovenije* is not nested and labelled on its own as a place name (*Slovenija*); the same goes for for *Društvo novinarjev Slovenije* [Journalists' Association of Slovenia], *Prešernova družba* [Prešeren's Society Publishing], *Direkcija za prehrano Beograd* [Belgrade Food Agency], or, e.g. *Fani* is NOT nested in *gospodična Fani*, but treated as a single-level personal proper name. A general dilemma often arises here as to whether the term should be referred to as a proper name or as a common noun.

6.2. Variation

In variation, the same content or value is expressed by multiple, interchangeable variants (Lüdeling, 2017). Variation can be due to extra-linguistic factors, such as the time period, genre, author/speaker of the text, or linguistic conventions.

Like ambiguity, variation is an inherent part of natural language and thus of corpus data. Indirectly related to variation is the case of ambiguity described above in 7.1.1. The descriptive name is not necessarily used exclusively for one and the same literary character; on the contrary, it usually alternates with the character's actual proper name.

Alternation in the mention of literary characters is very common; in fact, it is the rule. Some personal proper names (including their descriptive variants) occur as variants preceded by an attributive noun (always the same), usually referring to their professional or social status (e.g., Inspector Kos). When this type of designation is used consistently, we refer to the entire lexical unit as NAME / PER-LIT, but when the attributive noun (Inspector) becomes an independent descriptive variant, we refer to it as NAME / PER-DES.

Descriptive terms NAME / PER-DES may consist of one or more words, they may be a combination of "object nouns" and standard proper names (*inšpektor Kos*) or of two or more "common nouns" (*kurir Enoroki*), regardless of their capitalization, as long as they function as personal proper names when referring to or naming characters. The same character may be referred to by three, four, or more variants. In our case: inspector Kos, inspector or Kos.

Also treated as single variants are lexical units denoting proper names whose capitalization varies, e.g., *Ministrstvo za kulturo Republike Slovenije* vs. *ministrstvo za kulturo* (Ministry of Culture) and *Zveza borcev* vs. *zveza borcev* (Association of Freedom Fighters).

We are aware that when variants are expressed as a single interpretation, the property of variation as a whole is lost. However, a semantic annotation based on the function of linguistic elements is less prone to structural diversity than, for example, spelling variations in historical texts that reflect, for example, dialectal and/or temporal differences (cf. Beck et al., 2020), which is why, apart from our own specific research goals, we did not choose to preserve (proper name) variations.

6.3. Uncertainty

Uncertainty arises whenever there are multiple possible interpretations of data, but the relevant or reliable knowledge to make an informed decision about interpretation is not available (see Bonnea et al., 2014, in Beck et al. 2020). Most examples involve the inability to distinguish between the subtypes PER-REAL and PER-LIT in texts that do not provide sufficient clues to the "origin" of the character, although this seems to be rather rare.

In such cases, manual annotation provides the opportunity for discussion and collective decision, which we see as an advantage, since cases where the uncertainty (or ambiguity) cannot be resolved are reduced to the absolute minimum, for example:

> ***Maruška*** – [PER-REAL, author's wife] *peče domači kruh …*
> ***Milenko, Andraž, Marko, David*** – [PER-REAL, members of OHO Slovenian art group: Milenko Matanović, Andraž

Šalamun, Marko Pogačnik, David Nez – established on the basis of extra-textual knowledge].

7. Preliminary results

Apart from the problems encountered in the annotation itself, the preliminary research results of the annotated corpus can also contribute to the study of characters in a selected corpus of literary texts. Based on the query and the results in NoSketchEngine, Figure 1 shows the quantitative relationship between three subtypes of the type PERSON (literary names, descriptive names, and names of characters from the non-literary world). It can be seen that the majority are literary names (PER-LIT, 68 per cent) whose predominance was to be expected - followed quantitatively by descriptive names (PER-DES, 18 per cent, and then by names of characters from the non-literary world (PER-REAL, 14 per cent).

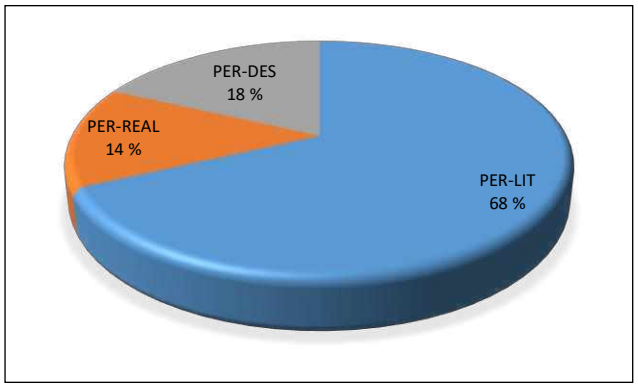


Figure 1: The ratio between the subtypes literary, descriptive and real of the PERSON type.

7.1. Categories of descriptive names and real names

Using the lists of the three types of personal names, we can create an approximate typology of character names according to the given typologies and evaluate the consistency of labelling. Because of their special characteristics, we limit ourselves to the subtypes descriptive and real, leaving aside the subtype literary, which includes mostly "ordinary" personal names.

Descriptive names are most often occupational (e.g., chief, inspector, captain, mayor; foreman, waitress, secretary, lab assistant); second are names expressing physical characteristics (e.g., one-armed, long-haired, "the one with the moustache" the handicapped), followed by names describing character (e.g., bully, beast, monster), beast, bloodthirsty),family relations (e.g., aunt, uncle, godmother), generational affiliation (e.g., old man, young man), while longer descriptive lexical strings are rarer (man with no name, brother in Christ, the long-haired one). Among the names for women, forms that formally express possession but function as gendered common proper names are frequent in Slovenian (e.g. Tomaž's (one), the manager's wife). This is statistically almost as significant as feminine names for occupations.

As can be seen from the annotated corpus, we identify five subcategories and include them in the subtype for real persons: 1. Real persons from social (Brutus, Lenin, Kidrič) and cultural history (Prešeren, Heidegger, Descartes, Shakespeare, Mozart); 2. Mythological figures (Cain, Poseidon, Ishtar); 3. Characters from other works of Slovenian and world literature (Pegam, Lambergar, Servant Jernej, Charlie Brown, Odysseus, Pinocchio); the last two groups are represented, on the one hand, by characters from the contemporary world of the authors, such as real-life celebrities (Tomaž Terček, Andraž Šalamun, Milenko Matanovič, Brigitte Bardot, Gérard Philipe, Giorgio Albertazzi, Sylvie Vartan) and, on the other hand, by characters from the authors' immediate (family) environment (Ana, Maruška).

The results show the least consistency for the descriptive name subtype with the lowest degree of intersubjectivity, especially with respect to the relationship between the transition from common noun to proper name and the aptronyms or nominative determinism, which Barthes considers a kind of "economic" characterization (Lahn and Maister, 2016). The relatively high presence of this subtype suggests a modernist blurring of the boundary between fiction and reality, which is reinforced by postmodernism.

7.2. Relationship between male and female characters

The second graph (cf. Figure 2) shows the quantitative ratio between male and female characters as they occur in the May68 Corpus (based on the number of tokens).

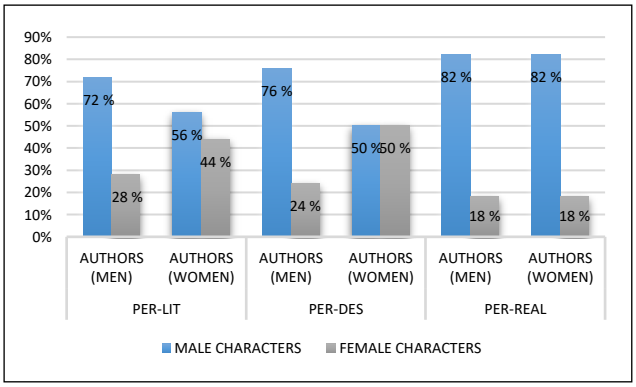


Figure 2: The quantitative relationship between male and female characters in the May68 Corpus.

The results confirm findings from other research (cf. Nagaraj and Kejriwal, 2022) that the proportion of male characters is significantly higher than that of women.

We supplement this account by comparing male and female characters by author gender, which gives a very disproportionate picture: Metadata analysis has shown the predominance of male authorship in the corpus (81 per cent) - only 7 per cent of authors are women, and there are no data for the remaining 12 per cent (Juvan, et al., 2021).

If we start from the gender of the authors when analyzing the occurrence of male and female characters, we find (see Figure 3) that in the works by men, male characters outnumber female characters by 44 per cent in the subcategory literary names, while this difference is much smaller in the works by women (12 per cent). In the category descriptive names, this ratio is difficult to assess due to the low occurrence among women authors, but a large difference between female and male characters in men authors goes in favour of the latter.

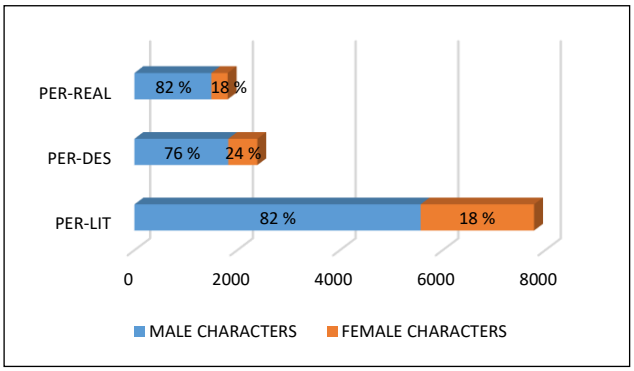


Figure 3: Male and female characters according to the gender of authors.

In the subcategory real, there is no significant difference in terms of author gender, which is probably due to the actual and undisputed presence of men and women in social and cultural history.

8. Conclusions and open challenges

The main goal of our annotation task was to provide an adequate representation of a specific set of semantic data (=Named Entities) and to fully exploit the potential of this type of corpus linguistic data in the context of future literary and linguistic analyses. To this end, we implemented a three-level annotation process. We conclude on the basis of high variation in referential expressions that in potential future projects an additional step should be linking the different names of the same character.

In the present work, we sought to identify and interpret different types of representational problems based on the model proposed by Beck et al. (2020) in order to improve our understanding of the linguistic and extra-linguistic properties of the texts in a (literary) corpus. It is hoped that this will lead to a more nuanced understanding of the challenges of NER, and that this in turn may inform future resources in ways that are more appropriate to the data they represent.

In the next phases of annotation, we plan to improve the segments that have the lowest level of consistency and agreement among annotators, such as common nouns that perform the referential function of proper names, seemingly operating as a representational continuum.

We have yet to work out the best approach to fully incorporate the various instances of PER-DES in the annotation scheme, but these are certainly worth considering as a special (sub)category of the NAME group.

9. Acknowledgements

ARRS (Slovenian Research Agency) J6-9384 “Maj 68 v literaturi in teoriji (May '68 in Literature and Theory)”

10. References

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias. In: The 14th Linguistic *Annotation Workshop*, pages 60–73, Barcelona, Spain, December 12, 2020.

Julian Brooke, Timothy Baldwin, and Adam Hammond. 2016. Bootstrapped Text-level Named Entity Recognition for Literature. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 344–350, Berlin, Germany, August 7–12.

John Burrows. 2004. Textual analysis. In: S. Schreibman, Ray Siemens, and John Unsworth, eds., *A Companion to Digital Humanities*. Blackwell, Oxford.

José Calvo Tello. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning*. Bielefeld University Press, Bielefeld.

Karine van Dalen-Oskam. 2022. *Distant Dreaming About European Literary History*. Evening keynote at the Distant Reading Closing Conference. https://www.distant-reading.net/events/conference-programme/

Jesse de Does, Katrien Depuydt, Karina van Dalen-Oskam, and Maarten Marx. 2017. Namescape: Named Entity Recognition from a Literary Perspective. In: J. Odijk, and A. van Hessen, eds., *CLARIN in the Low Countries*, pages 361–70. Ubiquity Press. https://www.ubiquitypress.com/site/chapters/10.5334/bbi.30/download/1046/.

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84. Osaka, Japan. The COLING 2016 Organizing Committee.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting Social Networks from Literary Fiction. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Tomaž Erjavec, Peter Holozan, and Nikola Ljubešić. 2015. Jezikovne tehnologije in zapis korpusa. In: V. Gorjanc, P. Gantar, I. Kosem and S. Krek, eds., *Slovar sodobne slovenščine: problemi in rešitve*, pages 262–76. Znanstvena založba Filozofske fakultete, Ljubljana.

Francesca Frontini, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. Named Entity Recognition for Distant Reading in ELTeC. In: *CLARIN Annual Conference 2020, Oct 2020*, str. 37–41, Virtual Event, France.

Marko Juvan, Andrejka Žejn, Mojca Šorli, Lucija Mandić, Andrej Tomažin, Andraž Jež, Varja Balžalorsky Antić, and Tomaž Erjavec. 2022. *Corpus of 1968 Slovenian literature Maj68 2.0*, ZRC SAZU. http://hdl.handle.net/11356/1430

Marko Juvan, Mojca Šorli, and Andrejka Žejn. 2021. Interpretiranje literature v zmanjšanem merilu: »Oddaljeno branje« korpusa »dolgega leta 1968«. *Jezik in slovstvo*, 66(4):55–76.

Nora Ketschik, André Blessing, Sandra Murr, Maximilian Overbeck, and Axel Pichler. 2020. Interdisziplinäre Annotation von Entitätenreferenzen. Von fachspezifischen Fragestellungen zur einheitlichen methodischen Umsetzung. In: N. Reiter, A. Pichler, and J. Kuhn, eds., *Reflektierte Algorithmische Textanalyse.*

Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt, pages 203–36, Berlin.

Benjamin Krautter, Janis Pagel, Nils Reiter, and Marcus Willand. 2018. In: T. Weitin, ed., *Eponymous Heroes and Protagonists – Character Classification in German-Language Dramas*. LitLab. Pamphlet # 7.

Silke Lahn, and Jan Christoph Meister. 2016. *Einführung in die Erzähltextanalyse*. Stuttgart, Metzler.

Nikola Ljubešić, Marija Stupar, and Tereza Jurič. 2012. Building Named Entity Recognition Models For Croatian And Slovene. In: T. Erjavec, and J. Žganec Gros, eds., *Proceedings of the Eighth Language Technologies Conference, October 8th-12th, 2012, Ljubljana, Slovenia: proceedings of the 15th International Multiconference Information Society - IS 2012, volume C*, pages 129–34. Ljubljana, Institut Jožef Stefan.

Anke Lüdeling. 2017. Variationistische Korpusstudien. In: M. Konopka, and A. Wöllstein, eds., *Grammatische Variation. Empirische Zugänge und theoretische Modellierung. IDS Jahrbuch 2016*, pages 129– 144. de Gruyter, Berlin.

Elena V. Mikhalkova, Timofei Protasov, Anastasiia Drozdova, Anastasiia Bashmakova, and Polina Gavin. 2019. Towards annotation of text worlds in a literary work. In: *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, pages 101–10. Issue 18, Supplementary Volume 18.

Franco Moretti. 2011. Network Theory, Plot Analysis. *New Left Review*, 68:80–102.

Akarsh Nagaraj, and Mayank Kejriwal. 2022. Robust Quantification of Gender Disparity in Pre-Modern English Literature using Natural Language Processing. arXiv:2204.05872v1 [cs.CY] 12 Apr 2022.

Sean Papay, and Sebastian Padó. 2020. RiQuA: A Corpus of Rich Quotation Annotation for English Literary Text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 835–841, Marseille, France. European Language Resources Association.

Janis Pagel, Nils Reiter, Ina Rösiger, and Sarah Schulz. 2020. Annotation als flexibel einsetzbare Methode. In: N. Reiter, A. Pichler, and J. Kuhn, eds., *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*, pages 125–142. Berlin.

Ranka Stanković, Diana Santos, Francesca Frontini, Tomaž Erjavec, and Carmen Brando. 2019. Named Entity Recognition for Distant Reading in Several Languages. In: G. Pálko, ed., *DH_Budapest_2019*. Budapest, ELTE. http://elte-dh.hu/dh_budapest_2019-abstract-booklet/

Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. 2007. Named Entities in Czech: Annotating Data and Developing NE Tagger. In: V. Matoušek, P. Mautner eds., *Text, Speech and Dialogue: 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3–7, 2007. Proceedings*. Berlin – Heidelberg, Springer-Verlag. https://ufal.mff.cuni.cz/~zabokrtsky/publications/papers/tsd07-namedent.pdf

Tadej Štajner, Tomaž Erjavec, and Simon Krek. 2013. Razpoznavanje imenskih entitet v slovenskem besedilu. In: *Proceedings of the Eighth Language Technologies Conference, October 8th-12th, 2012, Ljubljana, Slovenia: proceedings of the 15th International Multiconference Information Society - IS 2012, volume C*, pages 191–96, Ljubljana, Institut Jožef Stefan.

Katja Zupan, Nikola Ljubešić, and Tomaž Erjavec. 2017. Annotation guidelines for Slovenian named entities: Janes-NER. *Technical report, Jožef Stefan Institute, September*. https://www.clarin.si/repository/xmlui/bitstream/handle/11356/1123/SlovenianNER-eng-v1.1.pdf.

A Transformer-based Sequence-labeling Approach to the Slovenian Cross-domain Automatic Term Extraction

Thi Hong Hanh Tran*†, Matej Martinc†, Andraž Repar†, Antoine Doucet‡, Senja Pollak†

*Jožef Stefan International Postgraduate School,
Jamova cesta 39, 1000 Ljubljana, Slovenia

† Jožef Stefan Institute,
Jamova cesta 39, 1000 Ljubljana, Slovenia

‡ University of La Rochelle,
23 Av. Albert Einstein, La Rochelle, France

Abstract

Automatic term extraction (ATE) is a popular research task that eases the time and effort of manually identifying terms from domain-specific corpora by providing a list of candidate terms. In this paper, we treat terminology extraction as a sequence-labeling task and experiment with a Transformer-based model XLM-RoBERTa to evaluate the performance of multilingual pretrained language models in the cross-domain sequence-labeling setting. The experiments are conducted on the RSDO5 corpus, a Slovenian dataset containing texts from four domains, including Biomechanics, Chemistry, Veterinary, and Linguistics. We show that our approach outperforms the Slovene state-of-the-art approach, achieving significant improvements in F1-score up to 40 percentage points. This indicates that applying multilingual pretrained language models for ATE in less-resourced European languages is a promising direction for further development. Our code is publicly available at https://github.com/honghanhh/sdjt-ate.

1. Introduction

Terms are single- or multi-word expressions denoting concepts from specific subject fields whose meaning may differ from the same set of words in other contexts or everyday language. They represent units of knowledge in a specific field of expertise and term extraction is useful for several terminographical tasks performed by linguists (e.g., construction of specialized term dictionaries). Most of these tasks are time- and labor-demanding, therefore recently several automatic term extraction approaches have been proposed to speed up the process.

Term extraction can also support and improve several complex downstream natural language processing (NLP) tasks. The broad range of downstream NLP tasks to which term extraction could benefit include, for example, glossary construction (Maldonado and Lewis, 2016), topic detection (El-Kishky et al., 2014), machine translation (Wolf et al., 2011), text summarization (Litvak and Last, 2008), information retrieval (Lingpeng et al., 2005), ontology engineering and learning (Biemann and Mehler, 2014), business intelligence retrieval (Saggion et al., 2007; Palomino et al., 2013), knowledge visualization (Blei and Lafferty, 2009), specialized dictionary creation (Le Serrec et al., 2010), sentiment analysis (Pavlopoulos and Androutsopoulos, 2014), and cold-start knowledge base population (Ellis et al., 2015), to cite a few.

In the attempt to ease the time and effort needed to manually identify terms from domain-specific corpora, automatic term extraction (ATE), also known as automatic term recognition (Kageura and Umino, 1996) or automatic term detection (Castellví et al., 2001), thus became an essential NLP task. However, despite the importance of term extraction and the research attention paid to the task, identifying the correct terms remains a notoriously challenging problem with the following not yet solved hurdles. First, despite several different definitions to describe the meaning of a term, the explicit distinction between terms and common words is in many cases still unclear. In addition, the characteristics of specific terms can vary significantly across domains and languages. Furthermore, the gold standard term lists and manually labeled domain-specific corpora for training and evaluation of ATE approaches are generally scarce for less-resourced languages including Slovenian, due to the large amount of work required for the construction of these resources.

Deep neural approaches towards ATE have been only recently proposed, but their evaluation in less-resourced languages has not yet been sufficiently explored and remains a research gap worth investigating. Inspired by the success of Transformer-based models in ATE from the recent TermEval 2020 competition's ACTER dataset (Hazem et al., 2020; Lang et al., 2021), we propose to exploit and explore the performance of XLM-RoBERTa pretrained language model (Conneau et al., 2019), which addresses the ATE as a sequence-labeling task. Sequence-labeling approaches have been successfully applied to a range of NLP tasks, including Named Entity Recognition (Lample et al., 2016; Tran et al., 2021) and Keyword Extraction (Martinc et al., 2021; Koloski et al., 2022). The experiments are conducted in the cross-domain setting on the RSDO5 corpus[1] (Jemec Tomazin et al., 2021a) containing Slovenian texts

[1] http://hdl.handle.net/11356/1470

from four domains (Biomechanics, Chemistry, Veterinary, and Linguistics).

The main contributions of this paper can be summarized in the following points:

- We systematically evaluate the performance of the Transformer-based pretrained model, namely XLM-RoBERTa, on the term extraction task, formulated as a supervised cross-domain sequence-labeling on the RSDO5 dataset containing texts from four different domains.
- We demonstrate that the proposed cross-domain approach surpasses the performance of the current state of the art (Ljubešić et al., 2019) for all the combinations of training and testing domains we experimented with, therefore establishing a new state-of-the-art (SOTA) method for the ATE on Slovenian corpus.

This paper is organized as follows: Section 2. presents the related work in the field of term extraction. Next, we introduce our methodology in Section 3., and the experimental details in Section 4.. The results with further error analysis are discussed in Section 5. and 6., before we conclude and present future works in Section 7..

2. Related Work

The history of ATE has its beginnings during the 1990s with research done by Damerau (1990), Ananiadou (1994), Justeson and Katz (1995), Kageura and Umino (1996), and Frantzi et al. (1998). ATE systems usually employ the following two-step procedure: (1) extracting a list of candidate terms; and (2) determining which of these candidate terms are correct using supervised or unsupervised approaches. Recently, neural approaches have been proposed.

Traditionally, the approaches were strongly based on linguistic knowledge and distinctive linguistic aspects of terms in order to extract possible candidates. Several NLP tools, such as tokenization, lemmatization, stemming, chunking, PoS tagging, full syntactic parsing, etc., are employed in this approach to obtain linguistic profiles of term candidates. As a heavily language-dependent approach, the better the quality of the pre-processing tools (e.g., FLAIR (Akbik et al., 2019), Stanza (Qi et al., 2020)), the better the quality of linguistic ATE methods.

Meanwhile, several studies preferred the statistical approach or combined linguistic and statistical approaches. Some of the measures include the termhood (Vintar, 2010), unithood (Daille et al., 1994) or C-value (Frantzi et al., 1998). Many current systems still apply some variation of this approach, most commonly in hybrid systems combining linguistic and statistical information (Repar et al., 2019; Meyers et al., 2018; Drouin, 2003; Macken et al., 2013; Šajatović et al., 2019; Kessler et al., 2019, to cite a few.).

Recently, advances in embeddings and deep neural networks have also influenced the term extraction field. Several embeddings have been investigated for term extraction, for example, uni-gram term representations constructed from a combination of local and global vectors (Amjadian et al., 2016), non-contextual word embeddings (Wang et al., 2016; Khan et al., 2016; Zhang et al., 2017), contextual word embeddings (Kucza et al., 2018), and the combination of both representations (Gao and Yuan, 2019).

In the recent ATE challenge, namely TermEval 2020 (Rigouts Terryn et al., 2020), the use of language models became very important. The winning approach on the Dutch corpus used pretrained GloVe word embeddings fed into a bi-directional LSTM based neural architecture. Meanwhile, the winning approach on the English corpus (Hazem et al., 2020) relied on the extraction of all possible n-gram combinations, which are fed into a BERT binary classifier that determines for each n-gram inside a sentence, whether it is a term or not. Besides BERT, several other variations of Transformer-based models have also been investigated. For example, RoBERTa and CamemBERT have been used in the TermEval 2020 challenge (Hazem et al., 2020). Another recent method is the HAMLET system (Rigouts Terryn et al., 2021), which proposes a hybrid adaptable machine learning approach that combines the linguistic and statistical clues to detect terms and is also evaluated on the TermEval data.

Meanwhile, Conneau et al. (2019) and Lang et al. (2021) take advantage of XLM-RoBERTa (XLM-R) to compare three different approaches, including a binary sequence classifier, a sequence classifier, and a token classifier employing the sequence-labeling approach (also under research by Kucza et al. (2018)), as we do in our research. Finally, Lang et al. (2021) proposes to use a multilingual encoder-decoder model called mBART (Liu et al., 2020), which is based on denoising pre-training, that generates sequences of comma-separated terms from the input sentences.

Annotated Corpora for Term Extraction Research (ACTER) dataset was released for the TermEval competition as a collection of four domain-specific corpora (Corruption, Wind energy, Equitation, and Heart failure) in three languages (English, French, and Dutch). However, when it comes to ATE for less-resourced languages, there is still a lack of gold standard corpora and limited use of neural methods. In recent years, the Slovene KAS corpus was compiled (Erjavec et al., 2021), and most recently the RSDO corpus that we use in our study (Jemec Tomazin et al., 2021b). Regarding the Slovenian language on which we focus in our study, the current SOTA was proposed by Ljubešić et al. (2019) that extracts the initial candidate terms using the CollTerm tool (Pinnis et al., 2019), a rule-based system employing a complex language-specific set of term patterns (e.g., POS tag,...) from the Slovenian SketchEngine module (Fišer et al., 2016), followed by a machine learning classification approach with features representing statistical term extraction measures. Another recent approach by (Repar et al., 2019) focuses on term extraction and alignment, where the main novelty is in using an evolutionary algorithm for the alignment of terms. On the other hand, the deep neural approaches have not been explored for Slovenian yet. Another problem very specific for less-resourced languages is that the open-sourced code is often not available for most current benchmark systems, hindering their reproducibility (for Slovenian, only the code by Ljubešić et al. (2019) is available).



Figure 1: An example of the (B-I-O) mechanism on a text sequence from Slovenian corpus.

3. Methodology

We consider ATE as a sequence-labeling task where the model returns a label for each token in a text sequence. We use the (B-I-O) labeling mechanism (Rigouts Terryn et al., 2021; Lang et al., 2021) where B stands for the beginning word in the term, I stands for the word inside the term, and O stands for the word not part of the term. The terms from a gold standard list are first mapped to the tokens in the raw text and each word inside the text sequence is annotated with one of the three labels (see examples in Figure 1). The model is first trained to predict a label for each token in the input text sequence (e.g., we model the task as token classification) and then applied to the unseen text (test data). Finally, from the tokens or token sequences labeled as terms, the final candidate term list for the test data is composed.

We experiment with XLM-RoBERTa [2] (Conneau et al., 2019), a Transformer-based model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages. With the proliferation of non-English models (e.g., CamemBERT for French, Finnish BERT, German BERT, etc), XLM-RoBERTa, the multilingual version of RoBERTa (Liu et al., 2019), is a generic cross-lingual sentence encoder that achieves benchmark performance on multiple downstream NLP tasks, including ATE for rich-resourced languages (e.g. English) (Rigouts Terryn et al., 2020). Due to this well-documented SOTA performance on several related tasks, we opted to employ XLM-RoBERTa in a monolingual setting on our low-resourced Slovenian corpus. The overall architecture of our approach is presented in Figure 2.

In our experiments, we use a multilingual pre-trained language model in order to leverage the general knowledge the model obtained during pretraining on the huge multilingual corpus. First, we divide the dataset into train-validation-test splits. We also investigate the effectiveness of cross-domain learning, where the main idea is to test the transfer of knowledge from one domain to another and therefore evaluate the capability of the model to extract terms in new unseen domains as well as the ability to learn the relations between terms across domains given the assumption that they have terminologically-marked contexts. Therefore, we fine-tune the model on two domains (e.g., Biomechanics, Chemistry) as the train split, validate on a third domain (e.g., Veterinary) as the validation split, and test on the fourth domain that does not appear in the train set (e.g., Linguistics). The train split is used for fine-tuning the pre-trained language model. The validation split is applied to prevent over-fitting during the fine-tuning phase. Finally, the test split, which is not adopted during training, is used for the evaluation of the method.

[2] https://huggingface.co/xlm-roberta-base

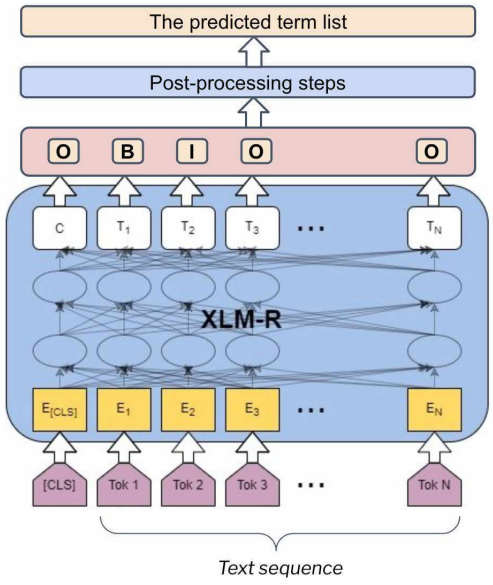


Figure 2: The overall architecture.

The model is fine-tuned on the training set to predict the probability for each word in a word sequence whether it is a part of the term (B, I) or not (O). To do so, an additional token classification head containing a feed-forward layer with a softmax activation is added on top of the model.

4. Experimental Setup

Here, we describe the dataset, the experimental details, and the metrics that we apply for the evaluation.

4.1. Dataset

The experiments are conducted on the Slovenian RSDO5 corpus version 1.1 (Jemec Tomazin et al., 2021a), which is a less-resourced Slavic language with rich morphology. As a part of the RSDO national project, the RSDO5 corpus was manually compiled and annotated and contains 12 documents with altogether about 250,000 words from the fields of Biomechanics (bim), Chemistry (kem), Veterinary (vet), and Linguistics (ling). The data were collected from diverse sources, including Ph.D. theses (3), a Ph.D. thesis-based scientific book (1), graduate-level textbooks (4), and journal articles (4) published between 2000 and 2019. Apart from the manually annotated terms, RSDO5 is also annotated with Universal Dependency tags (e.g. tags annotating tokens, sentences, lemmas, morphological features, etc.). However, in our research, we only leverage the original text with the term labels, where we consider all terms and do not distinguish between in-domain and out-of-domain terms.

In Table 1, we report on the number of documents, tokens, and unique terms across domains. Given the same

Languages	Biomechanics (bim)			Chemistry (kem)			Veterinary (vet)			Linguistics (ling)		
	# Docs	# Tokens	# Terms	# Docs	# Tokens	# Terms	# Docs	# Tokens	# Terms	# Docs	# Tokens	# Terms
Slovenian	3	61,344	2,319	3	65,012	2,409	3	75182	4,748	3	109,050	4,601

Table 1: Number of documents, tokens, and unique terms per domain in Slovenian RSDO5 dataset.

Languages	Biomechanics (bim)				Chemistry (kem)				Veterinary (vet)				Linguistics (ling)			
	B	I	O	% Term	B	I	O	% Term	B	I	O	% Term	B	I	O	% Term
Slovenian	7,070	6,835	47,439	22.67	7,614	4,486	52,912	18.61	10,953	6,261	57,968	22.90	12,348	6,079	90,623	16.89

Table 2: Label distribution and the proportion of terms appearing per domain in the Slovenian RSDO5 dataset.

number of collected documents for each domain, the documents from the Linguistics and Veterinary domains are longer (i.e., have more tokens) and also contain more terms than the domains of Biomechanics and Chemistry. In addition, Figure 3 presents the frequency of terms of different lengths per domain. Veterinary, Chemistry, and Linguistics share a similar term length distribution with most terms made of one to three words and only a few (less than three) terms longer than seven words (an example of a long term found in the corpus would be *"kaznivo dejanje zoper življenje , telo in premoženje"*, which means a crime against life, body, and property). Meanwhile, the Biomechanics domain distribution has a longer right tail, containing several terms with more than three words.

Furthermore, the corpus contains several nested terms, i.e., they also appear within larger terms and vice versa, a multiword term may contain shorter terms. For example, in the Biomechanics domain, term *"navor"* (torque) appears in terms such as *"sunek navora"* (torque shock), *"zunanji sunek navora"* (external torque shock), and *"izokinetični navor"* (isokinetic torque), to mention a few. This makes the labeling harder and the classifier needs to infer from the context whether a specific term is part of a longer term.

4.2. Implementation Details

We experiment with several combinations of training, validation, and testing data where two domains are used for training, the third one for validation, and the fourth one for testing (i.e., we train 12 models covering all possible domain combinations). We consider term extraction as a sequence-labeling or token classification task with a (B-I-O) annotation scheme. Table 2 presents the distribution across label types and the proportion of (B) and (I) labels in the total number of tokens per domain in the dataset. On average, the number of tokens annotated as terms (or parts of the term) only represents about one-fifth of the total tokens in the corpus, which means that there is a significant imbalance between (B) and (I) tokens, and tokens labeled as not terms (O).

We employ the XLM-RoBERTa token classification model and its "fast" XLM-RoBERTa tokenizer from the Huggingface library [3]. We fine-tune the model for up to 20 epochs regarding model convergence (i.e., we also employ the early stopping regime) with the learning rate of 2e-05, training and evaluation batch size of 32, and sequence length of 512 tokens, since this hyperparameter configuration performed the best on the validation set. The documents are split into sentences and the sentences containing more than 512 tokens are truncated, while the sentences with less than 512 tokens are padded with a special $< PAD >$ token at the end. During fine-tuning, the model is evaluated on the validation set after each training epoch, and the best-performing model is applied to the test set.

The model predicts each word in a word sequence whether it is a part of a term (B, I) or not (O). The sequences identified as terms are extracted from the text and put into a set of all predicted candidate terms. A post-processing step to lowercase all the candidate terms is applied before we compare our derived candidate list with the gold standard using the evaluation metrics discussed in Section 4.3..

4.3. Evaluation Metrics

We perform the global evaluation on our term extraction system by comparing the list of candidate terms extracted on the level of the whole test set with the manually annotated gold standard in the test set using Precision, Recall, and F1-score. Precision refers to the percentage of the extracted terms that are correct. Meanwhile, Recall indicates the percentage of the total correct terms that are extracted. Low Precision means a lot of noise in extraction whereas low Recall indicates the presence of lots of misses in extraction. Besides, F_1-score is a measure that computes an overall performance by calculating the harmonic mean between Precision and Recall). These evaluation metrics have been used also in the related work, including the TermEval 2020 shared task (Hazem et al., 2020; Rigouts Terryn et al., 2020; Lang et al., 2021).

5. Results

Table 3 presents the results achieved by the multilingual XLM-RoBERTa pre-trained language model on the Slovenian RSDO5 dataset. Note that the results in the table are grouped according to the model's test domain for better comparison between different settings. Our cross-domain approach proves to have relatively consistent performance across all the combinations, achieving Precision of more than 62%, Recall of no less than 55%, and F1-score above 61%. The model performs slightly better for the Linguistics and Veterinary domains than for Biomechanics and Chemistry. The difference in the number of terms and length of terms per domain pointed out in Section 4.1. might be one of the factors that contribute to this behavior. In addition, a significant performance boost can be observed for the Linguistics domain when the model is trained in the Chemistry

[3] https://huggingface.co/models

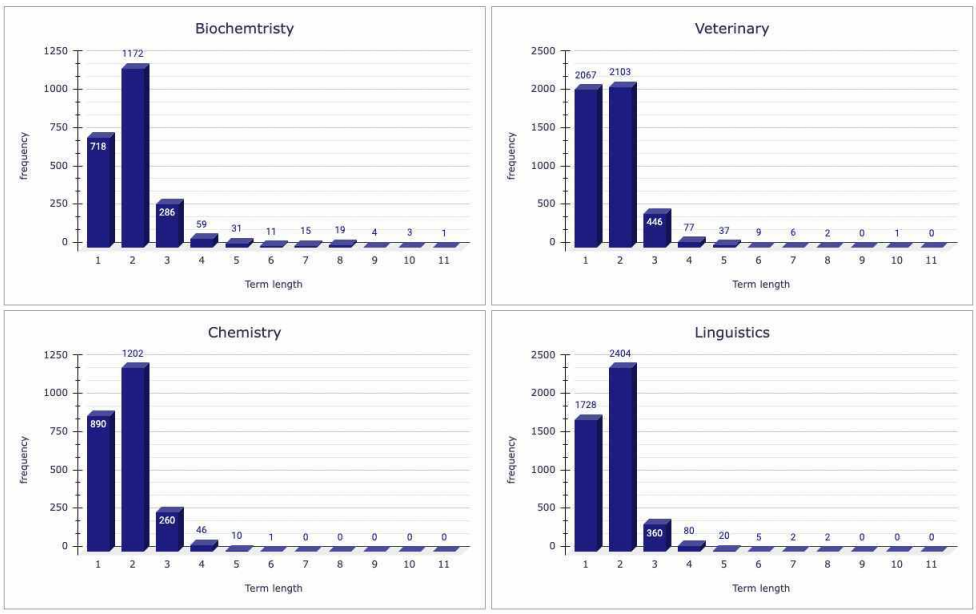


Figure 3: The frequencies of terms of specific length per each domain in a Slovenian dataset.

and Veterinary domains, and for the Veterinary domain, when the model is trained in Biomechanics and Linguistics. In these two settings, the model achieves an F1-score of more than 68%.

Training	Validation	Testing	Precision	Recall	F1-score
bim + kem	vet	ling	**69.55**	64.05	66.69
bim + vet	kem	ling	69.48	**73.66**	**71.51**
kem + vet	bim	ling	66.20	72.38	69.15
Ljubešić et al. (2019)		ling	52.20	25.40	34.10
bim + kem	ling	vet	71.06	66.72	68.82
bim + ling	kem	vet	**72.66**	65.59	**68.94**
ling + kem	bim	vet	69.3	**68.07**	68.68
Ljubešić et al. (2019)		vet	66.90	19.30	29.90
bim + vet	ling	kem	68.67	55.13	61.16
bim + ling	vet	kem	70.14	**60.27**	**64.83**
ling + vet	bim	kem	**70.23**	59.24	64.27
Ljubešić et al. (2019)		kem	47.80	31.40	37.80
vet + kem	ling	bim	**63.51**	**66.80**	**65.11**
vet + ling	kem	bim	62.25	65.20	63.69
ling + kem	vet	bim	62.35	63.99	63.16
Ljubešić et al. (2019)		bim	53.80	24.80	33.90

Table 3: Term extraction evaluation in a cross-domain setting on a Slovenian RSDO5 dataset.

We also present results for the current SOTA approach from Ljubešić et al. (2019) by reproducing their methodology in the same RSDO5 dataset. In general, our approach outperforms the approach proposed by Ljubešić et al. (2019) by a large margin on all domains and according to all evaluation metrics. The margin is especially large when it comes to Recall. Given the training process applied on RSDO5 corpus, Ljubešić et al. (2019) approach has low performance in F1-score due to the high imbalance between the Precision and Recall. This is most likely due to the fact that the methods employed by Ljubešić et al. (2019) rely heavily on the frequency and are thus not suitable for discovering low-frequency terms of which there are a lot in the RSDO5 corpus. In their own experiments, Ljubešić et al. (2019) discard all term candidates with a frequency below 3, hence why their results on their corpus are higher than on RSDO5.

Overall, we achieve results roughly twice as high as the approach proposed by Ljubešić et al. (2019) in terms of F1-score for all test domains. The results demonstrate the predictive power of contextual information in language models such as XLM-RoBERTa over the machine learning approach with features representing statistical term extraction measures as in Ljubešić et al. (2019).

6. Error Analysis

In this section, we analyze the predictions of XLM-RoBERTa in the RSDO5 corpus to get a better understanding of the model's performance and discover possible avenues for future work. First, we analyze the predictive power of our approach for terms of different lengths by calculating the Precision and Recall separately for terms of length k = {1,2,3,4, equal or more than 5}. The number of predicted candidate terms, number of ground truth terms, number of correct predictions (TPs), Precision, and Recall regarding different terms of length k and different test domains are presented in Tables 4, 5, 6, and 7. Note that these statistics are collected for the train-validation-test combinations that perform the best on each domain according to the F1-score.

Results across Tables 4 to 7 show that our models are good at predicting short terms containing up to three words in all four domains. The best model applied to the Linguistics test domain also shows competitive performance for the prediction of longer terms, achieving 75.00% Precision and a decent 31.03% Recall for terms with at least 5 words. Despite the relatively high Precision achieved by the models on long terms in the Veterinary and Biomechanics test domains, the Recall is pretty low, most likely due to the small

k	#Predictions	#Ground truth	#TPs	Precision	Recall
1	2,078	1,728	1,300	62.56	75.23
2	2,631	2,404	1,858	70.62	77.29
3	322	360	7,191	59.32	53.06
4	57	80	31	54.39	38.75
≥5	12	29	79	75.00	31.03

Table 4: Performance in Precision and Recall per term length in Linguistics domain.

k	#Predictions	#Ground truth	#TPs	Precision	Recall
1	2,159	2,067	1,472	68.18	71.21
2	2,062	2,103	1,448	70.22	68.85
3	314	446	182	57.96	40.81
4	28	77	10	35.71	12.99
≥5	3	55	2	66.67	3.64

Table 5: Performance in Precision and Recall per term length in Veterinary domain.

k	#Predictions	#Ground truth	#TPs	Precision	Recall
1	943	890	580	61.51	65.17
2	1,073	1,202	768	71.58	63.89
3	164	260	93	56.71	35.77
4	26	46	11	42.31	23.91
≥5	3	11	0	0.00	0.00

Table 6: Performance in Precision and Recall per term length in Chemistry domain.

k	#Predictions	#Ground truth	#TPs	Precision	Recall
1	1,079	718	22	48.38	72.70
2	1,153	1,172	822	71.29	70.14
3	223	286	124	55.61	43.36
4	26	59	11	42.31	18.64
≥5	11	84	5	45.45	5.95

Table 7: Performance in Precision and Recall per term length in Biomechanics domain.

amount of longer terms in the dataset on which the models are trained. When it comes to predictions in the Chemistry domain, there are no correct term predictions that consist of more than five words.

In addition, as the corpus contains many nested terms, the very common mistake the model makes is to predict a shorter term nested in the correct term of the gold standard (Pattern 1). Vice versa, the model sometimes generates incorrect predictions containing the correct nested terms (Pattern 2). Furthermore, in some cases, the model predicts a single prediction made out of two consecutive terms (Pattern 3). We report some examples of these incorrect patterns in Table 8, where the first column refers to the pattern type, the second one refers to our predicted candidate term, and the last column presents the true term from the gold standard. The presented candidate terms are extracted from the final list of predicted terms for the Linguistics test domain.

7. Conclusion

In summary, we investigated the performance of the multilingual Transformer-based language model, XLM-RoBERTa, in the monolingual cross-domain sequence-labeling term extraction task. The experiments were conducted on the representative Slovenian RSDO5 corpus, which contains texts from four specific domains, namely Biomechanics, Chemistry, Veterinary, and Linguistics. Our cross-domain sequence-labeling approach with XLM-RoBERTa had consistent performance across all the combinations of training, validation, and test set, achieving the performance of up to 72.66% in terms of Precision, up to 73.66% in terms of Recall, and up to 71.51% in terms of F1-score. The model performed slightly better in extracting terms from the Linguistics and Veterinary domains than from Biomechanics and Chemistry. Moreover, our approach outperformed the current state of the art on the Slovenian language (Ljubešić et al., 2019) by a large margin according to all three evaluation metrics, in some cases achieving three times higher Recall and roughly two times higher F1-score. As a consequence, our approach is the new SOTA approach on the RSDO5 dataset.

However, we believe that there remains room for improvement in the field of supervised term extraction. In the future, we would like to pre-train the model on the intermediate task (e.g., machine translation) resembling term extraction before fine-tuning it on the target downstream task, in order to boost the extraction performance. In addition, we will also investigate the performance of the models in the zero-shot cross-lingual setting, multi-lingual setting, and the combination of both settings in comparison with our current monolingual setting. Lastly, we suggest the integration of active learning into our current approach to improve the output of the automated method by dynamical adaptation after human feedback. By learning with humans in the loop, we aim at getting the most information with the least amount of term labels. We will also evaluate the contribution of active learning in reducing the annotation effort and determine the robustness of the incremental active learning framework across different languages and domains.

8. Acknowledgements

The work was partially supported by the Slovenian Research Agency (ARRS) core research program Knowledge Technologies (P2-0103) and project TermFrame (J6-9372), as well as the Ministry of Culture of the Republic of Slovenia through project Development of Slovene in Digital Environment (RSDO). The first author was partly funded by Region Nouvelle Aquitaine. This work has also been supported by the TERMITRAD (2020-2019-8510010) project funded by the Nouvelle-Aquitaine Region, France.

9. References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair:

Patterns	Our predictions	The gold standards
1	*"klasična analogna telefonska zveza"* (classic analog telephone connection)	*"klasična analogna telefonska zveza pot"* (classic analog telephone connection path)
	"končnica neprve slovarske oblike" (suffix of non-first dictionary form)	*"končnica"* (suffix)
	...	...
2	*"brezžično slušalk v ušesu"* (wireless in-ear headphones)	*"brezžično slušalk"* (wireless headphones)
	"elektromehanska uporaba električne energije" (electromechanical use of electrical energy)	*"električne energije"* (electrical energy)
	...	...
3	*"batne parne stroje za pogon"* (reciprocating steam engines)	*"batne parne stroje" , "pogon"* (piston steam engines), (propulsion)
	"elektrarna na atomski pogon" (nuclear power plant)	*"elektrarna", "atomski pogon"* (power plant), (nuclear power plant)
	"besedilnim tipom strokovnega jezika" (text type professional language)	*"besedilnim tipom", "strokovnega jezika"* (text type), (professional language)
	"eksperimentalno modeliranje dinamičnih sistemov" (experimental modeling of dynamic systems)	*"eksperimentalno modeliranje", "dinamičnih sistemov"* (experimental modeling), (dynamic systems)
	...	...

Table 8: Examples of unlemmatised predictions in the Linguistics test domain.

An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Ehsan Amjadian, Diana Inkpen, Tahereh Paribakht, and Farahnaz Faez. 2016. Local-Global Vectors to Improve Unigram Terminology Extraction. In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pages 2–11.

Sophia Ananiadou. 1994. A methodology for automatic term recognition. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Chris Biemann and Alexander Mehler. 2014. *Text mining: From ontology learning to automated text processing applications*. Springer.

David M Blei and John D Lafferty. 2009. Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*.

M Teresa Cabré Castellví, Rosa Estopa Bagot, and Jordi Vivaldi Palatresi. 2001. Automatic term detection: A review of current systems. *Recent advances in computational terminology*, 2:53–88.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. In *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*.

Fred J Damerau. 1990. Evaluating computer-generated domain-oriented vocabularies. *Information processing & management*, 26(6):791–801.

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare Voss, and Jiawei Han. 2014. Scalable topical phrase mining from text corpora. *arXiv preprint arXiv:1406.6312*.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M Strassel. 2015. Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and results. In *TAC*.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2021. The kas corpus of slovenian academic writing. *Language Resources and Evaluation*, 55(2):551–583.

Darja Fišer, Vit Suchomel, and Miloš Jakubícek. 2016. Terminology extraction for academic slovene using sketch engine. In *Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2016*, pages 135–141.

Katerina T Frantzi, Sophia Ananiadou, and Junichi Tsujii. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. In *International conference on theory and practice of digital libraries*, pages 585–604. Springer.

Yuze Gao and Yu Yuan. 2019. Feature-less End-to-end Nested Term extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 607–616. Springer.

Amir Hazem, Mérieme Bouhandi, Florian Boudin, and Béatrice Daille. 2020. TermEval 2020: TALN-LS2N System for Automatic Term Extraction. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 95–100.

Mateja Jemec Tomazin, Mitja Trojar, Simon Atelšek, Tanja Fajfar, Tomaž Erjavec, and Mojca Žagar Karer. 2021a.

Corpus of term-annotated texts RSDO5 1.1. Slovenian language resource repository CLARIN.SI.

Mateja Jemec Tomazin, Mitja Trojar, Mojca Žagar, Simon Atelšek, Tanja Fajfar, and Tomaž Erjavec. 2021b. Corpus of term-annotated texts rsdo5 1.0.

John S Justeson and Slava M Katz. 1995. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural language engineering*, 1(1):9–27.

Kyo Kageura and Bin Umino. 1996. Methods of Automatic Term Recognition.A Review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2):259–289.

Rémy Kessler, Nicolas Béchet, and Giuseppe Berio. 2019. Extraction of terminology in the field of construction. In *2019 First International Conference on Digital Data Processing (DDP)*, pages 22–26. IEEE.

Muhammad Tahir Khan, Yukun Ma, and Jung-jae Kim. 2016. Term Ranker: A Graph-Based Re-Ranking Approach. In *FLAIRS Conference*, pages 310–315.

Boshko Koloski, Senja Pollak, Blaž Škrlj, and Matej Martinc. 2022. Out of thin air: Is zero-shot cross-lingual keyword detection better than unsupervised? *arXiv preprint arXiv:2202.06650*.

Maren Kucza, Jan Niehues, Thomas Zenkel, Alex Waibel, and Sebastian Stüker. 2018. Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks. In *INTERSPEECH*, pages 2072–2076.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3607–3620.

Annaïch Le Serrec, Marie-Claude L'Homme, Patrick Drouin, and Olivier Kraif. 2010. Automating the compilation of specialized dictionaries: Use and analysis of term extraction and lexical alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 16(1):77–106.

Yang Lingpeng, Ji Donghong, Zhou Guodong, and Nie Yu. 2005. Improving retrieval effectiveness by using key terms in top retrieved documents. In *European Conference on Information Retrieval*, pages 169–184. Springer.

Marina Litvak and Mark Last. 2008. Graph-based keyword extraction for single-document summarization. In *Coling 2008: Proceedings of the workshop multi-source multilingual information extraction and summarization*, pages 17–24.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. Kas-term: Extracting Slovene Terms from Doctoral Theses via Supervised Machine Learning. In *International Conference on Text, Speech, and Dialogue*, pages 115–126. Springer.

Lieve Macken, Els Lefever, and Veronique Hoste. 2013. Texsis: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 19(1):1–30.

Alfredo Maldonado and David Lewis. 2016. Self-tuning ongoing terminology extraction retrained on terminology validation decisions. In *Proceedings of The 12th International Conference on Terminology and Knowledge Engineering*, pages 91–100.

Matej Martinc, Blaž Škrlj, and Senja Pollak. 2021. Tnt-kid: Transformer-based neural tagger for keyword identification. *Natural Language Engineering*, page 1–40.

Adam L Meyers, Yifan He, Zachary Glass, John Ortega, Shasha Liao, Angus Grieve-Smith, Ralph Grishman, and Olga Babko-Malaya. 2018. The Termolator: Terminology Recognition Based on Chunking, Statistical and Search-Based Scores. *Frontiers in Research Metrics and Analytics*, 3:19.

Marco A Palomino, Tim Taylor, and Richard Owen. 2013. Evaluating business intelligence gathering techniques for horizon scanning applications. In *Mexican International Conference on Artificial Intelligence*, pages 350–361. Springer.

John Pavlopoulos and Ion Androutsopoulos. 2014. Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 44–52.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, Tatjana Gornostaja, Špela Vintar, and Darja Fišer. 2019. Extracting data from comparable corpora. In *Using Comparable Corpora for Under-Resourced Areas of Machine Translation*, pages 89–139. Springer.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Andraž Repar, Vid Podpečan, Anže Vavpetič, Nada Lavrač, and Senja Pollak. 2019. TermEnsembler: An Ensemble Learning Approach to Bilingual Term Extraction and Alignment. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 25(1):93–120.

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. TermEval 2020: Shared Task on

Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)*, pages 85–94. European Language Resources Association (ELRA).

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2021. HAMLET: Hybrid Adaptable Machine Learning approach to Extract Terminology. *Terminology*.

Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva. 2007. Ontology-based information extraction for business intelligence. In *The Semantic Web*, pages 843–856. Springer.

Antonio Šajatović, Maja Buljan, Jan Šnajder, and Bojana Dalbelo Bašić. 2019. Evaluating automatic term extraction methods on individual documents. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 149–154.

Thi Hong Hanh Tran, Antoine Doucet, Nicolas Sidere, Jose G Moreno, and Senja Pollak. 2021. Named entity recognition architecture combining contextual and global features. In *Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings*, page 264. Springer Nature.

Spela Vintar. 2010. Bilingual Term Recognition Revisited: The Bag-of-equivalents Term Alignment Approach and its Evaluation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 16(2):141–158.

Rui Wang, Wei Liu, and Chris McDonald. 2016. Featureless Domain-Specific Term Extraction with Minimal Labelled Data. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 103–112.

Petra Wolf, Ulrike Bernardi, Christian Federmann, and Sabine Hunsicker. 2011. From statistical term extraction to hybrid machine translation. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*.

Ziqi Zhang, Jie Gao, and Fabio Ciravegna. 2017. SemRe-Rank: Incorporating Semantic Relatedness to Improve Automatic Term Extraction Using Personalized PageRank. *arXiv preprint arXiv:1711.03373*.

Metapodatki o posnetkih in govorcih v govornih virih: primer baze Artur

Darinka Verdonik,* Andreja Bizjak,* Andrej Žgank,* Simon Dobrišek†

* Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška 46, 2000 Maribor
darinka.verdonik@um.si, andreja.bizjak1@um.si, andrej.zgank@um.si
† Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1000 Ljubljana
simon.dobrisek@fe.uni-lj.si

Povzetek

Ob združevanju različnih govornih jezikovnih virov se pojavljajo težave, ki izhajajo iz vsebinske nezdružljivosti zabeleženih metapodatkov o posnetkih, govorcih oz. govoru nasploh (npr. tip govora, vrsta govornega dogodka, lokacija in čas snemanja, spol, izobrazba, regija govorca). Ti metapodatki se zajemajo po eni strani zato, da omogočajo preverjanje uravnoteženosti govornega vira glede na različne govorce in govorne situacije, po drugi strani pa zato, da omogočajo razvrščanje govornih podatkov v kategorije, potrebne bodisi za jezikoslovne analize bodisi za učenje algoritmov razpoznavanja govora ipd. Najpogostejše razlike med zabeleženimi metapodatki o posnetkih in govorcih v obstoječih prosto dostopnih govornih virih za slovenščino so v kategorizacijah vrste govora in lokacije snemanja oziroma v kategorizacijah in oznakah regije govorca. Različne kategorije se pojavljajo tudi v zvezi s starostnimi in izobrazbenimi skupinami govorcev. Veliko vrst metapodatkov se pojavlja samo v posameznih virih, v drugih pa ne. Prispevek poleg pregleda razlik podaja tudi predloge za njihovo premostitev.

Metadata on recordings and speakers in spoken language resources: The case of the Artur database

When merging data from different spoken language resources, problems arise due to incompatibility of metadata on recordings, on speakers or on speech in general (e.g., information about the speech type or speech event, time and place of the recording, the gender, education, region of speaker). These metadata are captured on the one hand to ensure the balance of speech samples according to different speakers and speech situations, and on the other hand to enable the classification of speech data into categories needed either for linguistic analysis or for learning speech recognition algorithms. The most common differences in metadata on recordings and on speakers in the existing freely available speech resources for Slovene relate to categorizations of the type of speech and the location of the recording as well as to categorizations and designations of the speaker's region. Different categories also emerge in relation to age and educational groups of speakers. Many types of metadata are recorded only in particular resources. In addition to reviewing the differences we also give some suggestions how to overcome them.

1. Uvod

Govorni jezikovni viri so pomembni tako za razvoj jezikoslovja in celostno poznavanje jezika kot tudi za razvoj govornih tehnologij, kot je razpoznavanje ali sinteza govora. Poleg posnetkov in zapisa govora vsebujejo običajno tudi manjše ali večje število podatkov o tem, kje, kdaj, kako so posnetki nastali in kakšne so lastnosti govorcev glede na spol, starost, izobrazbo ipd. Čeprav Text Encoding Iniciative – TEI vključuje tudi standardizacijska priporočila s področja govora, pa so vsebinske odločitve, katere kategorije tovrstnih podatkov zajeti in kako podrobno jih opisati, zelo odvisne od vrste gradiva in namena govornega vira. Tako se ob združevanju virov, nastalih v različnih časovnih obdobjih z delno različnimi cilji in vključujoč različne tipe govora, pojavljajo težave, ki izhajajo iz vsebinske nezdružljivosti popisanih podatkov o posnetkih, govorcih oz. govoru nasploh.

S ciljem, da se tovrstne težave v prihodnje zmanjšajo, bomo v tem prispevku pregledali, potrebe po katerih podatkih so se pojavljale v različnih vedah, s poudarkom na dosedanjih slovenskih govornih virih (poglavji 2 in 3), podrobneje predstavili strukturo teh podatkov na primeru govorne baze Artur, ki predstavlja najnovejši in hkrati najobsežnejši in najbolj heterogen govorni vir za slovenščino ta trenutek (poglavje 4), ter izpostavili tiste vrste podatkov, kjer so vsebinska razhajanja največja, in podali predlog za njihovo uskladitev (poglavje 5).

2. Metapodatki o posnetkih in govorcih v govornih korpusih

Korpus GOS je predstavljal enega prvih večjih projektov, namenjenih zagotovitvi obsežnejšega govornega vira za raziskave slovenskega jezika. Izdan je bil leta 2011 v obsegu ca. 112 ur posnetkov in je sledil za tisti čas aktualnim korpusnojezikoslovnim prizadevanjem po dopolnjevanju referenčnih pisnih korpusov z referenčnimi govornimi korpusi (npr. Burnard, 2007; Allwood et al., 2000; Oostdijk et al., 2002; Pořízka, 2009). Njegov namen je bil torej predvsem zagotoviti podatke o govorjeni slovenščini za leksikografske, slovnične in druge jezikoslovne raziskave, za poučevanje slovenščine, za poklicne govorce ali pisce oz. tudi za širšo zainteresirano javnost. Vseboval je kolikor mogoče reprezentativen nabor različnih govornih situacij, s ciljem, da bi zajeli vzorčne primere različnih govornih situacij in različnih govorjenih diskurzov, demografsko reprezentativen vzorec govorcev slovenskega jezika in tiste govorne situacije, v katerih so uporabnika jezika najbolj pogosto produktivno ali pa samo receptivno udeleženi (Verdonik in Zwitter Vitez, 2011: 17).

GOS je bil poleg transkripcij dopolnjen tudi s posnetki ter s številnimi podatki o posnetkih in govorcih (po katerih lahko uporabniki korpusa tudi filtrirajo zadetke). Podobna je praksa v drugih, tujih govornih korpusih. Običajni podatki o situaciji, ki je posneta, so datum, lokacija, vrsta interakcije, kontekst, tematika, udeleženci, trajanje, uporabljena oprema za snemanje, vir ipd. Podatki o

udeležencih so običajno identitifikacijska koda, starost, spol, narodnost oz. prvi jezik, regija oz. narečje, poklic, lahko pa tudi še mesto rojstva, trenutna lokacija, drugi jeziki ipd. (Zemljarič Miklavčič, 2008; Cresti in Moneglia, 2005; Ehmer in Martinez, 2014; Love et al., 2017).

V korpusu GOS so metapodatki o posnetkih vključevali (Verdonik in Zwitter Vitez, 2011):

- podatke o gradivodajalcu oz. viru posnetka,
- podatke o vrsti govora, institucionalnem okviru, govornem dogodku, prosti opis govorne situacije in število aktivnih udeležencev govornega dogodka,
- podatke o času in kraju snemanja, pri čemer je bil kraj snemanja opredeljen tako z imenom kraja kot umeščen v širše (registrsko) območje.

Podatki o govorcih so zajemali:

- spol,
- starost, razdeljeno v 7 kategorij,
- izobrazbo, razdeljeno v 4 kategorije,
- regijo govorca, opredeljeno glede na registrsko območje, pri čemer je bila možnost opredelitve več regij v primeru, da je govorec več kot eno leto bival v različnih regijah (npr. zaradi študija, službe ipd.),
- prvi jezik govorca.

Korpusu GOS je v letih 2016–2019 v več izdajah sledila manjša govorna baza Gos Videolectures (Verdonik, 2018), ki je v nasprotju s korpusom GOS zajema področno omejeno gradivo javnih predavanj, dostopnih prek portala Videolectures.net. V svoji zadnji, četrti različici obsega skupno 22 ur posnetkov javnih predavanj, uravnoteženih glede na tematska področja družboslovja, humanistike, medicine, tehnike ter naravoslovja/matematike. Prav tako nastopajoči govorci enakomerno predstavljajo oba spola, starejše in mlajše govorce ter grobo opredeljene različne regije Slovenije.

Metapodatki o posnetkih in govorcih so sledili shemi, zastavljeni v korpusu GOS, vendar zaradi omejenega dostopa do informacij niso bili beleženi z isto natančnostjo. Če je bila starost govorcev v korpusu GOS deljena v 7 kategorij, je v Gos Videolectures samo v 2, pa še to predvsem na podlagi vizualnega vtisa, ne na podlagi neposredne, točne informacije. Prav tako ni bilo neposrednih podatkov o regiji govorca, ampak so bili pod to postavko zabeleženi slušni vtisi o značilnostih govora. Nekateri podatki pa niti niso bili opredeljeni, saj bodisi niso bili dostopni (izobrazba) bodisi niso bili relevantni (prvi jezik, ki je za vse govorce slovenščina). Ker je bila govorna baza Gos Videolectures namenjena tudi razvoju tehnologije razpoznavanja govora, so se pokazale potrebe še po beleženju kvalitete posnetka, ki je bila dodana zgolj kot subjektivna ocena transkriptorja na podlagi slušnega vtisa.

3. Metapodatki o posnetkih in govorcih v govornih bazah za razpoznavanje govora

Z vidika razvoja govornih tehnologij oziroma razpoznavalnikov govora je glavni razlog za zbiranje podatkov o govorcih in posnetkih predvsem ta, da se v govorni bazi zagotovi čim bolj ustrezna reprezentativna zastopanost vseh izrazitih govornih značilnosti, ki se spreminjajo med različnimi govorci in različnimi govornimi okoliščinami. Relevanten je torej katerikoli podatek o govorcu ali govornem posnetku, ki lahko nosi informacijo o govornih značilnostih samega govorca oziroma njegovih govornih okoliščinah, za katere se predpostavi, da imajo vpliv na akustične in jezikovne značilnosti posnetega govora. Z računskimi metodami obdelave signalov se namreč iz govornih signalov lahko izlušči različne govorne značilke, pri katerih se predpostavlja hierarhična razvrščenost pri njihovem odražanju tako nizkonivojskih anatomskih značilk človekovih govoril kot tudi višjenivojskih dialoških in semantičnih značilk.

Za razvoj samodejnih razpoznavalnikov govora je torej iz celotnega nabora metapodatkov smiselno ohraniti predvsem tiste, ki lahko prispevajo k boljšemu akustičnemu in jezikovnemu modeliranju govora. Pri razvoju govornih baz za razpoznavanje govora so bili tako v preteklosti metapodatki ključna informacija, na osnovi katere se je poskušala doseči ustrezna zastopanost vseh kategorij govorcev in govora, kot je bilo predvideno v specifikacijah. Glavni namen zbiranja teh metapodatkov je bil predvsem ta, da se v govorni bazi čim bolj realno odražajo okoliščine in scenariji možnih uporab samodejnih razpoznavalnikov govora (Kolář in Švec, 2008). Takšen pristop je zelo pomemben predvsem pri govornih bazah, ki obsegajo od vsaj nekaj 10 do več 100 ur govora oziroma govorcev.

Hiter tehnološki razvoj informacijsko-komunikacijskih sistemov je omogočil zbiranje in obdelavo vse večjih količin podatkov. Hkrati je prišlo tudi do izrazitega povečanja razpoložljivih računskih zmogljivosti sodobnih računalnikov, predvsem z razvojem zelo zmogljivih grafičnih procesnih enot (GPU), s katerimi se učinkovito izvajajo numerično zahtevni algoritmi t. i. globokega učenja (Gondi in Pratap, 2021). Posledica tega napredka je tudi ta, da so se za jezike z velikim številom govorcev začele pridobivati obsežne govorne baze, ki obsegajo tudi več kot 10.000 ur posnetkov govora. Tukaj gre praviloma za govorne baze, ki se pridobijo iz zelo različnih virov, kot so npr. razni mediji, spletne platforme, zvočne knjige idr. Zaradi velikega obsega takšnih baz se pridobljeni govorni posnetki pogosto ne označujejo in ne transkribirajo ročno. Za učenje razpoznavalnikov govora se potem uporabljajo nenadzorovani ali delno nadzorovani pristopi, ki ne zahtevajo ročno narejenih oznak in transkripcij govornih posnetkov (Hershey et al., 2017). Tako postane v večini primerov zelo obsežnih govornih baz dosledno uravnoteževanje govornih posnetkov na osnovi metapodatkov drugotnega pomena. Glede na zelo različne možne vire in načine zbiranja govornih posnetkov namreč pogosto tudi ni možno pridobivati relevantnih metapodatkov. V primerih, ko so metapodatki sicer na voljo, vendar jih je v govorni bazi težko uravnotežiti, pa pride v ospredje znamenit izrek Roberta Mercerja iz leta 1985, da ni boljših podatkov, kot je več podatkov.

Novi metapodatkovno neuravnoteženi pristopi k izdelavi govornih baz so dobili dodatno podporo pri postopkih globokega učenja, kjer se vse bolj pogosto uporabljajo metode samodejnega povečevanja obsega in plemenitenja učnih podatkov. Izvorni govorni posnetki se lahko tako s pomočjo sodobnih metod digitalne obdelave signalov modificirajo v različne simulirane oblike. Takšni osnovni pristopi so, denimo, pohitritve ali upočasnitve govora v izvornih govornih posnetkih. Z vidika metapodatkov, ki se navadno upoštevajo pri razvoju razpoznavalnikov govora, pa so se razvili tudi zahtevnejši pristopi, pri katerih se simulirajo različne snemalne okoliščine (npr. značilnosti kanala, nivo šuma, kodirniki, zvočna ozadja, prostor idr.). S takšnimi pristopi lahko učinkovito dopolnimo obseg izvornih govornih posnetkov

in uravnotežimo primanjkljaj določenih vrst govornih posnetkov (Karafiát et al., 2017).

Pri zasledovanju osnovnega cilja, da govorna baza čim bolje odraža možne okoliščine in scenarije uporabe razpoznavalnikov govora, je smiselno postaviti določene prioritete pri upoštevanju metapodatkov in njihovi uravnoteženosti. Za razvoj splošnega samodejnega razpoznavalnika govora je tako priporočljivo upoštevati predvsem naslednje metapodatke:

- Oznaka govorca: enoznačno identificira vse posnetke istega govorca v bazi. To omogoča učinkovito izvajanje metod prilagajanja modela razpoznavalnika govora na posamezne govorce (npr. metode MLLR, SAT, iVector idr.) (Povey et al., 2008; Cardinal et al., 2015), kar lahko prispeva k znatnemu izboljšanju pravilnosti samodejnega razpoznavanja govora.
- Prvi jezik: samodejno razpoznavanje govora za določen jezik je navadno bistveno manj uspešno pri govorcih, ki jim ta jezik ni prvi. Zato se pri razvoju splošnega razpoznavalnika govora njihov govor navadno izloči iz učnega postopka in se potem izvajajo posebne prilagoditve splošnega razpoznavalnika takšnim govorcem.
- Narečna skupina (Draxler in Kleiner, 2017): metapodatek je še posebej pomemben v primerih spontanega nejavnega govora. V primeru izrazitega narečnega govora je namreč možno uporabiti različne pristope adaptacije razpoznavalnika govora na narečja govorcev, s čimer se lahko do neke mere odpravi poslabšanje rezultatov.
- Snemalne zvočne okoliščine (Zhang et al., 2018): imajo lahko bistven vpliv na zanesljivost samodejnega razpoznavanja govora. Njihov vpliv je delno možno tudi simulirati ali ga odstranjevati s postopki robustne obdelave in izboljševanja kakovosti govornih signalov.
- Spol in starost govorca: v primeru splošnega razpoznavalnika govora je pri tvorjenju akustičnega modela govora pomembna uravnoteženost govorcev po teh dveh kategorijah. Adaptacija razpoznavalnika govora na spol in starost govorca se sicer redko izvaja, saj se uporablja predvsem sprotno prilagajanje modela razpoznavalnika govora na posameznega govorca. Se pa ta informacija lahko uporabi pri razvoju in preizkušanju tovrstnih metod za ugotavljanje njihove odvisnosti od teh dveh metapodatkov.

Če predstavljeni metapodatki v neki govorni bazi niso na voljo, jih je z določeno zanesljivostjo možno tudi naknadno samodejno določiti z različnimi postopki samodejnega razpoznavanja govornih vzorcev, kot so postopki biometričnega razpoznavanja in grozdenje govorcev ali razpoznavanje prvega jezika govorca. Takšni naknadno samodejno določeni metapodatki seveda lahko vsebujejo tudi napake, kar je potrebno upoštevati pri njihovi uporabi.

4. Metapodatki o posnetkih in govorcih v govorni bazi Artur

Leta 2020 se je začel nacionalni projekt Razvoj slovenščine v digitalnem okolju,[1] ki sta ga sofinancirala Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020. Projekt je izvajal konzorcij 12 partnerjev, od tega 6 javnih raziskovalnih zavodov in 6 podjetij. Naslavljal je več sklopov jezikovnih tehnologij, med njimi tudi govorne tehnologije, kjer je bilo veliko pozornosti namenjene izdelavi govorne baze za razvoj razpoznavanja govora v obsegu 1000 ur. Pomanjkanje ustrezno velike, zahtevam razpoznavanja govora prilagojene in prosto dostopne govorne baze se je namreč pokazalo kot osrednja ovira pri razvoju razpoznavanja govora za slovenski jezik. Pri izdelavi govorne baze so sodelovali Univerza V Mariboru (FERI), Univerza v Ljubljani (FE in FRI), ZRC SAZU, Alpineon in STA. Vključuje 4 večje sklope različnih vrst govora: brani govor po pisnih predlogah (500 ur), javni govor (javni dogodki, mediji ipd. – 200 ur), parlamentarni govor (Državni zbor RS – 200 ur) in nejavni govor (terenski posnetki prosto govorjenih monologov in dialogov).

Podatki o posnetkih in govorcih so v bazi Artur organizirani kot tsv-datoteka in v obliki xml-zapisa po standardu TEI. V primerjavi s predhodnimi govornimi viri za slovenščino vključujejo predvsem zelo podroben popis tehničnih lastnosti posnetkov (npr. podatke o lastnostih izvornih posnetkov in tehnični opremi, uporabljeni za snemanje) ter vseh okoliščin, ki bi lahko na te lastnosti vplivale (od velikosti prostora snemanja, prisotnosti hkratnega govora vse do uporabe maske pri govorcih, ki je bila pogosta v času epidemije COVIDA-19).

Končni seznam metapodatkov o posnetkih v govornih bazi Artur je naslednji:

I. Identifikacijski podatki in kategorizacija posnetkov:

- ID-posnetka: je sestavljen iz imena baze (Artur), podatka o tipu govora (brani – B, javni – J, nejavni – N in parlamentarni govor – P), štirimestne identifikacijske številke govorca (Gxxxx), šestmestne identifikacijske številke posnetka (Pxxxxxx) ter podatka o vrsti datoteke (-avd). Pri posnetkih javnega govora, na katerih se običajno pojavlja večje število govorcev, je namesto štirimestne identifikacijske številke govorca navedba Gvecg (s pomenom *več govorcev*). Primer ID-posnetka: *Artur-N-G5134-P600134-avd*.
- Vrsta govornega dogodka: označuje, ali gre za javni, nejavni, parlamentarni ali brani govor (Žganec Gros in Vesnicer, 2020).
- Opisi govornih dogodkov oz. topiki: Pri parlamentarnem govoru je govorni dogodek vedno označen kot seja državnega zbora. Pri javnem govoru so govorni dogodki opredeljeni kot okrogle mize, intervjuji, nagovori na dogodkih, novinarske konference ipd. oziroma kot spletni dogodek, kadar gre za posnetke, posnete na daljavo. Pri branem govoru so govorni dogodki opisani kot branje vnaprej pripravljenih pisnih predlog ali kot dva različna tipa črkovanja. Izbrani nabor kratic so govorci črkovali z dodajanjem samoglasnikov (npr. *ef a ku*), vnaprej določene pare imen in priimkov pa z dodajanjem polglasnikov (npr. *jə o nə a sə*). Če je govorec črkoval na nepredviden način, je topik poimenovan kot črkovanje s samoglasniki (oz. soglasniki) z odstopanjem *(*npr. *ef fa ku*), če je med branjem tudi kaj

[1] https://www.slovenscina.eu/

dodal ali komentiral, pa kot črkovanje s samoglasniki (oz. polglasniki) s komentarjem. Pri nejavnem govoru sta za govorne dogodke uporabljeni oznaki prosti dialog med dvema sogovornikoma in prosti monološki govor – pri slednjem govorec prosto opisuje različne stvari, recimo svoj najljubši film. Za potrebe razvoja specializiranih razpoznavalnikov v projektu Razvoj slovenščine v digitalnem okolju so v bazi Artur opredeljeni še govorni dogodki, kjer je snemanje potekalo po vnaprej pripravljenih scenarijih z dveh področij: opisovanje obrazov in upravljanje pametnega doma.

II. Podatki o okoliščinah snemanja:

- Datum snemanja je zapisan v obliki »mesec leto« (npr. *april 2021*).
- Podatek o občini snemanja temelji na seznamu občin v Republiki Sloveniji v času snemanja (2020–2022).
- Prostor snemanja natančneje opredeljuje, kje je govorni dogodek posnet, na primer v stanovanju ali pisarni, studiu ali premičnem snemalnem studiu, v dvorani, parlamentu ali pa je snemanje potekalo v odprtem prostoru.
- Velikost prostora je razdeljena v tri kategorije: do 20 m^2, od 20 do 80 m^2 in nad 80 m^2.
- Prisotnost šuma označuje, ali se na posnetku občasno pojavlja šum v ozadju, kot je šelestenje, šumenje, prometni hrup, zvok ventilatorja ipd. Če se šum po osebni presoji validatorja posnetkov pojavlja v preveliki meri, je tak posnetek uvrščen v skupino izločenih posnetkov.
- Presluh se občasno pojavi pri 2-kanalnem snemanju nejavnega govora, ko je spontani pogovor dveh sogovornikov posnet z dvema ločenima mikrofonoma. Prisotnost presluha je označena, če se pri takem snemanju pogosto in jasno sliši govor govorca z drugega kanala.
- Pogost hkratni govor je zabeležen pri nejavnem govoru, ko je sneman zasebni pogovor med dvema sogovornikoma, ki pogosto hkrati govorita.
- Podatek o tem, ali govorec nosi masko, je bil aktualen v času epidemije COVIDA-19, ko je veliko javnih dogodkov potekalo ob uporabi obrazne maske. To pomembno vpliva na akustične značilnosti posnetka. Posamezni redkejši posnetki te vrste, ki so bili uvrščeni v bazo Artur, so zato ustrezno označeni.

III. Podatki o formatu izvornih posnetkov:

- Najpogostejši formati izvornih posnetkov so WAV, MP3 in M4A.
- Čeprav so vsi posnetki v bazi Artur pretvorjeni v enotni format WAV, 44,1 kHz, pcm, 16-bit, mono, so bili posamezni posnetki, pridobljeni iz nelastnih virov, posneti v drugačnih formatih. Kadar so bile informacije dostopne, je bil izvorni format posnetkov popisan glede na frekvenco vzorčenja, bitno hitrost in bitno ločljivost.

IV. Podatki o opremi, uporabljeni za snemanje:

- Najpogosteje uporabljene snemalne naprave za posnetke v bazi Artur so prenosni ali namizni računalnik, prenosni snemalnik, pametni telefon, kamera in diktafon.
- Podatki o tehničnih lastnostih snemalne opreme zajemajo: opis naprave (npr. *MacBook PRO, Asus Vivobook, Zoom H4n, Zoom H1n*), naziv operacijskega sistema (npr. *iOS 14.2.1, Windows 10*), podatek o morebitnem mešalniku zvoka (npr. *Focusrite Scarlett 2i2 3rd Gen*), adapterju in opisu njegovega modela (npr. *Yamaha Audiogram 6*), vrsti mikrofona (npr. *namizni, vgrajeni* ali *studijski mikrofon*), modelu mikrofona (npr. *Samson Q2U*) in snemalnem programu (npr. *Adobe Audition 12, Audacity 2.3.2, Premiere Pro 14.0, Zoom, MS Teams*).

V. Podatki o viru posnetkov:

- Vir posnetka je lahko lastni posnetek, ki ga je naredila ekipa govorne baze Artur namensko za to bazo – to so vsi posnetki branega in nejavnega govora. V primeru parlamentarnega in javnega govora pa gre za arhivsko ali drugo gradivo, pridobljeno od različnih gradivodajalcev: Državni zbor RS, STA, Arnes, ZRC SAZU, Univerza v Mariboru, SDJT, Radio Štajerski Val idr.
- Pri javnem govoru je za posnetke večkrat na voljo tudi spletna povezava do videa.

Mnogi metapodatki o posnetkih večkrat niso bili dostopni. To velja zlasti za posnetke, ki niso bili lastni, ampak pridobljeni iz drugih virov, torej pri javnem in parlamentarnem govoru. Posnetki so bili uvrščeni v bazo, tudi če so kakšni metapodatki o njih manjkali, saj zlasti za javni govor ne moremo pričakovati, da bodo že obstoječi posnetki dokumentirani z metapodatki tako podrobno, kot je to mogoče, kadar snemamo namenoma za uvrstitev posnetka v govorno bazo.

Končni seznam metapodatkov o govorcih v govorni bazi Artur je naslednji:

I. Identifikacijski in sociodemografski metapodatki:

- ID-govorca zajema ime baze (Artur), oznako vrste govornega dogodka (B, J, N in P) ter vnaprej določeno štirimestno identifikacijsko številko govorca (Gxxxx). Primer ID-govorca: *Artur-N-G5097*.
- Spol (moški, ženski, drugo) je minimalno določljiv metapodatek o govorcih, tudi ko govor ni bil posnet kot lastni vir in govorci svojih sociodemografskih podatkov niso sami posredovali.
- Izobrazba je ločena v 9 kategorij: osnovna šola – nedokončana; osnovna šola – dokončana; nižje poklicno izobraževanje; srednje poklicno izobraževanje; gimnazije, SSI in PTI; višješolski programi, VS in UNI programi (1. bolonjska stopnja); magisterij stroke (2. bolonjska stopnja); magisterij znanosti (pred bolonjsko reformo); doktorat znanosti.
- Metapodatek o starosti je razvrščen v skupine: 12–17 let, 18–29 let, 30–49 let, 50–59 let, 60+ let.

II. Metapodatki o regiji govorca:

- Občina stalnega bivališča vključuje tako občine v Republiki Sloveniji kot stalno bivališče v tujini.

- Čim celovitejša demografska uravnoteženost govorcev branega in nejavnega govora je upoštevana tudi pri statistični regiji njihovega stalnega bivališča.
- Metapodatek o občini bivanja v otroštvu pokriva diahroni vidik morebitnih narečnih vplivov na govor govorca.
- Prvi jezik. Poleg govorcev, katerih prvi jezik je slovenščina, so v bazo Artur v manjši meri vključeni tudi govorci, katerih prvi jezik je hrvaščina, srbščina, makedonščina, bosanščina, ruščina, madžarščina idr. Podatek je izpolnjen samo pri govorcih, od katerih je pridobljen neposredno, pri javnih govorcih pa samo, če se lahko z veliko verjetnostjo sklepa, da je prvi jezik slovenski.
- Značilnosti govora se nanašajo na socialno zvrstnost jezika in so bile opredeljene s strani transkriptorja standardiziranega zapisa ali validatorja posnetkov. Namenjene so v pomoč pri morebitnem prilagajanju modelov razpoznavanja govora regionalnim značilnostim, prav tako so lahko v pomoč pri analizah zvrstnosti slovenskega govora. Niso pa mišljene kot točna strokovna opredelitev zvrsti govora govorca na posnetku. Ker je podrobna teorija socialne zvrstnosti za slovenščino (Toporišič, 2000) na empiričnem gradivu težko enoumno in robustno uporabljiva, je bila poenostavljena v tri osnovne kategorije: standardni jezik, pogovorni jezik in narečje. Glede na okoliščine govora je bilo predvideno, da se v javnem in parlamentarnem govoru pojavljata bodisi standardni jezik bodisi pogovorni jezik, pri čemer smo za pogovorni jezik šteli situacijo, ko so bili v govoru govorca pogosto prisotni sistematični glasoslovni pojavi, značilni za nestandardne zvrsti. Za standardni jezik pa je bil na primer označen tudi govor, ki je imel sicer prepoznavno regionalno obarvano melodiko, vendar je bil hkrati razviden zavesten večji odmik od vsakdanjega pogovornega jezika govorca proti standardnemu – to velja zlasti za govorce iz obrobja Slovenije ali drugih neosrednjih delov Slovenije. Razlike v izgovorjavi so bile zaznane tudi pri branem govoru, ki ga pa zaradi okoliščin (branje vnaprej napisanih povedi) težko ločimo na standardni in pogovorni jezik, zato sta bili pri branem govoru uporabljeni oznaki standardna izgovorjava in nestandardna izgovorjava. Predvsem v nejavnem govoru pa je lahko prisotna tudi oznaka narečje. V kolikor je bila izbrana, je dodana tudi oznaka o vrsti narečja, ki je določena na podlagi metapodatka o občini bivanja govorca v otroštvu.
- Zadnja oznaka se nanaša na opazne izgovorne težave. Pri posameznih govorcih se namreč pojavijo kakšne posebnosti, ki so povezane na primer z izgovorom glasov r, l ali podobno.

Navedeni metapodatki bodo v bazi Artur predstavljeni s slovenskimi poimenovanji kot tudi s prevodi v angleški jezik.

5. Razhajanja v metapodatkih o posnetkih in govorcih

Govorni korpusi, ki nastajajo za potrebe jezikoslovnih raziskav, in govorne baze, pripravljene za namene razpoznavanja govora, so praviloma zelo podobni govorni viri. Zato je smiselno, da se iščejo sinergijski učinki in se vsaj del gradiva uporabi v oba namena (Žgank et al., 2014). Tako se je že baza Gos Videolectures delala z mislijo na uporabo tudi za razpoznavanje govora (Verdonik, 2018), vendar je v metapodatkih še dokaj dosledno sledila zastavljeni shemi v korpusu GOS. Tudi v projektu Razvoj slovenščine v digitalnem okolju je bil iz velikega obsega posnetkov za govorno bazo Artur izbran primeren del za nadgradnjo govornega korpusa GOS. Ob tem pa se je v veliki meri ravno v zvezi z metapodatki o govorcih in posnetkih zgodilo precej razhajanj, ki so večinoma posledica bolj natančnega popisovanja podatkov, specifik ali pa namena baze, povzročajo pa težave ob združevanju gradiv. Katere vrste metapodatkov so take, pri katerih se najpogosteje pojavljajo različne odločitve?

5.1. Metapodatki o posnetkih

Obstajajo različne kategorizacije posnetega govora, saj se te praviloma izvedejo na podlagi tega, kaj vse neki govorni vir vsebuje. GOS je tako ločeval štiri tipe diskurza: javni informativno-izobraževalni, javni razvedrilni, nejavni nezasebni in nejavni zasebni. Če primerjamo to s kategorizacijo v bazi Artur, vidimo, da se tam pojavi še kategorija parlamentarni govor, manjka pa javni razvedrilni, ki se v Arturju tako rekoč ne pojavlja, pač pa se lahko celoten javni govor uvrsti kot javni informativno-izobraževalni. Prav tako ni nejavnega nezasebnega, ki se nanaša na različne uradovalne, storitvene, trgovalne in druge podobne nezasebne govorne situacije v vsakdanjem življenju. Je pa prisoten brani govor, ki se nanaša na zelo specifično, za namene snemanja posnetkov za bazo Artur ustvarjeno govorno situacijo, v kateri govorci berejo vnaprej pripravljene povedi eno po eno.

Poleg krovne kategorizacije posnetkov v manjše število krovnih kategorij se tako v korpusu GOS kot v bazi Artur uporabljajo še bolj podrobne opredelitve posnetega govora glede na govorni dogodek. V korpusu GOS je zabeleženih več kot 20 vrst govornih dogodkov, prav tako v bazi Artur, pri čemer pa jih je približno polovica namenjenih opredelitvi gradiva, ki je zelo specifično za potrebe razpoznavalnikov govora (črkovanje, področno specifični razpoznavalniki za pametni dom in opisovanje obrazov). Opredelitev vrste govornega dogodka je nadvse pomembna, saj omogoča po potrebi tudi naknadno prekategorizacijo zbranega gradiva ob združevanju različnih virov, zato je verjetno eden najbolj bistvenih metapodatkov o tipih posnetkov za vsak govorni vir, bolj pomemben kot širša, krovna kategorizacija, ki se lahko naknadno tudi spreminja na podlagi razvrščanja informacij o vrstah govornih dogodkov ali deloma tudi na podlagi informacij o viru.

Obvezna metapodatka o posnetkih v govornih virih sta čas in lokacija snemanja. Medtem ko so pri času lahko razhajanja samo v večji ali manjši natančnosti zabeleženega časa, pa se pri opredelitvi lokacije pojavljajo razlike, na katere enote se pri tem naslonimo. V korpusu GOS je bil ta metapodatek opredeljen dvojno: kot kraj, torej z imenom mesta ali vasi, ki pa skozi spletni konkordančnik ni dostopen zaradi varovanja identitete govorcev, in kot regija

snemanja, ki pa jo lahko opredelimo zelo različno. V korpusu GOS se je označila na podlagi registrskih območij. V bazi Artur je metapodatek o lokaciji zabeležen kot občina snemanja. V slovenskem kontekstu se zdi (glede na veliko število in razdrobljenost občin) informacija lokacije snemanja skozi občino ustrezen kompromis. V slovenskem podeželskem okolju lahko namreč navajanje točnega kraja z imenom vasi razkriva identiteto govorcev, enote, večje od občine (npr. upravna enota, registrsko območje ali statistična regija) pa niso več zadosti natančne in skladne z narečno razpršenostjo, ki je v Sloveniji pregovorno velika.

Metapodatek o viru prinaša informacijo o izvornem nosilcu avtorskih pravic. Podobno kot za pisna besedila namreč tudi za govorjena besedila velja, da so njihovi tvorci hkrati tudi avtorji z avtorskimi pravicami[2] nad besedili in pogosto obstajajo pogodbene zaveze, da bo ta podatek v jezikovnem viru ustrezno naveden. Pri posnetkih govora se v zvezi z avtorskimi pravicami in navajanjem vira srečujemo s štirimi vrstami situacij: (1) Če gre za posnetek na terenu, ki je bil narejen za namene govornega vira in zajema avtentični govor v vsakdanjih situacijah, govorci prenesejo avtorske pravice praviloma na nosilca projekta, v katerem nastaja govorni vir. Praksa je, da je v takih primerih kot vir označeno *terenski/lastni posnetek*. (2) Če gre za posnetek, ki je bil predvajan prek radia ali televizije, so pogosto nosilci avtorskih pravic medijske hiše in so posledično te navedene kot vir. Tudi pri spletnih virih (npr. posnetki na Youtubu[3]) je treba pogosto urediti avtorske pravice z njihovim/-i nosilcem/-i in v metapodatkih ustrezno navesti vir. Če gre za spletne dogodke, ki jih sicer organizira in objavi neka institucija (npr. spletne konference, delavnice, seminarji), je pogosto treba urejati avtorske pravice z neposrednimi tvorci teh besedil. Pri tem se pojavi vprašanje, kako je najbolj smiselno definirati metapodatek o viru: kot posameznika/e, ki je/so pravice odstopil/-i in nastopa/-jo na posnetku, ali kot institucijo, ki je organizirala in objavila spletni dogodek. V bazi Artur je bila pri tovrstnih posnetkih izbrana druga možnost. (3) Določeni internetni viri že imajo urejene avtorske pravice na način, ki omogoča nadaljnjo uporabo, in sicer pod pogoji katere od licenc Creative Commons. Taka večja vira posnetkov v slovenščini sta portala Videolectures.net in Arnes Video. V takih primerih se v obstoječih bazah za slovenščino kot vir navaja kar ime portala. (4) Določena govorjena besedila niso avtorsko varovana. V skladu z 9. členom ZASP so taka »uradna besedila z zakonodajnega, upravnega in sodnega področja«. Čeprav še ni tovrstne sodne prakse ali doktrine, se lahko kot tovrstna med drugim štejejo govorjena besedila, ki nastajajo v Državnem zboru RS v okviru zakonodajnih postopkov. V tem primeru se kot vir v bazi Artur, kjer se pojavljajo tovrstni posnetki, navaja kar Državni zbor Republike Slovenije.

Druge vrste metapodatkov o posnetkih, kot smo jih predstavljali v poglavjih 2, 3 in 4, se v določenih govornih virih pojavljajo, v drugih ne, odvisno od specifičnega namena govornega vira. Pri združevanju govornih virov se lahko bodisi izpustijo bodisi ostanejo nedefinirani, če niso bili zabeleženi in niso na voljo.

5.2. Metapodatki o govorcih

Čeprav so metapodaki o govorcih manj raznovrstni kot metapodatki o posnetkih, pa se razlike, kako jih opredelimo, pojavljajo tako rekoč pri vseh kategorijah razen pri spolu.

Najzahtevnejše vprašanje je povezano s potrebo, da se zabeležijo različni regionalni vplivi na govor posameznika. V zvezi s tem sta problematični naslednji točki:

1. Opredelitev regionalnih vplivov na govor govorca ni nujno enoznačna. Tako so se na primer v dodatku h govornemu delu korpusa BNC (British National Corpus) iz leta 2014, v katerem so zajemali samo vsakdanje pogovore, prepustili govorcem, da so sami s svojimi besedami opisali svoj dialekt, in nato te opise preslikali v shemo statističnih teritorijskih enot Velike Britanije (Love et al., 2017). Tudi v slovenskih govornih virih se je uveljavila praksa, da se regija govorcev beleži skozi geopolitične, in ne geolingvistične kategorije. Razlog je bržkone ta, da lahko zanesljive geolingvistične kategorizacije naredi samo stroka, in to naknadno, na podlagi zbranih podatkov. V korpusu GOS so bile tako kategorije za regijo govorcev definirane na podlagi registrskih območij, ki jih je za Slovenijo skupno 11, k temu pa so bile dodane še kategorije za zamejske Slovence (Avstrija, Italija, Madžarska) in govorce, ki jim slovenščina ni prvi jezik (tujina). Taka razdelitev je izredno ohlapna in nenatančna v primerjavi s slovensko dialektalno razpršenostjo. Tudi sam koncept »regionalna pripadnost«, zveden na registrsko označbo na avtomobilu, se zdi neustrezen, čeprav ima za teren zelo koristno lastnost robustnosti. V bazi Artur se je zato iskala bolj natančna, enoznačna, enostavna in manj sporna opredelitev metapodatka, ki bi nosil informacije o regiji govorcev. Ker smo ime kraja, zlasti ko gre za podeželsko okolje, že izpostavili kot problematično zaradi potencialnega razkrivanja identitete govorca, je bila kot osnovna enota izbrana občina. Slovenija je v času zbiranja posnetkov za bazo Artur razdeljena na 212 občin. Prednost te kategorije je tudi ta, da je mogoče občine enostavno enoznačno preslikati na širše geopolitične enote – 12 statističnih regij Slovenije, kot jih v času nastajanja baze definira Statistični urad Republike Slovenije.
2. Marsikdo danes ne živi vse življenje v nekem omejenem geografskem prostoru, ki je govorno homogen, pač pa je veliko ljudi mobilnih, bodisi z dnevnimi/tedenskimi migracijami zaradi šolanja ali zaposlitve bodisi zaradi selitev. Slika regionalnih vplivov na govor govorca je zato lahko pri določenih posameznikih izredno kompleksna in hkrati včasih tudi zelo specifična. Korpus GOS je tako omogočal, da so govorci zase izbrali skupno tudi do pet »regionalnih pripadnosti«. Tako nastane precej kompleksna slika, saj

[2] Termin avtorske pravice tukaj uporabljamo za vse materialne avtorske pravice, druge pravice avtorja v skladu z ZASP in avtorski sorodne pravice, ki utegnejo nastati pri snemanju. O vprašanjih osebnostnih pravic in varstva osebnih podatkov, ki so prav tako pomembna za vsako uporabo posnetkov v govornih virih, tukaj ne razpravljamo, saj ni relevantno v kontekstu tega članka. Bralca samo opozarjamo, da uporabo posnetkov za govorne vire ovira tudi ta pravni vidik.

[3] Sama licenca Youtube ne omogoča uporabe posnetkov za govorni vir.

dobimo poleg govorcev s samo eno regijo še precej govorcev z zelo različnimi kombinacijami regij, med katerimi pa posamezna kombinacija ne zajema veliko govorcev. Na koncu je za slednje najbrž smiselno zabeležiti samo eno skupno kategorijo »različni regionalni vplivi«, kot naredijo v korpusu C-ORAL-ROM (Cresti in Moneglia, 2005). V bazi Artur je bila opredelitev geografske mobilnosti skozi čas poenostavljena na dve vrsti metapodatkov, prva se nanaša na občino bivanja v otroštvu, druga na občino trenutnega stalnega bivališča. S tem se izgubi precej informacij o morebitni dodatni mobilnosti posamezne osebe, ki bi sicer bile pomembne za podrobno analizo govora posameznega govorca, vprašljivo pa je, koliko so relevantne za (kvantitativno) korpusno analizo ali za morebitno prilagajanje razpoznavalnika govorcem po regijah.

Določenemu delu govorcev slovenščina ni prvi jezik. Tudi to je podatek, ki je za govorni vir, če se v njem tovrstni govorci pojavljajo, zelo pomemben. Niti iz korpusa GOS niti iz baze Artur se nematerni govorci slovenščine niso izključevali, pač pa nasprotno – namenoma vključevali. S tem je v obeh virih bistven tudi metapodatek o prvem jeziku govorca.

Niti metapodatek o geografski pripadnosti govorca niti metapodatek o prvem jeziku pa še ne povesta, kakšen je dejansko govor nekega govorca v govornem viru z vidika socialnozvrstne delitve. Slednjo lahko ugotavljamo šele na podlagi (zlasti) slušne analize govora. Ne gre torej za metapodatek, ki ga zabeležimo na terenu, ampak za naknadno interpretacijo govornih podatkov. V korpusu GOS se ni delala, v bazi Artur pa je bila izražena tovrstna želja za potrebe razpoznavalnikov govora.

Izobrazba in starost govorcev sta metapodatka, preko katerih predvsem zagotavljamo ustrezno demografsko razpršenost govorcev, zajetih v govorni vir. Za posnetke javnega govora večinoma niti nista dostopna in posledično za velik del posnetkov v korpusu GOS in bazi Artur teh metapodatkov ni. Kjer pa sta na voljo, so skupine glede na starost in izobrazbo delno različno opredeljene in različno podrobne, kar otežuje združevanje virov. Minimalne kategorije starostnih skupin so po našem mnenju skupina najstnikov (okvirno do 19 let), skupina upokojencev (okvirno nad 60 let) in vse ostalo vmes. V kategoriji izobrazbe imamo 4-stopenjsko delitev v GOS-u in 9-stopenjsko delitev v Arturju. Po našem mnenju minimalna delitev je vsaj v dve skupini glede na to, ali je oseba zaključila izobraževanje po srednji šoli ali pa nadaljevala šolanje. Večja podrobnost metapodatkov o govorcih bi bila zanimiva verjetno predvsem za sociolingvistične raziskave, zato je potrebna ustrezna previdnost pred prehitrim posploševanjem v zelo grobe kategorije.

6. Zaključek

V prispevku smo obravnavali metapodatke o posnetkih in govorcih, ki se tipično uporabljajo v govornih jezikovnih virih. Osredotočili smo se na obstoječe prosto dostopne govorne vire korpusnega tipa za slovenski jezik, tj. referenčni govorni korpus GOS, bazo Gos Videolectures in govorno bazo v nastajanju znotraj projekta Razvoj slovenščine v digitalnem okolju, Artur. Govorni podatki iz teh treh baz namreč predstavljajo vir podatkov za razširitev referenčnega govornega korpusa GOS, ob tem pa se kažejo težave z združevanjem, ki med drugim[4] izhajajo tudi iz razlik v popisu in kategorizacijah metapodatkov o posnetkih in govorcih.

V prihodnje bi si želeli večjo homogenizacijo metapodatkov o posnetkih in govorcih zlasti tam, kjer gre za ključne metapodatke, ki so bistveni tako za spremljanje uravnoteženosti gradiv kot za kategoriziranje govornih podatkov. Pri posnetkih so taki ključni metapodatki: (1) opis govornega dogodka, ki mora biti zadosti podroben in se lahko razume v smislu govornih situacij, ki imajo večje število skupnih kontekstnih lastnosti, vključno z vrsto lokacije, vrsto razmerja med tvorci in naslovniki, namenom in kanalom komunikacije; (2) čas in lokacija snemanja, pri čemer je zlasti pri lokaciji pomembno, da je zadosti podrobna, npr. ime kraja ali občine, kjer poteka snemanje; (3) vir posnetka, pomemben zaradi korektne obravnave avtorskih pravic, v pomoč je lahko tudi pri sortiranju govornih podatkov po tipih, zaradi naknadnega dostopa do video vsebine pa je skoraj nujna tudi povezava do videoposnetka, če obstaja; (4) vedno koristni in zaželeni, a morda manj nujni pa so tudi vsi razpoložljivi podatki o snemalni opremi in tehničnih lastnostih posnetka. Pri govorcih so ključni metapodatki o: (1) identifikaciji, (2) spolu, (3) starosti, (4) prvem jeziku in (5) regiji/-ah, pri čemer mora biti slednja zadosti podrobno opredeljena (npr. na ravni kraja ali občine) in vsaj v grobem upoštevati tudi diahroni vidik. Pogosto prisoten je tudi metapodatek o (6) stopnji izobrazbe, medtem ko beleženje metapodatkov o poklicih, socialnem sloju ali pripadnostih različnim družbeno-kulturnim skupinam v slovenskih govornih virih do zdaj ni bilo prakticirano.

V članku med drugim predstavljamo tudi podroben opis metapodatkov o posnetkih in govorcih v govorni bazi Artur. Pri določanju metaoznak se je pokazalo, da pri čisto vseh kategorijah vnosi niso enoznačni in enostavno določljivi. Pri metaoznakah, nanašajočih se na govorce, je bila največji izziv kategorija značilnosti govora, saj je bil odločevalec pogosto soočen z dilemo, ali je jezik še standardni ali pogovorni oz. ali je pogovorni ali narečje. Kot pišemo v poglavju 4, so bili sistematični glasoslovni pojavi, značilni za nestandardne zvrsti, odločilni kriterij, da gre za pogovorni jezik; in nasprotno, opazno prizadevanje govorca, da bi uporabljal standardni jezik, čeprav je v njegovem govoru še vedno mogoče zaznati regionalno obarvano melodiko, je bilo odločilno za oznako standardni jezik. Če je bil govor označen kot narečni, smo se za točno določitev vrste narečja oprli na podatek o občini bivanja v otroštvu. Preostali metapodatki o govorcih so bili bodisi pridobljeni neposredno od govorcev bodisi jih nismo določali. Izjema je spol govorca, ki smo ga določili na podlagi posnetka, tudi ko ni bilo neposredne informacije. Pri javnih govorcih, za katere nismo imeli neposrednih informacij, a smo lahko z veliko verjetnostjo sklepali, da je njihov prvi jezik slovenski, je lahko bil dodan tudi ta podatek. Veliko izzivov je bilo tudi pri pridobivanju metapodatkov o posnetkih, saj je v primeru, ko ni podatkov s terena, izjemno težko sklepati o vrsti in velikosti prostora snemanja ali identificirati podatke o datumu in občini

[4] Določene razlike so sicer tudi v pravilih zapisovanja govora. V tem članku se osredotočamo samo na metapodatke o posnetkih in govorcih.

snemanja dogodka ter nemogoče zagotoviti natančen tehnični popis snemalne opreme. V bazi Artur so bili ti metapodatki vpisani samo, ko so bili znani.

Baza Artur je prioritetno namenjena razvoju modelov razpoznavanja govora, vendar lahko s svojim izredno podrobnim popisom metapodatkov predstavlja izhodišče pri morebitni nadgradnji ali razvoju podobnih govornih virov v prihodnosti. Po zaključku, od novembra 2022 naprej, bo prosto dostopna prek repozitorija CLARIN.SI pod licenco Creative Commons.

7. Literatura

Jens Allwood, Maria Björnberg, Leif Grönqvist, Elisabeth Ahlsen in Cajsa Ottesjö. 2000. The spoken language corpus at the Linguistics Department, Göteborg University. *Forum Qualitative Social Research,* 1(3).

Lou Burnard, ur. 2007. *Reference guide for the British National Corpus (XML Edition)*. URL: http://www.natcorp.ox.ac.uk/XMLedition/URG/.

Patrick Cardinal, Najim Dehak, Yu Zhang in James Glass. 2015. Speaker adaptation using the i-vector technique for bottleneck features. V: *Proceedings of Interspeech 2015*, str. 2867–2871.

Emanuela Cresti in Massimo Moneglia, ur. 2005. *C-ORAL-ROM: Integrated reference corpora for spoken romance languages*. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Christoph Draxler, Stefan Kleiner. 2017. A cross-database comparison of two large German speech databases. V: *Proceedings of the 18th International Congress of Phonetic Sciences,* Glasgow, UK, 10.–15. avgust 2015. International Phonetic Association.

Oliver Ehmer in Camille Martinez. 2014. Creating a multimodal corpus of spoken world French. V: Sükriye Ruhi, Michael Haugh, Thomas Schmidt, Kai Wörner, ur., *Best Practices for Spoken Corpora in Linguistic Research*, str. 142–161. Cambridge Scholars Publishing, Newscastle upon Tyne.

Santosh Gondi in Vineel Pratap. 2021. Performance Evaluation of Offline Speech Recognition on Edge Devices. Electronics 2021, 10, 2697. MDPI, Basel, Switzerland.

John R. Hershey, Jonathan Le Roux, Shinji Watanabe, Scott Wisdom, Zhuo Chen in Yusuf Isik. 2017. Novel deep architectures in speech processing. V: *New Era for Robust Speech Recognition*, str. 135–164. Springer.

Martin Karafiát, Karel Veselý, Kateřina Žmolíková, Marc Delcroix, Shinji Watanabe, Lukáš Burget, Jan "Honza" Černocký in Igor Szőke. 2017. Training data augmentation and data selection. V: *New Era for Robust Speech Recognition*, str. 245–260. Springer.

Jáchym Kolář in Jan Švec. 2008. Structural Metadata Annotation of Speech Corpora: Comparing Broadcast News and Broadcast Conversations. V: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina in Tony McEnry. 2017. The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics,* 22(3):319–344.

Nelleke Oostdijk, Wim Goedertier, Frank Van Eynde, Lou Boves, Jean-Pierre Martens, Michael Moortgat in Harald Baayen. 2002. Experiences from the Spoken Dutch corpus project. V: M. González Rodriguez, C. Paz Suárez Araujo, ur., *Proceedings of the third international conference on language resources and evaluation* (LREC'02), str. 340–347. Las Palmas, Kanarski otoki. ELRA.

Petr Pořízka. 2009. Olomouc corpus of Spoken Czech: Characterization and main features of the project. *Linguistik online*, 38(2). http://www.linguistik-online.de/38_09/porizka.html.

Daniel Povey, Hong-Kwang J. Kuo in Hagen Soltau. 2008. Fast speaker adaptive training for speech recognition. V: *Proceedings of Interspeech 2008*, str. 1245–1248.

Jože Toporišič. 2000. Slovenska slovnica. Založba Obzorja, Maribor.

Darinka Verdonik. 2018. Korpus in baza Gos Videolectures. V: Darja Fišer, Andrej Pančur, ur., *Zbornik konference Jezikovne tehnologije in digitalna humanistika,* str. 265–268. Znanstvena založba Filozofske fakultete, Ljubljana.

Darinka Verdonik in Ana Zwitter Vitez. 2011. *Slovenski govorni korpus Gos*. Trojina, zavod za uporabno slovenistiko, Ljubljana.

Jana Zemljarič Miklavčič. 2008. *Govorni korpusi*. Znanstvena založba Filozofske fakultete, Ljubljana.

Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa in Wenyu Jin, Björn Schuller. 2018. Deep learning for environmentally robust speech recognition: An overview of recent developments. V: *ACM Transactions on Intelligent Systems and Technology (TIST) 9.5*, str. 1–28.

Jerneja Žganec Gros in Boštjan Vesnicer. 2020. Izbor fonetično uravnoteženih besedilnih predlog za bazo branega govora. V: Tanja Mirtič, Marko Snoj, ur., *Razprave II. razreda SAZU: 1. slovenski pravorečni posvet,* str. 111–119. Slovenska akademija znanosti in umetnosti, Ljubljana.

Andrej Žgank, Ana Zwitter Vitez in Darinka Verdonik. 2014. The Slovene BNSI broadcast news database and reference speech corpus GOS: Towards the uniform guidelines for future work. V: Nicoletta Calzolari, ur., *LREC 2014: proceedings of the Ninth International Conference on Language Resources and Evaluation*, str. 2644–2647, Reykjavik, Islandija. ELRA.

Uporaba Europeaninega podatkovnega modela (EDM) pri digitalizaciji kulturne dediščine: primer Skuškove zbirke iz Slovenskega etnografskega muzeja v projektu PAGODE-Europeana China

Maja Veselič,* Dunja Zorman,†

Oddelek za azijske študije, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
*maja.veselic@ff.uni-lj.si
† dunja.zorman @ff.uni-lj.si

Povzetek

V prispevku predstaviva podatkovni model baze Vzhodnoazijske zbirke v Sloveniji in njegovo prilagoditev Europeaninem podatkovnem modelu za potrebe objave podatkov več kot 900 predmetov kitajske kulturne dediščine, pretežno fotografij, v Europeani. Podrobno opiševa proces oblikovanja prilagojenega modela ter postopek priprave podatkov na uvoz, ki je potekal s pomočjo orodja MINT. Na koncu podava nekaj refleksij o pozitivnih učinkih te izkušnje na delo z izvorno bazo.

Application of the Europeana Data Model (EDM) in digitalization of cultural heritage: The example of the Slovene Ethnographic Museum's Skušek Collection in the PAGODE-Europeana China project

This paper first introduces the data model of the East Asian Collections in Slovenia database. It then details how it was adjusted to the Europeana Data Model for the purpose of publishing in Europeana, more than 900 objects of Chinese cultural heritage, mostly photographs. It recounts the steps taken in creating the adjusted model and the procedure of data preparation for the import by using the MINT tool. It concludes with a reflection on the positive impacts of this experience on the work with the original database.

1. Uvod

Zadnjih nekaj let države in nadnacionalne organizacije spodbujajo institucije, ki hranijo in varujejo kulturno dediščino – galerije, knjižnice, arhive in muzeje, k pospešeni digitalizaciji kulturne dediščine. Ta naj ne bi zgolj zaščitila in ohranjala kulturne dediščine ali olajšala dostop do materialne in nematerialne dediščine za raziskovalne, izobraževalne ali ljubiteljske namene, temveč naj bi stimulirala gospodarsko rast skozi promocijo kreativnosti in inovacij, npr. v turizmu z novimi digitalnimi turističnimi produkti ali kot vir idej in navdiha v t. i. kreativnih industrijah.[1]

Toda da bi bila digitalizirana dediščina resnično dostopna in uporabna, da bi torej uporabnik pri iskanju lahko dobil čim več zadetkov, ki čimbolj natančno zadostijo iskalnim parametrom, da bi prišel do relevantnih zadetkov, tudi če so podatki zapisani v drugem jeziku, kot je jezik iskanja in da bi zadetke lahko po različnih parametrih dodatno filtriral, je digitalizirane predmete nujno opremiti s čimbolj kakovostnimi metapodatki. Ti močno olajšajo selekcijo gradiva glede na specifične potrebe konkretnih uporabnikov, kar med drugim pripomore h kvalitetnejšemu kuriranju (npr. v obliki digitalnih galerij, razstav) in lažjemu vizualiziranju vsebin.

V prispevku predstaviva izkušnje s prilagoditvijo podatkovnega modela baze Vzhodnoazijske zbirke v Sloveniji (VAZ) Europeaninem podatkovnem modelu (Europeana Data Model, v nadaljevanju EDM) pri uvozu izbranih digitaliziranih predmetov iz Skuškove zbirke na evropsko digitalno knjižnico Europeana. Gre za del zbirke pretežno kitajskih predmetov, ki jih je iz Pekinga v Ljubljano leta 1920 prinesel mornariški častnik Ivan Skušek ml. in jih danes hrani Slovenski etnografski muzej. Ti predmeti so bili digitalizirani in v Europeani objavljeni v okviru projekta PAGODE-Europeana China (2020–2021, v nadaljevanju PAGODE),[2] medtem ko je analiza predmetov in ustvarjanje opisnih (vsebinskih) podatkov potekalo v okviru projektne skupine Vzhodnoazijske zbirke v Sloveniji (2018–2021, VAZ),[3] ki z istoimensko podatkovno bazo in portalom predstavlja tudi izvorno lokacijo digitalnih fotografij predmetov.

Za nekoga, ki se prvič srečuje metapodatki in podatkovnimi bazami in ob tem nima tehnično strokovnega znanja, je soočenje z EDM-om in obdelavo podatkov za uvoz in objavo v Europeani zastrašujoče. Skozi refleksijo lastnih napak in končnega uspeha, želiva tiste, ki oklevajo glede objave svojega gradiva v Europeani, k temu spodbuditi.

2. Europeanin podatkovni model (EDM)

Evropska digitalna knjižnica Europeana, ki jo financira Evropska unija, sodi med najpomembnejše zbirke digitalne kulturne dediščine v Evropi. Danes v Europeani najdemo gradivo več kot 4000 posameznih institucij, ki obsega nekaj deset milijonov slikovnih in tekstovnih datotek, skoraj milijon avdio- in videoposnetkov, pa tudi več kot 8000 3D modelov.[4] Poudariti je treba, da knjižnica na svojih strežnikih ne hrani digitaliziranih predmetov kulturne dediščine,[5] temveč so ti dostopni preko povezav na različne institucionalne in nacionalne repozitorije in baze. Europeana digitalizirane predmete le prikazuje in objavlja njihove (meta)podatke. Europeana tudi ni v neposrednem stiku s posameznimi institucijami, temveč podatke pridobi

[1] https://digital-strategy.ec.europa.eu/en/news/commission-proposes-common-european-data-space-cultural-heritage.
[2] https://photoconsortium.net/pagode/.
[3] https://vazcollections.si/.
[4] https://www.europeana.eu/en/about-us.
[5] Izraz predmet ne označuje zgolj materialnega predmeta, saj so v Europeani predstavljeni tako predmeti snovne kot nesnovne dediščine ter predmeti, ki so bili že ustvarjeni digitalno.

od agregatorjev, ki zbirajo, pregledujejo in pred uvozom v Europeano obogatijo podatke, ki jih posredujejo različne kulturne in dediščinske institucije ali organizacije (t. i. ponudniki vsebin). Številni, a ne vsi agregatorji so organizirani kot posredniki na nacionalni ravni. Za Slovenijo to vlogo opravlja Nacionalna in univerzitetna knjižnica v Ljubljani.[6]

V primeru Europeane torej poseben izziv predstavlja številčnost institucij, ki tam objavljajo svoje vsebine, in raznolikost načinov organizacije (meta)podatkov, ki so jih oblikovale skozi svoje institucionalne zgodovine in prakse. Nekateri metapodatkovni standardi so sicer močno razširjeni, na primer LIDO, ki ga je razvil Mednarodni muzejski svet (ICOM) in ga uporabljajo številni muzeji. Toda v Europeano prihaja gradivo različnih vrst kulturnih institucij, gradivo različnih tipov, poleg tega predstavlja knjižnica tudi večjezikovno okolje. Pri Europeani so zato razvili svoj lastni model metapodatkov ter vanj integrirali elemente pred tem uveljavljenih standardov, zlasti ORE, DCMI, SKOS in CRM.

EDM metapodatke deli na tri jedrne razrede (angl. *core classes*): (1) metapodatki, vezani na predmet kulturne dediščine, ki je digitaliziran (edm:ProvidedCHO), npr. kdaj in kje je predmet nastal, kdo ga je ustvaril, kakšne dimenzije ima, (2) metapodatki, vezani na spletni vir ali več virov, ki so vezani na predmet (edm:WebResource), npr. kakšen je format spletnega vira, kdo ga je prispeval, kakšne so avtorske pravice; ter (3) metapodatki, povezani z agregacijo, torej z mehanizmom, ki združuje zgornja dva razreda in predstavlja uvoz podatkov v Europeano, npr. kateri agregator prispeva podatke, kje so ti prikazani (ore:Aggregation) (Europeana, 2017).

Poleg tega EDM omogoča tudi kontekstualne razrede (angl. *contextual classes*). Sem sodijo metapodatki o agentu (edm:Agent), o prostoru (edm:Place), o časovnem obdobju (edm:TimeSpan), o konceptu (skos:Concept) in o licenci (cc:Licence). Med podatke o agentu na primer beležimo ljudi, ki jih je predmet v svojem življenju srečal oz. so z njim kakorkoli povezani, med tiste o prostoru pa mesta, kjer se je kdaj nahajal (Europeana, 2017).

Europeana poleg tega pri kvaliteti metapodatkov izpostavlja še dvoje: večjezičnost ter rabo nadzorovanih besednjakov. Europeana namreč prikazuje zbirke in predmete v štiriindvajsetih evropskih jezikih. V ta namen morajo biti v model vključeni podatki o jeziku, v katerem so vrednosti, tj. podatki v posameznem polju, zapisani. Poleg tega je za čim širšo jezikovno pokritost zaželena čim večja uporaba identifikatorjev iz povezanih odprtih podatkov in nadzorovanih besednjakov, kot so Wikidata, Gettyjev Arts & Architecture Thesaurus (AAT) ali geografska podatkovna baza GeoNames. Metapodatki vezani na identifikatorje se tako ne prikazujejo le v jeziku iskanja, temveč tudi v drugih evropskih jezikih, ki so vključeni v nadzorovane besednjake oz. povezane odprte podatkovne baze. Poleg tega identifikatorji služijo nadaljnjemu semantičnemu bogatenju metapodatkov. To je odlično za končnega uporabnika, saj povečuje število ključnih besed, preko katerih lahko uporabnik najde določen predmet v Europeaninem brskalniku.

Bogatost in strukturiranost podatkov torej vplivata na to, kako najdljivi so predmeti. V Europeani različne poti do predmetov imenujejo scenariji za odkrivanje[7] (angl. *discovery scenarios*) in ločijo štiri osnovne načine najdljivosti: glede na čas oziroma časovni razpon, glede na teme in tipe, glede na agente ter glede na lokacije.

Da bi spodbudili ponudnike k objavljanju čim bolj bogatih in čim bolje strukturiranih metapodatkov, so pri Europeani v zadnjih letih razvili tristopenjski standard kakovosti metapodatkov, pri čemer vsaka od ravni omogoča določeno uporabniško izkušnjo. Raven A omogoča le osnovno iskanje konkretnih predmetov, raven B omogoča raziskovanje vsebin, raven C pa predstavlja platformo znanja, saj omogoča številna presečna iskanja, med drugim tudi po specifičnih temah, motivih, barvah in drugih lastnosti predmeta kulturne dediščine (Europeana 2019b). Čeprav se v projektu PAGODE nismo zavezali k določeni ravni metapodatkov, si je večina partnerjev prizadevala doseči stopnjo C.[8]

3. Projekt PAGODE – Europeana China

Mednarodni projekt PAGODE – Europeana China (PAGODE),[9] ki je trajal 18 mesecev (2020–2021) in ga je vodilo italijansko ministrstvo za gospodarski razvoj, je združil javne in zasebne institucije s področja znanosti, kulturne dediščine in kulturnega menedžmenta z namenom, da bi obogatili, izpostavili in dodatno osvetlili kitajsko kulturno dediščino v Europeani. V projektu je sodelovalo 6 partnerjev ter 15 pridruženih partnerjev. Glavni cilj je bil v Europeano dodati 10.000 novih digitaliziranih predmetov kitajske kulturne dediščine, avtomatsko obogatiti metapodatke že obstoječim 20.000 predmetom, še 2000 predmetom pa metapodatke dodati z ročno anotacijo v obliki množične skupnostne kampanje. Drugi osrednji cilj je bil kitajsko dediščino uporabnikom Europeane predstaviti skozi kurirane vsebine – galerije, bloge, razstave ter posebno vozlišče za kitajsko dediščino.[10]



Slika 1: Kurirane vsebine kitajske kulturne dediščine na Europeani pod skupno tematsko vstopno točko.

[6] http://www.agregator.si/.
[7] Nujne metapodatkovne kategorije za posamezen scenarij so predstavljene v Charles, Isaac in Hill (2015).
[8] Europeanin okvir za objavljanje (Europeana Publishing Framework) loči tudi različne nivoje kakovosti vsebine (od 1 do 4), pri čemer merijo kakovost digitalnega zapisa (fotografije, zvočnega posnetka itd.) ter možnost njegove ponovne uporabe glede na avtorske pravice (Europeana 2019a).
[9] Projekt je financirala Evropska komisija v okviru mehanizma Connecting Europe Facilities.
[10] https://www.europeana.eu/en/chinese-heritage. Vozlišče predstavlja osrednjo zbirno točko za kurirane vsebine o kitajski dediščini na Europeani tudi po koncu projekta PAGODE.

V projektu PAGODE je za večino ponudnikov vsebin agregacijo opravil partner in akreditirani agregator Photoconsortium,[11] ki je sicer specializiran za fotografske vsebine v Europeani. Poleg tega, da je veliko sodelujočih ponudnikov vsebin v Europeano dodalo prav fotografsko gradivo, je tovrstna agregacija omogočala boljši nadzor na izpolnjevanjem ambicioznih ciljev glede novih vnosov in avtomatskega bogatenja.

Kot partner je v projektu sodeloval tudi Oddelek za azijske študije Filozofske fakultete UL, bolj natančno tri članice nacionalnega raziskovalnega projekta Vzhodnoazijske zbirke v Sloveniji (2018–2021).[12] Naša naloga je bila vzpostavitev semantične sheme projekta, ki naj bi vodila tako izbor novih predmetov (opredelitev, kaj sploh je kitajska dediščina v Evropi), kot obogatitev že obstoječih predmetov (ključne besede, ki opredeljujejo kitajsko dediščino). Čeprav z Europeano nismo imele izkušenj, pa tudi projekt VAZ se je šele dobro začel, nas je povabilo k sodelovanju pritegnilo predvsem, ker je obljubljalo dostop do dodatnih sredstev za digitalizacijo predmetov, ki smo jih nameravali vključiti v digitalno bazo VAZ.

Največja zbirka kitajskih predmetov pri nas je Skuškova zbirka v Slovenskem etnografskem muzeju (SEM), ki obsega več kot 500 predmetov manjših in večjih dimenzij, med njimi pohištvo, okrasne stene, porcelan, tekstil, slike, kipce, kadilne pripomočke, kovance, knjige, fotografije. Predmete je Ivan Skušek ml. (1877–1947), mornariški častnik, ki se je med prvo svetovno vojno znašel v internaciji v Pekingu, skupaj s svojo bodočo ženo, na Kitajskem živečo Japonko Kondō Kawase Tsuneko (1893–1963), kasneje krščeno Marija Skušek, leta 1920 pripeljal v Ljubljano. Skušek je doma nameraval postaviti muzej kitajske kulture, a so mu finančne težave to preprečile. Po moževi smrti je Marija Skušek zbirko zapustila državi in večina predmetov je pristala v Slovenskem etnografskem muzeju. Le nekaj jih je razstavljenih na stalni razstavi, mnogi med njimi pa donedavna niso bili niti spodobno inventarizirani.[13]

Na naš predlog se je projektu PAGODE kot pridruženi partner priključil SEM, ki je v ta namen pripravil digitalne fotografije približno 200 kovancev ter skenograme dveh na Japonskem izdanih tiskanih albumov arhitekturnih fotografij in skic cesarskega Pekinga, dveh naslikanih albumov s podobami kitajskega kaznovanja in vsakdana bogatih žensk in otrok in album s 450 prilepljenimi fotografijami Pekinga in okolice. Opisne podatke predmetov smo pripravili v projektu VAZ, prilagoditev podatkovne sheme, ki jo uporabljamo v bazi VAZ za potrebe uvoza v Europeano pa sva pripravili avtorici. V nadaljevanju prispevka tako najprej predstaviva podatkovno shemo, ki smo jo razvili v projektu VAZ, nato pa prilagoditev te sheme za uvoz v Europeano.

4. Podatkovna shema baze VAZ

Projekt VAZ je nacionalni raziskovalni projekt, ki je formaliziral večletna prizadevanja za sistematičen popis in znanstveno-strokovno obravnavo vzhodnoazijskih zbirk in predmetov v različnih slovenskih muzejih (Vampelj Suhadolnik, 2019). Skupna podatkovna baza in portal, ki sta osrednja rezultata projektnega dela, predstavljata neke vrste digitalno različico muzejev vzhodnoazijskih umetnosti in kultur, kakršne najdemo v številnih prestolnicah in velemestih po Evropi, v Severni Ameriki in drugod. Kot pobudnik in vodilni partner projekta si Oddelek za azijske študije Filozofske fakultete Univerze v Ljubljani prizadeva za trajno hrambo in dopolnjevanje ter posodabljanje baze in portala tudi po zaključku projekta, seveda v meri, ki jo bodo v bodoče dopuščale finančne zmožnosti in delovne obveznosti.[14]

Eden od naših osrednjih ciljev je bil vseskozi, da je baza s fotografijami in podatki javno dostopna in enostavna za uporabo, saj velika večina predmetov že več desetletij ni bila razstavljenih, prav tako pa so bili le redki med njimi deležni temeljitejše analize, saj slovenske muzejske institucije nimajo oseb z ustreznim specializiranim znanjem.[15] V okviru projekta smo obravnavali že omenjeno Skuškovo zbirko iz SEM-a, Zbirko Alme Karlin ter Zbirko predmetov iz Azije in južne Amerike iz Pokrajinskega muzeja Celje, vzhodnoazijske kose v zbirki keramike iz Narodnega muzeja ter album vzhodnoazijskih razglednic mornariškega superiorja Ivana Koršiča, ki ga hrani Pomorski muzej Piran.

Pri oblikovanju podatkovne sheme smo se najprej posvetovali s kustosi obravnavnih zbirk in nekaterimi njihovimi muzejskimi sodelavci. Vse sodelujoče institucije uporabljajo program Galis, ki so ga snovalci razvili v sodelovanju z domačimi in tujimi strokovnimi institucijami, tudi Europeano. Shema podatkov, ki jih je moč beležiti za posamezen predmet je izjemno obširna, vendar pa v praksi kustosi izpolnijo le nekaj osnovnih kategorij, pri čemer na izbor vplivajo tako tipi predmetov, za katere skrbijo, kot tudi njihove povsem individualne navade in ambicije. Tudi tuji strokovnjaki, s katerimi smo sodelovali – tako muzejski kustosi kot akademski raziskovalci specializirani za različne vidike vzhodnoazijske umetnosti, so nam svetovali, naj podatkovno shemo razvijemo glede na predmete, ki jih najdemo v slovenskih muzejskih zbirkah. Ti so v resnici izjemno raznovrstni in obsegajo keramiko in porcelan, kipe, pohištvo, tekstil, pahljače, numizmatiko, fotografije in razglednice, slike in lesoreze, orožje, arhitekturne modele ter različne predmete vsakdanje rabe. Po osnovnem pregledu izbranih zbirk smo si v raziskovalni skupini

[11] https://www.photoconsortium.net/.

[12] Projekt s polnim imenom *Vzhodnoazijske zbirke v Sloveniji: vpetost slovenskega prostora v globalno izmenjavo predmetov in idej z Vzhodno Azijo* (2018–2021) (št. J7-9429), je financirala Javna agencija za raziskovalno dejavnost Republike Slovenije (ARRS). Poleg avtoric prispevka je v projektu PAGODE sodelovala še Nataša Vampelj Suhadolnik.

[13] O poti zbirke od Kitajske do SEM-a pišeta Berdajs (2021) in Motoh (2021), o razlogih za pomanjkljivo obravnavo v muzeju pa Vampelj Suhadolnik (2021).

[14] Pridobitev novega nacionalnega raziskovalnega projekta *Osiroteli predmeti: obravnava vzhodnoazijskih predmetov izven organiziranih zbirateljskih praks v slovenskem prostoru* (2021–2024) (ARRS, št. J6-3133) zagotavlja sredstva za nadaljnje delo in tehnične izboljšave.

[15] V okviru projekta VAZ je analiza potekala pretežno s strani sinologinj, japonologinj in koreanista ob podpori pristojnih kustosinj in kustosa. Poleg tega smo organizirali več delavnic, na katerih so izbrane predmete ali skupine predmetov preučili tudi tuji strokovnjaki in strokovnjakinje.

razdelili tipe predmetov glede na rabo,[16] nato pa je vsak za dodeljeni tip predmetov pregledal na spletu dostopne podatkovne sheme različnih priznanih muzejev in arhivov. Pri tem smo bili seveda omejeni na institucije, ki so že digitalizirale dele svojih zbirk in jih ponudile na ogled javnosti, in na tiste vrste podatkov, ki so jih smatrale kot relevantne za obiskovalce in jih zato prikazovale na svojih straneh.

Po več krogih posvetovanj ter širjenj in oženj nabora podatkovnih elementov, smo izoblikovali spodnjo shemo, pri čemer smo metapodatke razdelili na tiste, ki bodo vidni obiskovalcem portala, in one, ki jih zbiramo za naše raziskovalne analize in administracijo. Podatki, jih na portalu ne prikazujemo, so zapisani ležeče. Z asteriskom so označeni podatki, ki jih na portalu uporabljamo kot filtre za prikazovanje.

Trenutno ima naša podatkovna baza obliko Excelove tabele z ločenimi listi za tipe predmetov, vendar je zaradi razvejanosti neprijazna za vnos in slabo pregledna. Smo v postopku tehnične prenove baze, tako da bo v bodoče vstopna točka zanjo spletna stran, vmesnik pa bo v obliki podatkovne kartice. Ob tem bomo dodali nekaj novih kategorij administrativnih podatkov, npr. avtorje zapisa o predmetu.

Administrativni podatki • *Zaporedna številka* (inventarna številka predmeta v naši bazi, označena s črkami za tip in zaporedno številko vnosa) • *Fotografija* (imena fotodatotek, ki prikazujejo predmet) • Copyright • *Podatki o procesu vnosa* (beležimo ali je določen vnos zaključen, lektoriran in prenesen na portal) **Lokacija predmeta** • Zbirka/album* • Muzej* **Provenienca in obravnava predmeta** • Trenutni lastnik • Čas pridobitve • Način pridobitve • Pretekli lastniki in obdobja lastništva • Stanje predmeta, obravnava, poškodbe • Zgodovina razstavljanja • Objave medijih • *Originalne inventarne številke* **Izvor predmeta** • Stoletje* • Obdobje* (dinastična obdobja) • Regija* • Kraj izdelave • Avtor (izvirnik in transkripcija) • Delavnica/tovarna/izdajatelj (izvirnik in transkripcija) • *Datacija (cesarji)*	**Opis predmeta** • Ime predmeta • Raba* • Sekundarna raba* • Material* • Sekundarni material* • Tekstualni opis • Opis materiala • Tehnika izdelave • Dimenzije • Napis – vsebina (izvirnik, transkripcija, prevod) • Podpis(i) (izvirnik, transkripcija) • Cenzor (izvirnik, transkripcija, prevod) • Žig (izvirnik, transkripcija, prevod) • Datum in kraj korespondence • Naslovnik • Pošiljatelj • Število delov/kosov • Ime v izvornem jeziku (izvirnik, transkripcija) • *Motiv* • *Format* • *Tehnika vezave* • *Stil kaligrafije* • *Lokacija napisa na predmetu* • *Lokacija podpisa na predmetu* • *Lokacija cenzorjevega podpisa na predmetu* • *Lokacija žiga na predmetu*

Tabela 1: Podatkovna shema VAZ.

Pri oblikovanju podatkovne sheme VAZ sta nas torej vodili dve vprašanji: katere vrste podatki so ali utegnejo postati koristni za naše raziskave in katere vrste podatkov so lahko zanimive za druge uporabnike, naj bodo strokovnjakinje ali ljubitelji. Čeprav smo imeli javnost nenehno v mislih, pa vse do konca projekta PAGODE nismo veliko razmišljali o načinih priprave kuriranih vsebin, še manj pa o pomenu metapodatkov v tem procesu. Prav tako tudi nismo razmišljali o tem, kako naj bodo podatki strukturirani, da jih bomo čim lažje in čim učinkoviteje raziskovalno obdelovali. Drugače povedano, čeprav je bila digitalizacija eden od osrednjih ciljev projekta VAZ, nismo poznali praks digitalne humanistike. Preveč osredotočeni na predmete kot muzejske predmete po eni ter njihov vzhodnoazijski izvor na drugi strani, nismo našli poti do tistih institucij in strokovnjakov in

[16] Glede na pogostnost v zbirkah smo naredili naslednjo tipologijo po rabi (po abecednem vrstem redu): arhitektura in modeli, glasbila in gledališki predmeti, igre in igrače, kipi, knjige in tiskani materiali, numizmatika, oblačila, obutev in dodatki, orožje in vojaška oprema, pahljače, pohištvo in notranja oprema, posodje in pribor, predmeti za osebno nego, pripomočki za kajenje in uživanje substanc, razglednice in fotografije, religijski predmeti, slike in grafike ter drugo.

strokovnjakinj v našem prostoru, ki se ukvarjajo z digitalizacijo kulturne dediščine, digitalnimi arhivi in digitalno humanistiko. Poleg tega podjetje, ki skrbi za tehnično podporo našega portala, nima izkušenj z razvijanjem podatkovnih baz, smo pa z njimi v preteklosti dobro sodelovali.

5. Prilagoditev podatkovnega modela VAZ za uvoz v Europeano

Vsi digitalni predmeti kulturne dediščine v projektu VAZ, ki smo jih nameravali objaviti v Europeani, so spadali v tip slik, saj je šlo za digitalne slikovne posnetke izbranih predmetov Skuškove zbirke. V projektu PAGODE smo se zavezali doseči višje nivoje na Europeanini lestvici vsebinske kakovosti, s katero označujejo predmete z visokim potencialom za rabo v izobraževanju, na odprtih platformah in v kreativnih industrijah (prim. Europeana 2019a). Fotografije oziroma skenogrami so zato morali izpolnjevati dve zahtevi: (1) njihova velikost ni smela biti manjša od 1200 x 1200 slikovnih točk, in (2) omogočen je moral biti prosti dostop ali uporaba pod licenco Creative Commons Priznanje avtorstva-Deljenje pod enakimi pogoji (CC BY SA).



Slika 2: Prikazovanje podatkov o predmetu na portalu VAZ.

Tudi pri premisleku, katere EDM-ove podatkovne kategorije zapolniti, so nas vodile ambicije po visoki ravni kakovosti metapodatkov, ki jo Europeana ocenjuje glede na večjezičnost, uporabo scenarijev za odkrivanje ter kontekstualne razrede. Ker smo želele nagovoriti širok spekter končnih uporabnikov – od strokovnjakov, do ljubiteljev kulturne dediščine in predstavnikov kreativnega sektorja, smo si za merilo postavile pogoje ravni C. To bi v praksi pomenilo, da bi uporabnik kovance iz Skuškove zbirke lahko našel s splošno poizvedbo »kitajski kovanci« ali zelo detajlno, poznavalsko poizvedbo o konkretnem tipu kovanca »Daoguang tongbao«, v pismenkah »道光通寶«. Za predstavnike kreativnega sektorja so po drugi strani še posebej koristni metapodatki, ki omogočajo iskanje po motivih, vzorcih in barvah.

Ciljna kakovostna raven metapodatkov je terjala, da je vsaj 75 odstotkov podatkov v podatkovnih elementih, ki jih Europeana uvršča med najbolj relevantna za iskanje, moralo imeti tudi metapodatek o jeziku ali jezikih, v katerih je vrednost zapisana. Poleg tega smo morale uporabiti tudi vsaj štiri različne elemente iz dveh različnih scenarijev za odkrivanje ter vsaj dva kontekstualna razreda z ustreznimi povezavami na odprte podatke oziroma nadzorovane besednjake.

Za začetek snovanja podatkovnega modela, smo definirale naslednja izhodišča:

- za osnovo vzamemo izvorno bazo projekta VAZ;
- identificiravao čim večje število podatkov v izvorni bazi, ki jih lahko prevedemo v EDM;
- dodamo administrativne metapodatke, ki jih zahteva EDM;
- v čim večjem obsegu metapodatkom dodamo identifikatorje iz nadzorovanih besednjakov;
- z metapodatki zajamemo tudi raznovrstnost končnih uporabnikov v Europeani.

Naša izvorna podatkovna baza VAZ vsebuje 46 kategorij. Od tega smo jih 23 kot elemente vključile v različne razrede EDM. Med izpuščenimi podatki so bili predvsem tisti, namenjeni tipom predmetom, ki niso bili vključeni PAGODE. V VAZ-u na primer zbiramo podatke, ki so namenjeni raziskovanju razglednic, kot so naslovnik in pošiljatelj. Ker razglednic nismo uvažali v Europeano, smo v pripravi modela za Europeano izločile te kategorije podatkov.

Za naše potrebe smo iz EDM-a uporabile vse tri jedrne razrede, Najprej smo v vsakem od njih identificirale minimalne zahtevane elemente za metapodatke (in si zabeležile njihove standardizirane lastnosti). Ti so (1) oblika digitalnega nadomestka (edm:type), (2) skrbnik podatkov (edm:dataProvider), (3) ime nacionalnega agregatorja ali druga institucija, ki je omogočila pretok podatkov v Europeano (edm:Provider) – v našem primeru je bil to Photoconsortium; in (4) pravice uporabe (edm:rights). Takoj zatem smo v model dodale še kontekstualne razrede za agenta (edm:Agent), časovni razpon (edm:TimeSpan) in koncept (skos:Concept).

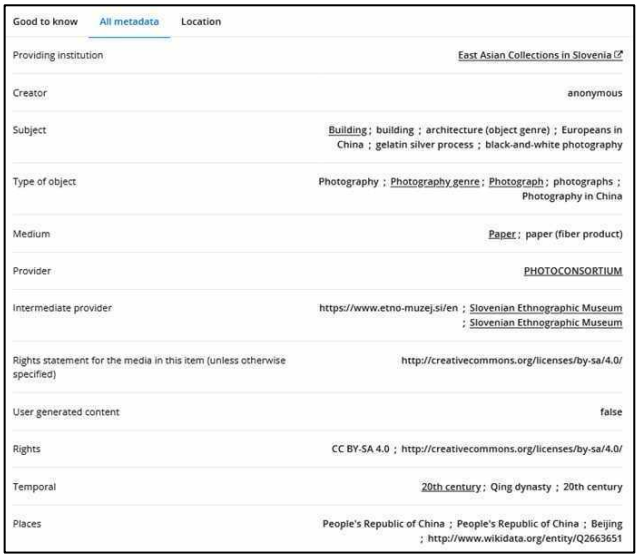
Nato smo v model vključile vse priporočene elemente za metapodatke, ki so sovpadali s posameznimi kategorijami iz baze VAZ, kot so opis predmeta (dc:description) in dimenzije (dcterms:extent). Sledilo je vključevanje priporočenih elementov za metapodatke, ki jih ni v izvorni bazi, a bi omogočali širok spekter uporabnosti. Tu se je zataknilo. Izkazalo se je, da v danem času ne bomo uspele napolniti modela z manjkajočimi podatki, da bi zadovoljile vse identificirane uporabnike Europeane. Čeprav smo prvotno želele vključiti tudi podatke za kreativni sektor, nam je za njemu namenjene elemente (motivi, vzorci, barve) manjkalo največ podatkov, zato smo ta del sheme opustile. So bili pa vnosi s podatki o barvah avtomatsko obogateni v procesu uvoza na Europeano, tako da je danes predmet iz Skuškove zbirke moč iskati in filtrirati tudi po tem kriteriju.

Na koncu smo pri 19 elementih dodale še identifikatorje iz nadzorovanih besednjakov. Dva elementa smo zaradi jezikovne dostopnosti prevedle še v angleščino in sicer ime predmeta (dc:title) in vmesnega ponudnika (edm:intermediateProvider). Podatkovni model za uvoz je na koncu vseboval 67 elementov.

Ko smo imele model zaključen, smo se lotile še pridobivanja manjkajočih metapodatkov. Med njimi so prevladovali identifikatorji. Ta del procesa smo opravile hitro. Zaradi ene naših osrednjih nalog v projektu PAGODE – priprave semantične sheme za avtomatsko in množično ročno anotacijo predmetov kitajske kulturne

dediščine, ki so že bili v Europeani, smo že nekaj mesecev pred tem oblikovale seznam nekaj manj kot 1000 ključnih besed s pripadajočimi identifikatorji iz Gettyjevega AAT in Wikidate.[17] Da smo to dosegle, smo predhodno pregledale vsaj trikrat večje število ključnih besed v omenjenih nadzorovanih besednjakih. Tako smo imele zelo dober pregled, kaj posamezen besednjak ponuja in kaj lahko uporabimo v našem podatkovnem modelu.

Poleg identifikatorjev smo morale v prilagojeno shemo vnesti tudi unikatne identifikatorje fotografij, objavljenih na spletni strani VAZ, saj Europeana digitalizirane predmete prikazuje neposredno s strežnikov institucij ali organizacij. Za konec smo dodale še metapodatke, ki so posamezne dele povezovali v celoten komplet (fotografije v album fotografij, vstavne liste v tiskane oz. slikane albume). Pri tem smo kot vrednosti vnesle unikatne identifikatorje fotografije predmeta objavljenega na spletni strani, ki je bil naslednji po zaporedju (edm:isNextInSequence).



Slika 3: Prikazovanje metapodatkov v Europeani.

Na koncu smo za mapiranje uporabile platformo MINT, ki jo redno uporabljajo projektni partnerji. Platforma omogoča mapiranje metapodatkov in polnjenje Europeane z novo vsebino brez programerskega predznanja o XML podatkovni strukturi, ki tehnično podpira agregacijo. MINT omogoča, da elemente v svojem podatkovnem setu povežeš z elementi EDM. Podatkovni model se uvozi na različne načine, med drugim z datoteko csv, kot smo storile me. Preko uporabniku prijaznega vmesnika se uredi mapiranje, ki ga program nato pretvori v XML obliko, ki omogoči dokončno polnjenje vsebin v Europeano. Poleg mapiranja metapodatkov smo vsakemu elementu metapodatkov v MINT-u ročno določile jezik, v katerem je zapisan, in v kontekstualni razred o agentu (edm:Agent) ročno vnesle imena agentov v različnih jezikih (npr. Marija Skušek/Kondō -Kawase Tsuneko/ 近藤常子).

S celotnim delom smo na koncu dosegle želeno raven C kvalitete metapodatkov, ki zagotavlja brskanje na precizen način, in omogoča, da Europeana deluje kot platforma znanja.

6. Refleksija

Delo na projektu PAGODE – tako priprava semantične sheme kot prilagoditev VAZ-ovega metapodatkovnega modela EDM-u, je bilo za sodelujoče raziskovalke izjemno dragocena izkušnja, skozi katero smo lahko ovrednotile in nato izboljšale tudi delo na projektu VAZ. Kot strokovnjakinje s področja vzhodnoazijskih študij za predmete, ki ji jih digitaliziramo v projektu VAZ, skušamo pridobiti čimbolj izčrpne podatke, ki jih organiziramo v razmeroma razvejano podatkovno shemo. Pri prilagoditvi naše sheme Europeaninemu podatkovnemu modelu, predvsem pa pri polnjenju te sheme s konkretnimi podatki smo zato imele veliko lažjo nalogo kot drugi ponudniki vsebin, ki so sodelovali v projektu PAGODE in niso imeli specializiranih znanj. Ko smo enkrat razumele opredelitve posameznih elementov v EDM-u, smo VAZ-ovim podatkovnim kategorijam hitro našle ustrezne vzporednice, so pa v VAZ-ovi shemi seveda manjkali podatki vezani na spletni vir oziroma agregacijo. Za potrebe projekta PAGODE smo v VAZ-ovo shemo dodale kategorijo avtorskih pravic fotografij, saj se mora ta informacija prikazovati tako na Europeanini kot na izvorni strani.[18] Podatke iz baze VAZ smo v EDM-u obogatile predvsem s povezavami na odprte podatke in nadzorovane besednjake, vendar teh zaenkrat ne nameravamo vključiti v bazo VAZ, saj je naša prioriteta dopolnjevanje baze z novimi vnosi.

Orodje MINT, ki smo ga uporabile za tehnično obdelavo podatkov za uvoz v Europeano, po drugi strani ni zahtevalo programerskih znanj, tako da smo tudi ta del lahko opravile same. Slabost takega načina objavljanja podatkov je, da gre za enkratni uvoz, zato se podatki ne bodo posodabljali hkrati z bazo VAZ. Uporabnik bo z Europeanine strani posameznega predmeta v Skuškovi zbirki sicer preko povezave lahko prispel na VAZ-ovo stran in tam videl najnovejšo verzijo, a podatki v Europeani ne bodo ažurirani, dokler ne bomo izvedli ponovnega uvoza. Če bi to vedele že na začetku, bi gotovo premislile o uvozu preko nacionalnega agregatorja, čeprav bi se glede na časovni pritisk in nizka finančna sredstva na koncu morda vseeno odločile za enostavnejšo agregacijo s pomočjo MINT-a. Poudariti morava, da podatki v Europeani ne bodo napačni ali nekakovostni, bodo le malenkost manj bogati kot v bazi VAZ, ki jo bomo dopolnjevali z novimi raziskovalnimi izsledki.

Skozi sodelovanje v projektu PAGODE smo postale tudi ambicioznejše glede kuriranja in vizualiziranja vsebin na portalu VAZ. Ob premlevanju idej, kako bi naše izsledke predstavili na dostopnejše, privlačnejše načine, sva avtorici ugotovili, da bi bilo boljše, če bi bila naša metapodatkovna shema še bolj razvejana in če bi elemente, ki jih sedaj pišemo skupaj, dodatno razdelili. Pri razglednicah na primer kraj in datum poštnega žiga vnesemo v isto polje, čeprav bi bilo za nadaljnjo obdelavo bolje, če bi jih ločili. Enako je pri provenienci, kjer so sedanji in pretekli lastniki

[17] V Wikidati smo okoli 80 gesel za potrebe projekta PAGODE tudi ustvarile.

[18] Naš prvotni načrt je bil, da bi bile vse fotografije v bazi VAZ prosto dostopne za rabo v nekomercialne namene, vendar so se sodelujoči muzeji skupaj odločili, da želijo ohraniti avtorske pravice. Tudi SEM je pravice spremenil le fotografijam pri predmetih, ki so bili dodani v Europeano.

našteti skupaj. Poleg tega so naju pri EDM-u navdušili kontekstualni razredi, ki bi nam zlasti pri gradivu, kjer imamo veliko podatkov o krajih, osebah in času, olajšali analizo in prikazovanje poti, mrež ter življenjskih zgodb predmetov.

7. Literatura

Tina Berdajs. 2021. Retracing the Footsteps: Analysis of the Skušek Collection. *Asian Studies*, 9(3): 141–166. https://doi.org/10.4312/as.2021.9.3.141-166.

Valentine Charles, Antoine Isaac in Timothy Hill, ur. 2015. *Discovery - User scenarios and their metadata requirements.* https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_WG/DataQualityCommittee/DQC_DiscoveryUserScenarios_v3.pdf

Europeana. 2017. *Europeana Data Model – Mapping Guidelines v2.4.* https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.4_102017.pdf.

Europeana. 2019a. Europeana Publishing Framework: Content. https://pro.europeana.eu/files/Europeana_Professional/Publications/Publishing_Framework/Europeana_publishing_framework_content.pdf.

Europeana. 2019b. Europeana Publishing Framework: Metadata. https://pro.europeana.eu/files/Europeana_Professional/Publications/Publishing_Framework/Europeana_publishing_framework_metadata_v-0-8.pdf

Helena Motoh. 2021. Lived-in museum. *Asian Studies*, 9(3): 119–140.

Nataša Vampelj Suhadolnik. 2021. Between Ethnology and Cultural History: Where to Place East Asian Objects in Slovenian Museums? *Asian Studies*, 9(3); 85–116. https://doi.org/10.4312/as.2021.9.3.85-116

Nataša Vampelj Suhadolnik. 2019. Zbirateljska kultura in vzhodnoazijske zbirke v Sloveniji. V: , uredili Andrej Bekeš, Jana S. Rošker in Zlatko Šabič, ur., *Procesi in odnosi v Vzhodni Aziji: zbornik EARL*, 93–137. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. https://doi.org/10.4312/9789610602699.

Human Evaluation of Machine Translations by Semi-Professionals: Lessons Learnt

Špela Vintar*, Andraž Repar†
* Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana
spela.vintar@ff.uni-lj.si

†Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
andraz.repar@ijs.si

Abstract

We report on two experiments in human evaluation of machine translations, one using the Fluency/Adequacy scoring and the other using error annotation combined with post-editing. In both cases the evaluators were students of translation at the Master's level, who received instructions on how to perform the evaluation but had previously had little or no experience with the evaluation of translation quality. The human evaluation was performed in the context of development and testing different MT models within the Development of Slovene in a Digital Environment (DSDE) project. Our results show that Fluency/Adequacy scoring is more efficient and reliable than error annotation, and a comparison of both methods shows low correlation.

1. Introduction

The design and evolution of a new machine translation system is invariably linked with regular quality assessments, using both automatic methods commonly known as metrics and human evaluations of the MT system's outputs. The context of this experiment is the development of a neural MT system for the English-Slovene language pair within the DSDE project, which involved work packages dedicated to data collection, implementation and testing of different NMT architectures and MT evaluation.

Throughout the project, different versions of the DSDE NMT system were regularly automatically evaluated using the BLEU metric, while later versions were also evaluated with a comprehensive set of scores available on the SloBench 1.0 evaluation platform. In parallel to the automatic ones we performed a set of human evaluations with several aims in mind: To validate the automatic scores with manual assessments, to gain insight into the performance of the system under development, but also to compare two human evaluation scenarios in terms of efficiency and reliability.

The manual evaluations of the DSDE MT engine were performed by students of MA Translation at the Department of Translation Studies, Faculty of Arts, University of Ljubljana. We refer to advanced students of translation as semi-professionals because of their high proficiency in both languages and their understanding of translation as a complex cognitive activity with many alternative solutions for each source text. On the other hand, their experience with translating is for the most part limited to the study environment, and they have received little or no formal training in post-editing or translation assessment.

Manual evaluation was performed using two common evaluation frameworks: the Adequacy/Fluency score and the MQM-DQF error annotation combined with post-editing.

The paper first presents the rationale for selecting the methodologies by referring to related work, then describes the MT system and its development within the DSDE project. We then present the evaluation setups and provide summaries of the results. In addition to quantitative results, for the error annotation and post-editing task we also give a brief summary of the most frequent observations. We conclude by discussing the findings from the perspective of translation quality assessment in MT development.

2. Related work

Evaluation of MT is a crucial part of development and improvement of MT systems, and it is traditionally divided into automatic evaluation using metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), and human or manual evaluation. Automatic evaluation is usually performed by comparing the candidate machine translated text to a reference translation produced by a human professional, whereby the comparison can be rather superficial and word-based such as with BLEU, or more linguistically informed such as with METEOR. The obvious advantage of automatic metrics is that they can be performed on the fly requiring no human effort, but the rate of correlation with human judgements remains a constant concern. Particularly since the emergence of NMT, some authors show that the reliability of metrics as indicators of translation quality may be faltering (Shterionov et al., 2018), or that metrics alone cannot adequately reflect the variety of linguistic issues which may affect quality. Manual evaluation therefore remains an integral part of MT quality evaluation and is annually included into the WMT shared task (Bojar et al., 2016).

Over time, many methods of human MT evaluation have evolved. The Adequacy/Fluency scoring was first adopted by the ARPA MT research program (White et al., 1994) as a standard methodology of scoring translation segments on a discreet 5- or 7-level scale. The adequacy evaluation is performed either by professional translators who are presented with the original and the machine translated segment and make judgments about the degree to which the information from the original can be found in MT output, or by monolingual speakers who are presented with the MT and a reference translation. For the fluency evaluation, no reference translation nor original is provided and the evaluators determine whether the translation

"reads" like good language, sounds natural and adheres to grammatical conventions of the language.

Other manual evaluation methods include task-based evaluation (Doyon et al., 1999), post-editing with targeted human annotation, also known as HTER (Snover et al., 2006), and error analysis using various error typologies. The most comprehensive translation error typology to date is the Multidimensional Quality Metrics (MQM) developed in the QT-Launchpad[1] project. The MQM guidelines provide a fine-grained hierarchy of quality issues and a mechanism for applying them to different evaluation scenarios, however the entire tagset is too complex to be used in a concrete evaluation task. Originating from the needs of the language industry, the TAUS Dynamic Quality Framework (DQF) proposed an error typology which has been harmonized with the MQM model in 2015 and is today integrated into most commercial translation tools (Marheinecke, 2016).

The annotation of translation errors can be a part of Linguistic Quality Assurance (LQA) in professional translation environments, in order to monitor quality on the corporate, project or individual levels. However, for the task of manual MT evaluation MQM and related methods are notoriously poor in inter-annotator agreement scores (Lommel et al., 2014). Some authors believe that pre-annotation training can significantly reduce disagreements, but the task apparently remains highly subjective.

Despite the labour intensity and low inter-annotator agreement, error annotation is still frequently employed in human MT evaluation because of the significance and depth of insight into translation issues it may provide. As Klubička et al. (2017) point out, Slavic languages are rich in inflection, case and gender agreement, and they have rather free word order compared to English. The motivation for using error analysis in MT evaluation is to see – in the process of developing and improving a new MT system – whether the particular grammatical issues occurring with Slavic languages are adequately addressed, resulting in overall quality improvement.

In line with related works we opted for two of the most commonly used manual evaluation methods, the Fluency/Adequacy score and the TAUS DQF-MQM metrics which has been further adapted for the DSDE project.

3. The DSDE MT system

The main goal of the machine translation work package in RSDO is to improve on the state-of-the-art model for the Slovene/English and English/Slovene language pairs developed within the TraMOOC project (Sennrich et al., 2017). To this end, various neural machine translation frameworks were evaluated, such as *MarianNMT* (Junczys-Dowmunt et al., 2018), *fairseq* (Ott et al., 2019) and *NeMo* (Kuchaiev et al., 2019). The same dataset consisting of publicly available parallel data as well as data collected within the DSDE project[2] was used to train the models on the selected frameworks.

4. Evaluation setup

Both types of manual evaulation were performed by students of MA Translation at the Department of Translation, University of Ljubljana. The translation environment of choice was memoQ, a tool which allows the project manager to select or define an LQA scheme with the fluency/adequacy scoring or the error categories respectively. The annotator performs the evaluation, error annotation and post-editing in a typical two-column setting with the segmented original on the left hand side and the machine translated segments already inserted into the target text on the right hand side via pre-translation. Annotators receive an outbound memoQ package which ensures that the source text, the raw MT and the evaluation/error annotation scheme are available and activated with no further setup, and the evaluated, post-edited and annotated texts may be returned to the project manager (in our case the experiment designer) as inbound return packages.

Five different source texts were used from the domains of chemistry, karst, economy, law and general news. The texts were of comparable length (~500 words) and consisted either of the entire text or a meaningful portion thereof. With the exception of the general news text dealing with US elections, all domain-specific texts were highly specialized with complex syntax and many terminological expressions.

For the fluency/adequacy scoring, both language pairs were evaluated by a group of five students over a period of several months. Each document was evaluated by two students. Once a new model was available, MemoQ packages were sent to the students who performed the evaluation in their home environment. Note that for the adequacy/fluency evaluation, no postediting took place – the students only had to score each translated segment on a scale of 0 to 3 (see Table 1).

	Adequacy	Fluency
0	None	Incomprehensible
1	Little	Disfluent
2	Much	Good
3	All	Flawless

Table 1: Adequacy/Fluency scoring.

For the error annotation, only the English-Slovene language pair was evaluated, with English as original and Slovene as target. Fifteen students participated, so that post-editing and error analysis were in the end performed by three students for each text. The experiment took place during a regular face-to-face seminar session in the presence of the lecturer. Students were using standard PCs and with memoQ 9.5 running Translator Pro licenses.

Students were requested to perform full post-editing of the machine translated text, and at the same time annotate each error using the preloaded TAUS/DSDE error typology. The latter proved somewhat wearisome, since the annotation of each single error involves opening a separate dialog box, selecting the category and resuming work, whereby the typical commands used during "normal" translation must be avoided (e.g. Control + Enter to confirm the segment). This invariably slows down the post-editing process and presumably affects the natural cognitive flow during post-editing.

[1] https://www.qt21.eu/launchpad/index.html

[2] Data collected within the DSDE project will be made available under a CC-BY-SA 4.0 license at the end of the project.

5. Results

5.1. Fluency/Adequacy scoring

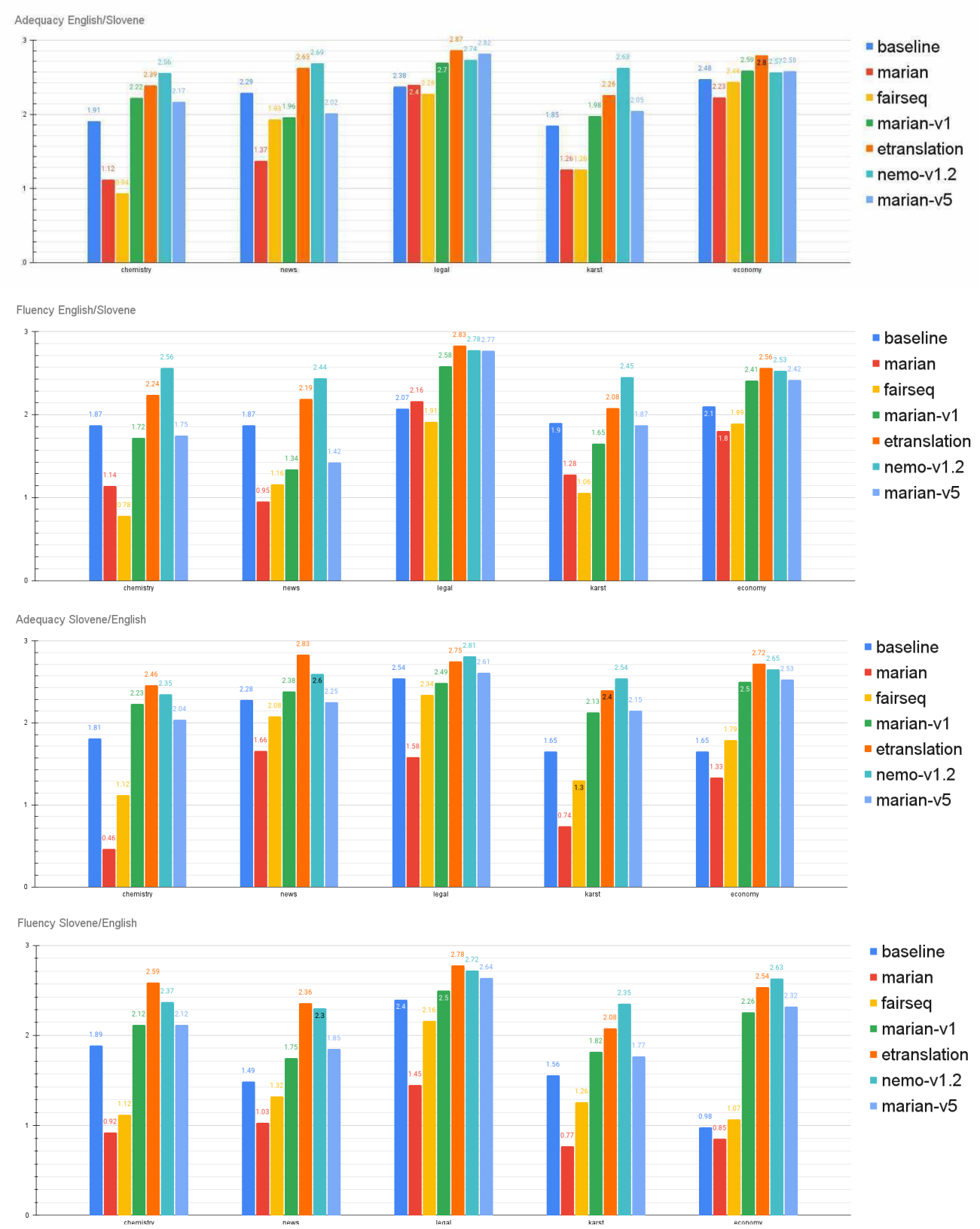


Figure 1. Adequacy and Fluency scores across five domains and two language pairs.

In addition to the baseline model, five models were evaluated using the Adequacy/Fluency methodology (three versions trained using the marianNMT framework, one using the fairseq framework and one using the NeMo framework. We also performed one round of evaluation of the *eTranslation* system developed by the European Commission.

The initial models (*marian* and *fairseq*) performed badly and did not exceed the scores of the baseline model in the DSDE project, but additional iterations performed

better. The overall best performance was exhibited by the NeMo model with best or close to best scores in all five domains. The latest version of the Marian model (*marian-v5*) also performed well in some domains (e.g. Legal) less well in others. When comparing the DSDE models with eTranslation, we can observe that the NeMo model offers competitive performance across all five domains (with the possible exception of the News domain for the Slovene/English language pair).

5.2. Error annotation with post-editing

The error annotation with post-editing was performed in order to gain insight into the translation issues most affecting MT quality, but also to assess the efficiency and reliability of this methodology when used with semi-professional translators. The evaluation took place in November 2021 using the output of the marian-v5 model.

.

Category	Subcategory	Severity 1 - Critical	Severity 2 - Major	Severity 3 - Minor
Accuracy	Category total	**56**	**68**	**37**
	Addition	1	2	3
	Mistranslation	50	63	30
	Omission	5	3	4
Language	Category total	**3**	**26**	**57**
	Grammar	3	18	37
	Spelling	0	8	20
Style	Category total	**13**	**18**	**80**
	Awkward	6	15	55
	Inconsistent	7	3	25
Terminology	Category total	**4**	**16**	**14**
Total		76	128	188

Table 2: Total errors by category.

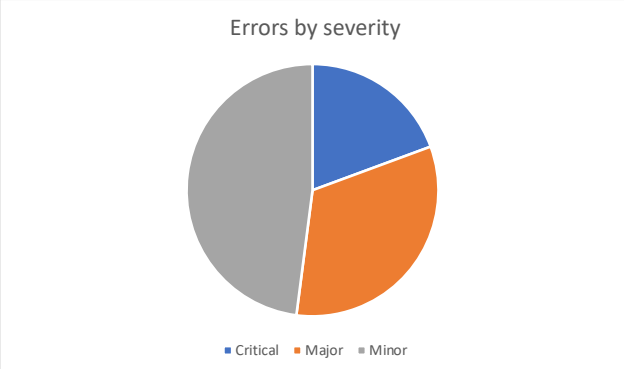


Figure 2: Errors by severity.

As shown in Table 2, the highest number of errors were marked in the Accuracy category, followed by Style, Language and Terminology. Given that four out of five texts were specialized, the low count of terminology errors is perhaps surprising but can be attributed to the fact that annotators frequently choose the Accuracy->Mistranslation category for errors related to specialized lexis. Minor errors are the most frequently selected severity level, with a majority of stylistic errors. Accuracy is also the source of the most critical errors which, in the opinion of annotators, completely change the meaning of the text.

5.2.1. Errors by text

On average, students would annotate ~30 errors per text, or 1.2 errors per segment. The differences in the number of errors between texts are small, with a maximum of 102 errors for the legal text (the sum for all three annotators) and a minimum of 90 for the text on karst.

Category	Subcategory	Chemistry	Economy	Karst	Legal	News
Accuracy	Category total	40	39	58	19	49
	Addition	10	1	0	0	1
	Mistranslation	30	36	56	13	44
	Omission	0	2	2	6	4
Language	Category total	30	14	16	26	18
	Grammar	19	13	15	11	14
	Spelling	11	1	1	15	4
Style	Category total	19	36	13	45	25
	Awkward	19	27	13	22	22
	Inconsistent	0	9	0	23	3
Terminology	Category total	6	5	3	12	9
Total		95	94	90	102	101

Table 3: Errors by text.

Chemistry: There is considerable variation in the number of errors marked by each annotator: 40 / 26 / 29. In all 3 annotations, the most frequent error types are Accuracy and Language, followed by Style and Terminology. Only one annotator found 2 critical errors, the majority of errors were markes as minor.

Economy: The number of errors marked by each annotator varies: 29 / 30 / 35. Similar to other texts, the highest number of errors were attributed to Accuracy->Mistranslation, followed by Style and Language, and only 5 terminology errors.

Karst: The three annotators diverged in the numbers of errors marked: 21 / 31 / 38. Contrary to other texts, here the majority of errors were found to be major or even critical, with only 22 errors categorized as minor. Given that the text was highly specialized, it is again surprising that the Terminology category was not selected more often.

Legal: For the legal text, variation and non-agreement between annotators is at its highest: they marked 21 / 54 / 27 errors each, and even more interesting is the distribution of errors amongst severity levels. For the most prolific annotator, only 4 errors were found to be critical, but for the annotator who spotted 21 errors, 11 were categorized as critical. The third annotator on the other hand found no critical errors.

News: The numbers of errors marked by each annotator were 28 / 33 / 40 respectively, with 12 / 10 / 6 critical errors. Despite the fact that this text was the least specialized of the five, annotators marked 9 errors as terminological, and the overall majority of errors were those pertaining to accuracy (49).

5.2.2. Analysing students' edits

Some texts were highly specialized and rich in terminology, the students however often perceive errors as minor and categorize terminology errors under Accuracy. In the Karst text for example, the original contains the term "precipitation" which is translated as "padavine". None of the annotators identified this as a critical error: in geology, precipitation is not a weather phenomenon but a type of sedimentation process, and the correct translation would read "precipitacija" or "usedanje". The word "test" in the original is most likely a typo and remains untranslated, while the translation of "algal crusts" into "drogovi" is another critical error.

> *In nature, many types of CaCO3 precipitation are linked to living organisms: test, shells, skeletons, stromatolites, algal crusts, etc.*

> *V naravi so številne **padavine** CaCO3 povezane z živimi organizmi: test, oklepi, skeleti, stromatoliti, drogovi itd.*

The students' edits are sometimes unnecessary or even wrong, as in the case of the correctly translated word "adduction" -> "addukcija" corrected into "adukcija" in one case, and in another into "uporaba".

Inconsistent translations are another common issue in machine translation. Thus, in the Economy text, "expenditure" is translated as "stroški", "izdatki", "poraba"; "plant" as "rastlina" and "naprava". A trained and alert post-editor would spot such inconsistencies and make sure they are consolidated in the final version, the students however focus on single segments and overlook such unwanted variation.

Easier to spot are untranslated words, such as "speleothem" in both the Karst original and the Slovene MT. All three annotators spotted the error and opted for "kapnik" in their edits, but the correction is inadequate because "kapnik" is a hyponym of "speleothem" and a better translation would be "speleotem" or "siga". Two annotators marked the error as Critical and one as Major.

It seems that students of translation are much more sensitive to grammatical errors than terminological ones, as the example below containing the correct phrase but in the wrong case was marked as a Major error by all three annotators.

> *Zaradi velike moči odpornosti proti svetlobi in trajnosti derivatov benzimidazolov se pogosto uporabljajo za proizvodnjo **akvarele in elektrofotografskih razvijalnih toner**.*

> *Zaradi velike moči odpornosti proti svetlobi in trajnosti derivatov benzimidazolov se pogosto uporabljajo za proizvodnjo **akvarelnih in elektrofotografskih razvijalnih tonerjev**.*

In many cases the annotators agree on the error itself or the portion of text which should be corrected, but categorize the error differently. A major error was unanimously marked by all three annotators in the Economy text, where the original "To repress these troubles" was machine translated to "Za ponoven tisk te težave". Corrections ranged from "spoprijemanje s težavo", "zmanjšanje teh težav" to "blaženje teh težav", but the error was categorized as Accuracy->Addition, Accuracy ->Mistranslation and Terminology respectively.

Disagreement in categories was frequent also in the non-specialized text, a news article reporting Trump's attempts to postpone elections. The MT version contains a fluent but inaccurate rendering of "November's presidential elections to be postponed", where the MT engine proposed "je predlagal predsedniške volitve v novembru". This is certainly a critical accuracy error, which should be categorized as omission since the postponement was missing in the target. Indeed all three annotators identify the error as critical, but one categorized as mistranslation and the other two as omission. Another severe mistranslation occurs in segment 4, where the MT reverts the meaning of "There is little evidence..." to "Ni malo dokazov..."; again all three annotators agree in the severity level but not in the category.

5.3. Comparing both evaluation methods

While the Fluency/Adequacy evaluation method gives little insight into the specific issues that may have been improved or aggravated from one MT model to another, it seems relatively consistent in the scoring of different models across domains. If we compare the Fluency/Adequacy scores obtained for each text translated by the marian-v5 model with the results of the error annotation, correlation is low. According to the former, the most adequate and fluent translation was that of the legal text, and the least of the karst text. According to the number of annotated errors and edits, karst was the best and legal the worst. (The number of errors in Figure 3 is normalized to allow for better visual comparison.)

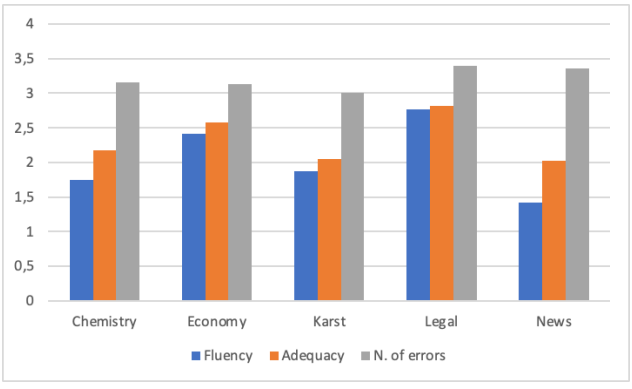


Figure 3: Comparing fluency, adequacy and number of errors per text.

6. Conclusion

We presented the results of human evaluation of MT using two well-known methodologies. The Fluency/Adequacy evaluation is relatively efficient and

fast, and the results are a useful indicator of the quality of different MT models. In general, the scores show high correlation with automatic metrics[3], with Nemo models achieving the highest automatic evaluation scores, followed by the Marian models and the baseline model, which is similar to what can be observed from the Adequacy/Fluency data. To measure the reliability of the Adequacy/Fluency ratings, we calculated the Cohen's kappa coefficient[4] for each document evaluated by a pair of evaluators. As somewhat expected, the agreement is fairly low with most of the values falling between 0.20 and 0.50. The fact that the evaluation was performed by students does not seem to significantly affect the results.

On the other hand, the evaluation through error annotation and post-editing requires a much higher level of effort, linguistic and extra-linguistic competence. Since each text was annotated by three students, a comparison of their decisions provides a valuable insight into the difficulty and subjectivity of the task. Agreement is low for all the parameters under observation: the number of errors marked, their categorization and their severity levels. Moreover, there is little correlation between the number of marked errors, their severity and the true quality of the machine translation. For the text which was the most specialized (Karst), contained a high number of un- or mistranslated terms and received the lowest Fluency/Adequacy score, the number of marked errors was the lowest of all. Student annotators with little or no expert knowledge of the domain will therefore find it difficult to correctly identify terminology errors, assess their severity or post-edit the text to a more accurate version.

Conversely, possibly owing to the fact that students of translation are still in the process of acquiring their language competence and are constantly reminded of the grammatical aspect of the texts they produce, their sensitivity to fluency-related issues is high, hence linguistic and stylistic errors are still often perceived as major. This might explain why the two texts which were most accessible and easy to understand received the highest number of marked errors.

In retrospect, the postediting and error annotation task was too difficult for advanced students of translation and failed to provide meaningful insights into MT quality, for several reasons: Firstly, the texts were too specialized and difficult to understand for non-experts. While students were free to use all available resources, some of the terminological expressions would require extensive research to resolve and the students lacked the time, motivation or skill to perform such research. Secondly, to ensure higher agreement in the severity and category of errors, students should have received training, a test run and much more comprehensive annotation guidelines with English-Slovene examples. Finally, the annotation environment in MemoQ with the rather fine-grained MQM/DSDE error typology is cumbersome and unintuitive, which probably affected the results.

We nevertheless believe that the experiments were valuable both for researchers and annotators. As researchers in MT development and evaluation we have gained experience which will allow us to better design evaluation runs, select texts and train annotators, and the student annotators have been subjected to translation quality assessment and postediting, both of which are tasks frequently encountered in professional translation.

7. Acknowledgments

The project Development of Slovene in a Digital Environment (Slovene: Razvoj slovenščine v digitalnem okolju, RSDO) is co-financed by the Republic of Slovenia and the European Union under the European Regional Development Fund. The operation is carried out under the Operational Programme for the Implementation of the EU Cohesion Policy 2014–2020.

The authors thank the students of MA Translation at the Faculty of Arts, University of Ljubljana, for their participation in the task.

8. References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72. Association for Computational Linguistics.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten years of WMT evaluation campaigns: Lessons learnt. In: *Proceedings of the LREC 2016 Workshop Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–34.

Jennifer Doyon, Kathryn B. Taylor, and John S. White. 1999. Task-based evaluation for machine translation. In: *Proceedings of Machine Translation Summit VII*, pages 574–578.

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics* 108, no. 1 (2017), pages 121–132.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building AI applications using neural modules. arXiv:1909.09577.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció* 12, pages 455–463.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In: *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Katrin Marheinecke. 2016. Can Quality Metrics Become the Drivers of Machine Translation Uptake? An Industry Perspective. *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 71–76.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael

[3] Automatic metric scores can be found at https://slobench.cjvt.si/

[4] Using the *cohen_kappa_score* function from the *sklearn* Python library.

Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. arXiv:1904.01038.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rico Sennrich, Antonio Valerio Miceli Barone, Joss Moorkens, Sheila Castilho, Andy Way, Federico Gaspari, Valia Kordoni, Markus Egg, Maja Popovic, Yota Georgakopoulou, Maria Gialama, Menno van Zaanen. 2017. TraMOOC—translation for massive open online courses: recent developments in machine translation. In: *20th Annual Conference of the European Association for Machine Translation*, EAMT.

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'dowd, and Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation* 32, no. 3, pages 217–235.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In: *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.

Automatic Predicate Sense Disambiguation Using Syntactic and Semantic Features

Branko Žitko,[*] Lucija Bročić,[*] Angelina Gašpar,[†] Ani Grubišić,[*] Daniel Vasić,[‡] Ines Šarić-Grgić[*]

[*]Faculty of Science
University of Split
Ruđera Boškovića 33, 21000 Split, Croatia
branko.zitko@pmfst.hr, lucija.brocic@pmfst.hr, ani.grubisic@pmfst.hr, ines.saric@pmfst.hr

[†]Catholic Faculty of Theology
University of Split
Ulica Zrinsko Frankopanska 19, 21000 Split, Croatia
angelina.gaspar@kbf-st.hr

[‡]Faculty of Science and Education
University of Mostar
Poljička cesta 35, Mostar, Bosnia and Herzegovina
daniel.vasic@fpmoz.sum.ba

Abstract

This paper focuses on Predicate Sense Disambiguation (PSD) based on PropBank guidelines. Different approaches to this task have been proposed, from purely supervised or knowledge-based, to recently hybrid approaches that have shown promising results. We introduce one of the hybrid approaches - a PSD pipeline based on both supervised models and handcrafted rules. To train three supervised POS, DEP and POS DEP models we used syntactic features (lemma, part-of-speech tag, dependency parse) and semantic features (semantic role labels). These features enable per-token classification, which to be applied to unseen words, requires handcrafted rules to make predictions specifically for nouns in light verb constructions, unseen verbs and unseen phrasal verbs. Experiments were done on newly-developed dataset and the results show a token-level accuracy of 96% for the proposed PSD pipeline.

1. Introduction

One of the main tasks of Natural Language Processing (NLP) is precisely understanding the meaning of the word and its specific usage in a sentence, task known as Word Sense Disambiguation (WSD). In this paper, we focus on predicate sense disambiguation, i.e. the correct meaning of a predicate in a given sentence. A predicate combines with a subject to form a sentence, expressing some situation, event or state. Predicates are often single or compound verbs, consisting of various part-of-speech (prepositions, adverbs, nouns, auxiliaries, etc.). Hence, the precise understanding of the meaning of a sentence lies in the correct disambiguation of different types of words, not just verbs. For example, the term light verb (LV) refers to a verb that gets its main semantic content from the noun that follows rather than the verb itself. Thus, the construction consisting of such a verb and noun is called Light Verb Construction (LVC). In the sentence “I take a walk in the park.”, ‘take a walk’ is the LVC in which the noun ’walk’ describes an action. It is non-compositional and its lexical-syntactic structure is not flexible. This example illustrates that word sense disambiguation can make Predicate Sense Disambiguation (PSD) more accurate, since splitting up the LVC and disambiguating the senses of its components individually neglects the semantic unity of the construction and fails to represent its single meaning. Namely, ’walk’ can have a meaning of moving forward, one foot in front of the other, but it can also be a term specific for baseball. Depending on the sense of a word ’walk’, the sense of the whole predicate changes.

Another important role of PSD is the one it plays in Semantic Role Labelling (SRL). The process of semantic role labelling typically consists of predicate identification and its sense disambiguation, followed by identification of semantic roles and finally their labelling. The state-of-the-art BERT models like AllenNLP’s models (Gardner et al., 2018) or InVeRo (Conia et al., 2020) perform all mentioned subtasks except for predicate sense disambiguation which is missing. Ideally, the tool would use predicate senses to label semantic roles. However, we lack the tool for PSD, so we use the opposite technique – attempting to predict role-set IDs from already annotated semantic role labels. Another shortcoming of mentioned state-of-the-art models is that they only label verbs as predicates, and as we have seen, it is necessary to label words of different part-of-speech in addition to verbs. Regarding the sentence "I take a walk in the park.", state-of-the-art models identify word ’take’ as a predicate, whereas they ignore the word ’walk’. The need for such a PSD tool arises during the question generation task in intelligent tutoring system (Grubišić et al., 2020) our research team is working on.

In this work, we describe our PSD pipeline, depicted in Figure 1, as well as the process it takes to create it. The approach we take is the combination of the supervised PSD trained with the Stochastic Gradient Descent method (Kiefer and Wolfowitz, 1952) and the knowledge

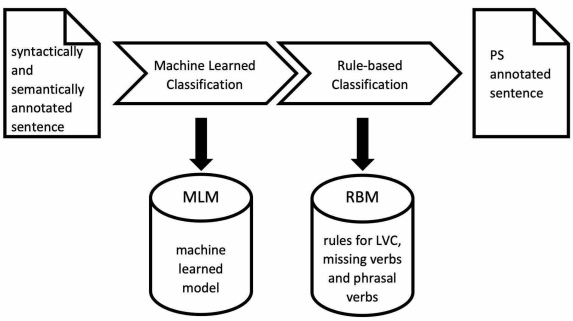


Figure 1: Our PSD Pipeline.

used to handcraft rules to compensate for the shortcomings of the data. We train supervised classifiers for each word to disambiguate senses based on extracted syntactic and semantic features, which play a significant role in many NLP tasks (e.g. text summarization, question generation, etc.). As for the syntactic features we use spaCy (Honnibal et al., 2020) annotated fine-grained POS (part-of-speech) tags and dependency tags. We employ the AllenNLP's BERT-based model (Gardner et al., 2018) to retrieve shallow semantics, represented by SRL labels. Thus, the proposed PSD pipeline consists of Machine Learned Classification (MLC) pipeline, based on Machine Learned Model (MLM), and Rule-Based Classification (RBC) pipeline, based on Rule-Based Model (RBM) including handcrafted rules for LVC, unseen verbs (verbs that don't occur in the OntoNotes dataset used for training the MLMs) and unseen phrasal verbs (phrasal verbs that don't occur in the OntoNotes dataset used for training the MLMs). We provide source code[1] with the spaCy integration of the proposed PSD pipeline.

Section 2 provides related works, which suggest that the WSD, which entails PSD, is a current problem encountered in various popular NLP tasks. Section 3 describes the dataset used for training PSD models and the modifications done to it. Section 4 describes the proposed PSD pipeline, providing detailed information on the training and evaluation of the models. Section 5 provides the conclusion of this paper and discussion about the given work.

2. Related Work

Word Sense Disambiguation and Predicate Sense Disambiguation are appealing NLP tasks for researchers in the field. Thus, they are the subject of many research activities, summarized in the up-to-date survey of recent trends in WSD (Bevilacqua et al., 2021). Among the various approaches to WSD, most popular are knowledge-based approaches, which often implement graph algorithms, and supervised approaches, which lately utilize neural networks.

Supervised WSD formulates the given task as classification task. Hence, it requires precisely labelled training data to learn the relationship between word annotations and senses. In contrast to a single classifier approach (Kawahara and Palmer, 2014), where one classifier is trained to make predictions for every word sense, there is also a per-verb approach (Chen and Eugenio, 2010). We implement the latter technique, where we train each classifier to disambiguate senses of only one word. Purely data-driven WSD is a straightforward approach when dealing with the comprehensive data. However, we find Supervised WSD approach that exploits relations between tokens more appealing. In that approach, some examples of improving the sense prediction might be by using contextual embeddings learned from Neural Language Model (Loureiro and Jorge, 2019), or by utilizing WordNet relations to create try-again mechanism to predict sense for ambiguous words (Wang and Wang, 2020).

On the other hand, knowledge-based WSD often implements various graph algorithms to extract from tokens and sentences their syntactic, semantic or any other features. These features are essential for modelling the Lexical Knowledge Base that algorithms use to predict senses. Although there are some high-scoring methods (Wang and Wang, 2020; Scozzafava et al., 2020) based on this approach, knowledge-based WSD systems still perform worse than supervised ones. However, lately there have been a few promising hybrid approaches that combine supervised and knowledge-based ones, as mentioned in the survey (Bevilacqua et al., 2021). Moreover, their high scores indicate that the hybrid approaches are currently the best solution to WSD (Barba et al., 2021). Besides the research done on WSD, there has also been some work concentrated specifically on Verb Sense Disambiguation (VSD). As verbal multiword expressions are semantically complex lexical items, there have been experiments to inspect the effect of the selection of semantic features in VSD. Research works like ours (Dang and Palmer, 2005; Dligach and Palmer, 2008; Ye and Baldwin, 2006) used SRL annotation, which is a distinctive characteristic of a predicate, to get better sense prediction.

3. Data Manipulation and Analysis

To build a good PSD model combining a supervised PSD approach and handcrafted rules, we need good data for the former and clear guidelines for sense disambiguation for the latter.

3.1. OntoNotes Data

We use an English corpus from the OntoNotes project as the train and test data for the supervised component of the model. The English dataset of the OntoNotes Release 5.0 (Weischedel et al., 2013) consists of 13109 annotated documents organized as .onf files, arranged into seven directories that correspond to files' sources. It is important to train the model on the content of assorted genres and types, therefore, OntoNotes was picked as it has the following seven categories: Broadcast Conversation (transcripts of talk shows from channels such as BBC, CNN and MSBNC), Broadcast News (news data collected from various news sources, such as ABC, NBC, CNN and Voice of America), Magazine (Sinorama Magazine), Newswire (data from sources such as Wall Street Journal newswire), Pivotal Corpus (biblical texts from the Old Testament and the New Testament), Telephone Conversation (conversational speech texts) and Web data (English web texts and

[1] https://github.com/lucijabrocic/PSD-pipeline

web text translated from Arabic and Chinese to English). The syntactic annotation of the sentences in the corpus followed the Penn TreeBank scheme and the predicate-argument structure followed the Proposition Bank (PropBank) annotation (Palmer et al., 2005). The OntoNotes English corpus consists of 143709 annotated sentences, most of which but not all have comprehensive annotation. Namely, some web texts selected to improve sense coverage were just tokenized and not even treebanked. Therefore, the corpus needed some refinement before further usage. The scripts (Bonial et al., 2014) provided by the Proposition Bank project enabled the conversion of original PropBank annotations (found in the OntoNotes project) to the new unified PropBank annotations. The files thus obtained were further modified by custom user-defined methods written for this work. Those methods mostly changed the aesthetics of the files, such as converting SRL annotation to utilize BIO notation and converting tree parses into dependency parse annotation. Finally, after the refinement and modifications, our corpus contains 7212 text files (137811 sentences), which follow the original OntoNotes directory structure based on files' sources.

3.2. The English PropBank

As already mentioned, the used data follows the English PropBank (Palmer et al., 2005) sense disambiguation guidelines. This research aims to predict the sense ID, also known as a frameset or roleset ID, for each word of any complex predicate structure in a sentence.

The English PropBank consists of 7311 .xml files called frame files, specifying the predicate-argument structure. One frame file, or frameset, consists of one predicate lemma or multiple different ones, and contains the information about roleset IDs that disambiguate various meanings of a predicate. Since diverse forms of a predicate can be under the same roleset ID, PropBank aliases can help to distinguish the correct sense from the wrong one. As our work required the English PropBank annotation information, we organized all the information for 10687 rolesets (and 7311 framesets) into easily loadable .json file.

No matter how large, representative, and carefully designed, no corpus can exhibit the same characteristics as a natural language. Having this in mind, we check the coverage of rolesets and framesets in the OntoNotes corpus. The analysis shows that the modified files miss 4922 rolesets and 3104 framesets, i.e. they cover 53.94% of rolesets and 57.54% of framesets that occur in the English PropBank. Even though the frequency of using missing framesets and rolesets might be low, the objective is to include as many framesets and rolesets as possible to increase the overall coverage. To achieve this objective, we add the handcrafted rules, explained more thoroughly in subsection 4.3.

4. The Proposed PSD Pipeline

This section describes the training process of three PSD models (POS, DEP and POS DEP) and their evaluation. We train each model by employing two approaches. In the first approach, we split the dataset into train and test sets, while in the second one, we use entire dataset for training.

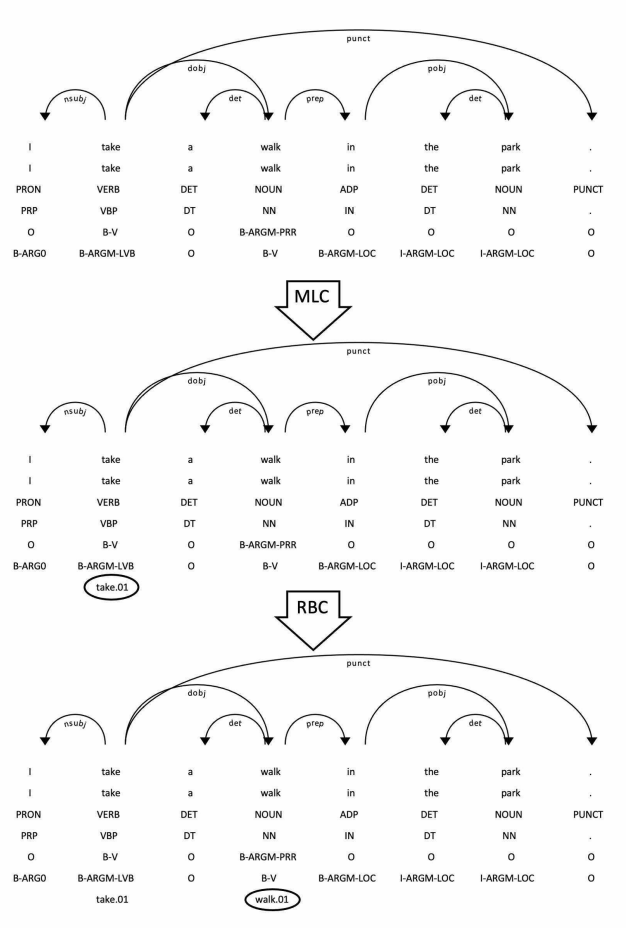


Figure 2: The syntactically and semantically annotated sentence "I take a walk in the park." enters MLC pipeline, which annotates the predicate sense for verb "take" as take.01. The annotated sentence then proceeds to the RBC pipeline, which annotates the predicate sense for noun "walk" as walk.01.

Figure 2 illustrates the PSD pipeline with an example input sentence, annotated with syntactic and semantic features. First the MLC pipeline extracts these features from the sentence and feeds them to the trained classifiers used to obtain predicate senses. Then, RBC pipeline takes the syntactically and semantically annotated sentence with predicted predicate senses. RBC pipeline applies handcrafted rules to the sentence to improve the prediction of predicates in light verb constructions, unseen verbs and unseen phrasal verbs. As a result of the proposed pipeline processing, each token in the sentence has a roleset attribute that stores the result.

4.1. Training the Models

We have 7212 OntoNotes files available to make the best use of while training our models. We first apply a typical supervised learning approach - splitting the dataset into the train and test sets and then performing the training and evaluation. The train-test split given in the PropBank (Bonial et al., 2014) resulted in 80% of the files (and sentences) in train set and 20% in the test set.

Table 1 shows that many framesets and roleset IDs oc-

	No. of files	No. of sentences	No. of framesets	No. of roleset IDs
Train set	5832	111104	3996	5455
Test set	1380	26707	2692	3609
Corpus	7212	137811	4208	5766

Table 1: Corpus composition.

cured in both train and test set. Out of 2692 framesets identified in the test set, 212 framesets did not appear in the train set. Likewise, out of 3609 roleset IDs detected in the test set, 311 of them failed to appear in the train set.

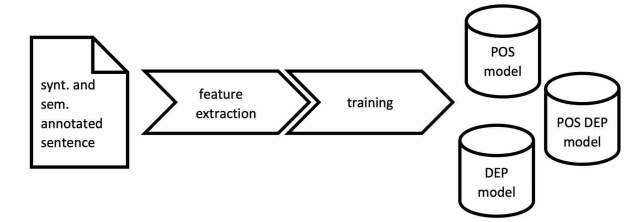


Figure 3: The models' training pipeline.

Figure 3 illustrates the training process. First, the syntactically and semantically annotated sentence is loaded and forwarded to feature extraction.

During the feature engineering and extraction phase, the most relevant token-level annotations for developing the models are selected. Those annotations are token text, its modified lemma that matched the English PropBank frameset, part-of speech (POS) tag, dependency parse and semantic role labels (SRL). The research (Dang and Palmer, 2005; Dligach and Palmer, 2008) shows that the predicate sense disambiguation could improve semantic role labelling. Ideally, word sense disambiguation would solve the problem of identifying the correct sense of a polysemic word based on context. However, the lack of comprehensive repository of senses and a tool for PSD prompted us to use the opposite technique - attempting to predict roleset IDs from already annotated semantic role labels. As for the POS and dependency annotation, previous studies show the performance of the SRL task heavily depends on the performance of dependency parsing (Mohammadshahi and Henderson, 2021) and POS tagging (Wilks and Stevenson, 1997) subtasks. We train three models and name them according to the features they used - POS, DEP and POS DEP. All three models utilize token text and lemma, but differ in the other used annotation(s): (i) the POS model utilizes the relation between SRL and fine-grained POS tag, (ii) the DEP model utilizes the relation between SRL and dependency tag, (iii) the POS DEP model utilizes the relation between SRL, fine-grained POS tag and dependency tag. In this research, we train and evaluate the three models in parallel.

To be more specific, we present featuresets of tokens "take" and "walk" in the Figure 2 used when employing the POS DEP model. Token "take" has only one SRL argument - token "walk" which is ARGM-PRR. On the other hand, token "walk" has three SRL arguments - token "I" that is ARG0, token "take" that is ARGM-LVB, and finally tokens "in", "the" and "park" that are ARGM-LOC. Therefore, the featureset for token "take" is $\big[\big(\text{text, take}\big), \big(\text{lemma, take}\big), \big(\text{ARGM-PRR, } \big[\langle \text{NN, dobj} \rangle\big]\big)\big]$, and for token "walk" $\big[\big(\text{text, walk}\big), \big(\text{lemma, walk}\big), \big(\text{ARG0, } \big[\langle \text{PRP, nsubj} \rangle\big]\big), \big(\text{ARGM-LVB, } \big[\langle \text{VBP, ROOT} \rangle\big]\big), \big(\text{ARGM-LOC, } \big[\langle \text{IN, prep} \rangle, \langle \text{DT, det} \rangle, \langle \text{NN, pobj} \rangle\big]\big)\big]$.

Then we vectorize extracted features and feed them into the classifiers. Dealing with PSD, we face a multiclass classification problem with more than 10000 classes. Instead of a single classifier, a common solution to a problem like this is training multiple binary classifiers, one for each class of the original problem. In the NLP-like domains, however, it is more suitable to use multiple classifiers which predict a constricted number of classes (Even-Zohar and Roth, 2001). Therefore, in this research, multiple multiclass classifiers perform the classification task, with one classifier for each frame file. Hence, the number of classifiers auguments to 7311, and each has to learn the nuances between roleset IDs within the same frame file. The model itself is essentially a collection of such classifiers.

Regarding the choice of classifier, we want to build a simple and fast model for this PSD task. Since the context we need is already assigned to a token through context-aware models (spaCy, AllenNLP), with some feature engineering we can utilize generated annotations (lemma, POS, dependency, SRL) as features for our model. Hence, we did not take a neural approach, but we decided on a linear classifier where learning is based on multinominal logistic regression with SGD optimization.

4.2. The evaluation of models' accuracy and performance

We evaluate our models on the OntoNotes test set containing 26707 sentences. Those sentences contain in total 504891 tokens, of which 75621 (or 14.98%) are predicate tokens, and 429270 (or 85.02%) are non-predicate tokens. When looking at the average sentence, it contains 18.90 tokens, of which 2.83 are predicate tokens and 16.07 are non-predicate tokens. We measure the accuracy of the three PSD models on this OntoNotes test set with three different metrics:

- the token-level accuracy (TLA) metric measures the number of (predicate and non-predicate) tokens the model predicted correctly (correct roleset ID or no prediction, depending on whether the token is a part of a predicate or not)
- the sentence-level accuracy (SLA) metric measures the number of sentences the model predicted completely correctly (all the tokens)
- the predicate-level accuracy (PLA) metric measures the number of predicate tokens the model predicted correctly

Besides accuracy, we also use predicate prediction coverage (PPC) metric, which represents the ratio of predicted predicate tokens and total predicate tokens (whether they are correctly predicted or not). When evaluating AllenNLP's BERT model on OntoNotes test set, we can obtain a measure similar to PPC. Looking at the ratio between

predicate tokens in OntoNotes test set for which AllenNLP annotates the SRL arguments and all predicate tokens in OntoNotes test set, we get a result of 88.02%. It is important to note that the remaining 11.98% are nouns for which AllenNLP's BERT model cannot annotate SRL labels. This coverage metric for AllenNLP puts into perspective the PPC measure of our models, given in Table 2.

	TLA (%)	**SLA (%)**	**PLA (%)**	**PPC (%)**
POS	98.50	76.91	90.01	97.49
DEP	98.71	79.74	91.37	97.82
POS DEP	98.73	80.04	91.54	97.97

Table 2: Evaluation of the Models.

Table 2 shows that the results of evaluation metrics on accuracy are similar for the three models, even though POS DEP model is the most accurate and obtained the highest PPC score. As explained in subsection 4.1., models encounter some framesets and roleset IDs in the test set alone. After the initial training and evaluation phase, we further train models on all 7212 modified OntoNotes files, assuming their performance would improve. To distinguish which results correspond to which model, we will use two terms: OntoNotes-split model and OntoNotes-whole model. The term OntoNotes-split model will denote model that is trained on OntoNotes train set and evaluated on OntoNotes test set, while OntoNotes-whole model will denote model that is trained on all of the 7212 OntoNotes files. The results given so far are for OntoNotes-split models.

4.3. PSD Pipeline

Even when trained on all available data, our PSD models cover only 53.94% of rolesets and 57.54% of framesets in the the English PropBank. Therefore, we handcraft rules to improve the predictive abilities of models.

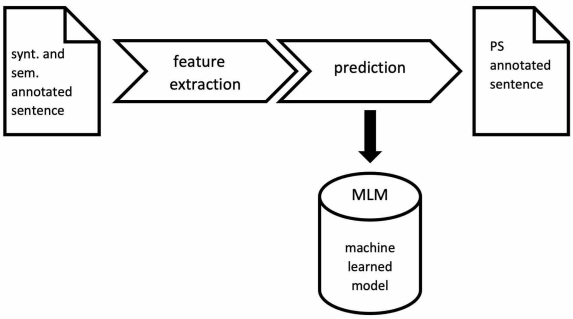


Figure 4: The MLC component of the PSD Pipeline.

Figure 4 presents the Machine Learned Classification (MLC) component of the PSD pipeline, which uses the ML model to make a predicate sense prediction. In model training phase, we use the OntoNotes annotation of sentences for feature extraction. However, when using the PSD pipeline "in the wild" on arbitrary sentences, spaCy's English RoBERTa-based transformer processing pipeline uses the raw input to retrieve syntactic features. The AllenNLP's BERT model is used to obtain semantic features, added to spaCy objects (Token, Span, Doc) via the custom SRL pipe. One thing to note is that we slightly modify both the spaCy pipeline and AllennNLP's BERT model. We improve spaCy's lemmatizer to better lemmatization of gerunds and contracted verbs. The modifications made to the AllenNLP's BERT model allow the presence of nouns in a predicate and adjustment of SRL labels for LVCs to the English PropBank guidelines.

Next, syntactic and semantic features are extracted in the same way as it has been described in the training phase (Subsection 4.1.). The prediction can be done using one of the three previously mentioned OntoNotes-whole models (POS, DEP, POS DEP), and each model is essentially a collection of classifiers that each corresponds to a Penn PropBank frameset. The output of MLC component is a sentence where predicate tokens are annotated with sense predicted via classifiers.

Figure 5 illustrates further processing of annotated sentences in the Rule-Based Classification (RBC) component based on the Rule-Based Model, including handcrafted rules for LVC, unseen verbs and unseen phrasal verbs to improve prediction.

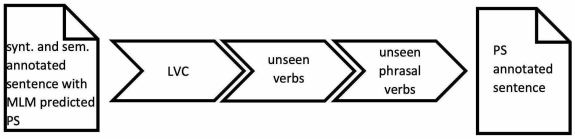


Figure 5: The RBC component of the PSD Pipeline.

Essentially, a sentence with classifier-predicted PSD annotation is forwarded to the RBC pipeline component to first handle the sense disambiguation of nouns in LVCs. Then the RBC component uses modified SRL labels to find both parts of an LVC and search for PropBank aliases to find the corresponding one. The pipeline component explores aliases labeled as nouns only if there are no aliases tagged as the light verbs. This way PropBank aliases help in finding the correct sense IDs.

The next step includes the sense disambiguation of unseen verbs. The RBC component searches for PropBank aliases tagged as verbs, attempting to find the potential sense (roleset ID) of verbs that do not occur in the training set.

In the last step, the pipeline component performs the sense disambiguation of two-word phrasal verbs. Phrasal verbs are easy to predict correctly using the rules. The RBC pipeline first checks if a verb has a dependant particle (eg. a preposition or an adverb) and searches the PropBank aliases tagged as verbs, to find a corresponding sense (roleset ID).

The RBC pipeline makes prediction in each step only if the observed token (i) has SRL labels (AllenNLP model identified the token as predicate), (ii) is not a modal verb (no sense disambiguation of modals) and (iii) has no prediction (the goal is to supplement the classifiers, not to overwrite their predictions).

Moreover, we introduce new annotations in the three steps of the RBC pipeline. For a better understanding, examples in Table 5 illustrate predictions of the PSD pipeline

components using the POS DEP model and their possible outcomes. The search for PropBank aliases can result in a lack of roleset ID matches, only one roleset ID match, or multiple roleset ID matches. Table 5 shows how each pipeline component resolves the roleset ID issue depending on the number of found rolesets matches.

When there is no corresponding roleset ID for the token, the actions of the RBC pipeline differ based on the predicate construction. If the token is a part of an LVC (e.g. picnic - None), the RBC pipeline predicts the sense disambiguation as the lemma of the token followed by ".00" (picnic - picnic.00). If the token is an unseen verb (e.g. overwrite - None) or a part of an unseen phrasal verb (e.g. clue - None), however, the sense remains unchanged (None).

If there is only one roleset ID match, components of the RBC pipeline choose that roleset ID.

If there are multiple roleset ID matches, components of the RBC pipeline choose the roleset ID with the lowest number, followed by the flag "X". However, this annotation indicates that the unique prediction is still not achievable.

Finally, our PSD pipeline incorporates final sense prediction into the spaCy's processing pipeline, into custom roleset attribute.

5. Experimental Results and Discussion

This section provides the results obtained on the gold standard dataset and discussion and suggestions for further work.

5.1. The Evaluation of the Model Performance on the Gold Standard Dataset

As all three OntonoNotes-split models perform similarily well, we further assess the accuracy of the OntoNotes-whole POS DEP model on a fresh set of sentences that represent our gold standard. The new dataset consists of manually annotated 664 sentences with syntactic (lemmatization, part-of-speech and dependency tags) and semantic (SRL) labels, and the predicate sense IDs which our model predicts. In Table 3 are given statistics for the dataset considering tokens, words and predicates. Tokens include both words and non-word parts of a sentence, e.g. punctuation. When expressed as a percentage, 18.46% tokens in the gold dataset are predicates.

	total	**per sentence**			
		mean	**std**	**min**	**max**
token	6853	10.320	4.770	2	65
word	5971	8.992	3.430	1	48
predicate	1265	1.905	0.890	0	12

Table 3: Gold dataset statistics.

The first step of evaluation process includes the predicate sense prediction using input sentence and the needed annotations obtained through system (spaCy transformer model and AllenNLP's BERT model) pipeline. In the second step, as some system annotations are erroneous, namely, wrong lemmatization and SRL labels, we use gold standard annotations to check if there is any difference in prediction.

	TLA (%)		**SLA (%)**		**PLA (%)**		**PPC (%)**
	X	no X	X	no X	X	no X	X & no X
pipeline	96.19	92.20	69.28	69.43	87.11	87.17	98.05
gold standard	97.63	97.67	78.01	78.46	89.75	89.94	100.00

Table 4: Evaluation of the POS DEP model on the gold standard dataset.

The evaluation results in Table 4 show that the OntoNotes-whole POS DEP model predicts better if fed with human-made annotations rather than with system-generated annotations. The most significant difference is in sentence-level accuracy, resulting from higher token-level and predicate-level accuracies.

To put the PPC measure given in Table 4 in perspective, we evaluate AllenNLP's BERT model on the gold standard dataset and obtain a measure similar to PPC. Looking at the ratio between predicate tokens in the dataset for which AllenNLP annotates the SRL arguments and all predicate tokens in the dataset, we get a result of 97.61%. When using system-generated annotations, our OntoNotes-whole POS DEP model relies on AllenNLP for discovering the predicates it needs to predict senses for. By deeper analysis, it is visible that there are certain errors in spaCy's system-generated annotations (namely lemma) that lower the original AllennNLP coverage of 97.61%. However, the modifications made to the AllenNLP's BERT model that allow presence of nouns in a predicate have increased our predicate coverage of 98.05%, and in the end improved the original AllenNLP's coverage of 97.61%.

The POS DEP model returns rolesets with "X" flag when it cannot decide between multiple different senses. To fully evaluate the model's performance, we calculated the four metrics on the predictions with removed "X" flag (no X). The slight increase in scores indicates that the roleset with the lowest ID number was often the right one.

5.2. Discussion and Further Work

We have shown our approach to predicate sense disambiguation utilizing POS, dependency and SRL annotations, and on the way presented the analysis of the coverage of the predicate senses in the OntoNotes corpus and the English PropBank contrastively. The integration of PSD pipeline into spaCy makes its usage straightforward - by adding a custom SRL and roleset components to the spaCy processing pipeline.

Another feature of the proposed PSD pipeline is its Machine Learned Models (MLMs). Each model consists of per-token classifiers, which implies some effort required to combine their outputs. However, the predicate sense prediction is fast since the pipeline only employs the classifiers corresponding to framesets found in the sentence. Moreover, changing the single classifier is simplified – if there is a change in annotation guidelines within one frame file, only one smaller classifier requires retraining. We have also presented different accuracy and prediction metrics used in evaluation of models' performance.

The scores in Table 4 suggest our PSD pipeline ob-

		LVC	Unseen verbs	Unseen phrasal verbs
Roleset ID doesn't exist	**Sentence**	Let's have a picnic in the park.	It will overwrite the files on your hard drive.	She'll clue you in on the latest news.
	MLC prediction	have – have.01 picnic - None	overwrite - None	clue - None
	MLC + RBC prediction	have – have.01 picnic - picnic.00	overwrite - None	clue - None
Unique roleset IDs exist	**Sentence**	He is having an affair.	Some people annotate as they read.	The cat scrunched up to sleep.
	MLC prediction	is – be.03 having - have.01 affair - None	annotate – None read – read.01	scrunched – None
	MLC + RBC prediction	is – be.03 having - have.01 affair – affair.01	annotate – annotate.01 read – read.01	scrunched – scrunch_up.01
Multiple roleset IDs exist	**Sentence**	We are making a plea to all companies.	John frowned when he heard the news.	They sluice the streets down every morning.
	MLC prediction	are – be.03 making – make.01 plea - None	frowned – None heard – hear.01	sluice - None
	MLC + RBC prediction	are – be.03 making – make.01 plea - plead.01X	frowned – frown.01X heard – hear.01	sluice – sluice_down.01X

Table 5: Examples for PSD pipeline.

tains satisfactory results, however, there is still room for improvement. More specifically, in our further work, we plan to enhance the Rule-Based Classification (RBC) component, particularly sense disambiguation of unseen words with multiple rolesets based on their part-of-speech tags. The PSD pipeline only chooses the roleset with the lowest roleset ID and adds the flag "X". We assume we can achieve better results if we create a more complex rule, as the one that utilizes PropBank guidelines on roleset sense IDs and their corresponding arguments in predicate-argument structure. Since there is a large number of missing rolesets and framesets (46.06% and 42.46% respectively), that will be no easy task and more in-depth analysis is necessary to figure out what mistakes does the model make and how to fix them.

We build our Rule-Based Models (RBMs) on three categories of words – nouns in Light Verbs Construction (LVC), unseen verbs and unseen phrasal verbs. Perhaps categories could be further disambiguated and thus, enable a better RBM. Another change that might be beneficial for improving the results is a selection of more features during the feature extraction phase. For a certain predicate, we use only POS and dependency tags of its arguments, but the accuracy might improve if we consider the text of the argument token as well.

Finally, the downstream task this PSD pipeline is created for is the question generation task in our intelligent tutoring system. Disambiguating predicate senses and capturing information about its arguments and characteristics will be useful when deciding on appropriate wh-word in a question.

Acknowledgements

The presented results are the outcome of the research project "Enhancing Adaptive Courseware based on Natural Language Processing (AC&NL Tutor)" undertaken with the support of the United States Office of Naval Research Grant (N00014-20-1-2066).

6. References

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In: Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. 2014. PropBank: Semantics of new

predicate types. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3013–3019, Reykjavik, Iceland. European Language Resources Association (ELRA).

Lin Chen and Barbara Di Eugenio. 2010. A Maximum Entropy Approach To Disambiguating VerbNet Classes. In: *Proceedings of Verb 2010, 2nd Interdisciplinary Workshop on Verbs, The Identification and Representation of Verb Features*.

Simone Conia, Fabrizio Brignone, Davide Zanfardino, and Roberto Navigli. 2020. InVeRo: Making semantic role labeling accessible with intelligible verbs and roles. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 77–84, Online. Association for Computational Linguistics.

Hoa Trang Dang and Martha Palmer. 2005. The role of semantic roles in disambiguating verb senses. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 42–49, Ann Arbor, Michigan. Association for Computational Linguistics.

Dmitriy Dligach and Martha Palmer. 2008. Novel semantic features for verb sense disambiguation. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, page 29–32, USA. Association for Computational Linguistics.

Yair Even-Zohar and Dan Roth. 2001. A sequential model for multi-class classification. In: *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Ani Grubišić, Slavomir Stankov, Branko Žitko, Ines Šarić-Grgić, Angelina Gašpar, Suzana Tomaš, Emil Brajković, and Daniel Vasić. 2020. Declarative Knowledge Extraction in the AC&NL Tutor. In: Robert A. Sottilare and Jessica Schwarz, editors, *Adaptive Instructional Systems*, pages 293–310, Cham. Springer International Publishing.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Boyd Adriane. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Daisuke Kawahara and Martha Palmer. 2014. Single classifier approach for verb sense disambiguation based on generalized features. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4210–4213, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jack Kiefer and Jacob Wolfowitz. 1952. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Alireza Mohammadshahi and James Henderson. 2021. Syntax-aware graph-to-graph transformer for semantic role labelling.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–46, Online. Association for Computational Linguistics.

Ming Wang and Yinglin Wang. 2020. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6229–6240, Online. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0.

Yorick Wilks and Mark Stevenson. 1997. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Natural Language Engineering*, 4.

Patrick Ye and Timothy Baldwin. 2006. Verb sense disambiguation using selectional preferences extracted with a state-of-the-art semantic role labeler. In: *Proceedings of the Australasian Language Technology Workshop 2006*, pages 139–148, Sydney, Australia.

Progress of the RETROGRAM Project: Developing a TEI-like Model for Croatian Grammars Books before Illyrism

Petra Bago

Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences,
University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
pbago@ffzg.hr

1. Background

RETROGRAM[1] (*Retro-digitization and Interpretation of Croatian Grammar Books before Illyrism*) is a 4-year research project that started in November 2019, co-funded by the Croatian Science Foundation (IP-2018-01-3585) and the Institute of Croatian Language and Linguistics. It is a linguistic heritage project that focuses on the digitization and interpretation of pre-Illyrian Croatian grammar books with the aim to serve as a repository of such works in the future as well as to offer a model and develop processes for future similar research on digitization of Croatian grammars. So far, no digitization projects have included Croatian grammar books from the pre-Illyrian period of the Croatian language i.e. before the establishment of the common standard language[2] and orthography (Horvat and Kramarić, 2021).

Croatian language comprises of a common standard language as well as its three dialects: Čakavian, Kajkavian and Štokavian. The standardization of the Croatian literary language and the orthography based on the Štokavian dialect variant began in the 17th century. The process was finalized in the 19th century during the time of Croatian National Revival or the Illyrian movement (i.e. Illyrism). The main goals of the movement regarding language was to introduce a common literary language and a spelling reform, as well as introducing the Štokavian dialect as a linguistic common standard in order to strengthen the national cultural identity. The grammars described in this article thus belong to the pre-Illyrian period of the Croatian language, containing Croatian literary languages that precede the modern Croatian common standard language. The first grammar books were written within the religious orders, of the Jesuits and Franciscans, and were used to teach Croatian or Latin language to the Franciscan and Jesuit youth. (Horvat and Kramarić, 2021)

The main goal of the project is to create a web portal of pre-Illyrian Croatian grammar books, which would include facsimiles of selected grammar books with basic bibliographic and processing information, transcription or translation, and an index of historical grammar and linguistic terminology. The portal will be equipped with thematic searching possibilities on the morphology level. The user will be able to browse grammar books facsimiles, read transcribed or translated text, and search it by predetermined parameters (which will allow conjugation and declension paradigms search). Links to facsimiles will enable comprehensive research on orthography and traductological aspects of the selected texts. An open-access portal will be developed and available to scholars and the general public.

The main objective of the project is to intensify research activities and the interpretation of the Croatian pre-Illyrian grammars within the scope of modern linguistic disciplines (e.g. cognitive approach), to complete existing knowledge about the morphological development of the Croatian language, its normative descriptions, and development of linguistic terminology in the pre-Illyrian period. Conclusions on the formation of the Croatian language grammar model will also be based on the analysis of the Latin language grammar structure. Contrastive analysis of Latin and Croatian grammar meta-text and terminology will lead to conclusions about the influence of Latin language description on Croatian linguistic concepts in the pre-Illyrian period. More on the project can be found in Horvat (2020) and Horvat and Kramarić (2021).

2. Dataset

RETROGRAM has selected eight Croatian grammar books for the digitization and enrichment process that span from the early 17th until the early 19th century. The grammar books cover two dialects (Štokavian and Kajkavian) of pre-Illyrian Croatian before there was an agreement on the common standard language and orthography. Even though not all are grammars of Croatian language, all contain Croatian as metalanguage and/or Croatian examples of morphological paradigms. The texts are transcriptions or translations of the originals in MS Word format, as all have been published as reference books by philologists from the project's research group.

[1] https://retrogram.jezik.hr/

[2] By "common standard language" we mean a standard language covering the entire Croatian speaking area.

The selected transcriptions or translations of grammar books used for the development of the annotation model are based on the following works:

- Bartol Kašić, *Institutionum linguae Illyricae libri duo*, Rome, 1604 (Kašić, 2002),
- Jakov Mikalja, *Gramatika talijanska ukratko ili kratak nauk za naučiti latinski jezik*, Loreto, 1649 (Mikalja, Horvat, and Gabrić-Bagarić, 2008),
- Ardelio Della Bella, *Istruzioni grammaticali della lingua illirica*, Venice, 1728 (Della Bella, Sironić-Bonefačić, and Gabrić-Bagarić, 2006),
- Blaž Tadijanović, *Svašta po malo iliti kratko složenje imena, riči u ilirski i njemački jezik*, Magdeburg, 1761 (Horvat and Ramadanović, 2012),
- Marijan Lanosović, *Uvod u latinsko riči slaganje s nikima nimačkog jezika biližkama za korist slovinskih mladića složen*, Osijek, 1776 (Perić Gavrančić, 2020),
- Ignacije Szentmártony, *Einleintung zur kroatischen Sprachlehre für Deutsche*, Varaždin, 1783 (Szentmártony, 2014),
- Josip Voltić, *Grammatica illirica*, Vienna, 1803 (Voltić, 2016),
- Francesco M. Appendini, *Grammatica della lingua Illirica*, Dubrovnik, 1808 (Appendini and Lovrić Jović, 2022).

3. Data Annotation Model

The eight selected Croatian grammar books are the basis for the development of the annotation model based on the *TEI Guidelines* (TEI Consortium, 2021b). The model addresses two annotation tasks: 1) annotation of historical grammar and linguistic terminology, and 2) the annotation of morphological paradigms. The annotation tasks will be performed manually by experts working on the project. The decision was made to keep the original text intact, and any enrichment to be done through elements and attributes. Each grammar book is a TEI document comprised of a header and the body of the grammar text. The header contains metadata relevant to the project and to the particular grammar book, such as a list of all annotated grammatical terms. The body of the TEI document contains all grammar text with grammatical terminology and morphological paradigms annotations.

3.1. Grammatical Terminology Model

One of the aims of the RETROGRAM project is to facilitate research into historical grammar and linguistic terminology via the web portal. We composed an index of contemporary Croatian terms to be used for normalization of the terminology. These terms are also used in the morphological paradigms annotation task. We have identified 87 terms related to the inflected parts-of-speech. The list of terms is encoded in the TEI header. In the Example 1. we present the encoding of the term "noun"(*imenica* in Croatian) in the index to be used in the annotation model. The example is extracted from Mikalja's grammar book.

```
<encodingDesc>
 <classDecl>
  <taxonomy>
   <category xml:id="imenica">
    <catDesc>imenica</catDesc>
   </category>
    ...
  </taxonomy>
 </classDecl>
<encodingDesc>
```

Example 1: Encoding of the term "noun" (*imenica* in Croatian) in the index of Mikalja's grammar.

To annotate the term in the grammar text, we use the element `<term>`[3] that is, according to the *TEI Guidelines*, used to encode a technical term. In the Example 2 you can find encoding of the historical grammar term *IMENA* that Mikalja used to describe nouns and adjectives, hence two attribute values. The model developed for annotating grammar terminology adheres to the *TEI Guidelines*.

[3] https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-term.html

```
<p>OD <term corresp="#imenica #pridjev">IMENA</term></p>
```

Example 2: Encoding of the term "noun" (imenica in Croatian) in the grammar text of Mikalja's grammar.

3.2. Morphological Paradigms Model

For the development of the morphological paradigms model, we analyzed the following inflected parts-of-speech: nouns, pronouns, adjectives, numbers and verbs. In the *TEI Guidelines*, there is no specific module for encoding grammar texts. However, we have decided to customize the dictionary module (TEI Consortium, 2021a) since it already contains elements that group morphosyntactic information of a lexical item. Interestingly, we were not the only ones with the same idea, as Toma Tasovac and Laurent Romary addressed the issue as part of the TEI Lex-0 initiative[4]. Often the morphological paradigms are presented in a table format. For the purposes of the RETROGRAM project, we decided to disregard the presentation mode of the paradigm, and encode only the implicit information contained in the tables.

To encode one lexical item in a paradigm, we use the element `<form>`[5], which usually "groups all the information on the written and spoken form of one headword" in a dictionary. According to the *TEI Guidelines*, the element is allowed to be contained by elements grouping information on one or more entries. We violate the guidelines by allowing this element to occur in a paragraph. Except for the violation of the guidelines regarding where the element `<form>` can occur, all other child elements adhere to the TEI documentation albeit are not encoding information on a headword, but on a lexical unit of a morphological paradigm. We have defined mandatory and optional information for each inflectional parts-of-speech to be annotated as part of the RETROGRAM project, and developed a customized TEI schema. In Example 3. an encoding of two cases of the noun *vojnik* (soldier in English) as part of the paradigm is presented.

```
<p>Kad ga imenujemo, rečemo
 <form type="inflectedForm" xml:lang="hr">
  <orth>vojnik</orth>
  <gramGrp>
    <gram type="pos" corresp="#imenica"/>
    <gram type="nounType" corresp="#I_opca"/>
    <gram type="gender" corresp="#muski"/>
    <gram type="number" corresp="#jednina"/>
    <gram type="case" corresp="#nominativ"/>
    <gram type="inflectionType" corresp="#I_a_sklonidba"/>
    <gram type="animacy" corresp="#I_zivo"/>
  </gramGrp>
 </form>
 il soldato</p>
<p>Kad se pita čigovo je, rečemo
 <form type="inflectedForm" xml:lang="hr">
  <orth>vojníka</orth>
  <gramGrp>
   <gram type="pos" corresp="#imenica"/>
   <gram type="nounType" corresp="#I_opca"/>
   <gram type="gender" corresp="#muski"/>
   <gram type="number" corresp="#jednina"/>
   <gram type="case" corresp="#genitiv"/>
   <gram type="inflectionType" corresp="#I_a_sklonidba"/>
   <gram type="animacy" corresp="#I_zivo"/>
  </gramGrp>
 </form>
 del soldato</p>
...
```

Example 3: Encoding of two cases of the noun *vojnik* as segment of a morphological paradigm in Mikalja's grammar.

[4] https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Resources/grammars-in-TEI
[5] https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-form.html

4. Future Plans and Conclusion

We are currently conducting the manual annotation tasks based on the two models. Once the annotation tasks are complete, the next step is to create a web portal where all enriched grammar texts will be open and freely available with various search options.

In this extended abstract we present progress of RETROGRAM, a linguistic heritage project that focuses on the digitization and interpretation of pre-Illyrian Croatian grammar books with the aim to serve as a repository of digital Croatian grammars as well as to offer a model and develop processes on digitization of such works. Analyzing eight grammar texts published from the 17th until the 19th century, we developed two models: 1) a model for annotation of historical grammar and linguistic terminology, 2) a model for annotation of morphological paradigms. We composed a taxonomy consisting of 87 terms to be used in both models. To implement the models, we consulted the *TEI Guidelines*, the *de facto* standard in the digital humanities. Our first model adheres to the guidelines. However, our second model is a TEI-like model that we developed based on the dictionary module of the same guidelines. We hope that the morphological paradigm model will serve as a basis for the development of a TEI module for grammars, a model that is presently missing, but could be incorporated in the TEI infrastructure by expanding the dictionary module.

5. Acknowledgements

RETROGRAM is generously co-financed by the Croatian Science Foundation under the program "Research Projects" with grant agreement IP-2018-01-3585 and by the Institute of Croatian Language and Linguistics. We wish to thank all our research associates as well as Toma Tasovac for their feedback and help.

6. References

Francesco Maria Appendini and Ivana Lovrić Jović. 2022. *Appendinijeva Gramatika ilirskoga jezika: Jezična studija s prijevodom i transkripcijom uz faksimil*. Institut za hrvatski jezik i jezikoslovlje, Nacionalna i sveučilišna knjižnica u Zagrebu, Zagreb.

Ardelio Della Bella, Nives Sironić-Bonefačić, and Darija Gabrić-Bagarić. 2006. *Istruzioni grammaticali della lingua illirica, 1728: Gramatičke pouke o ilirskome jeziku*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

TEI Consortium (ed.). 2021a. 9 Dictionaries. In: *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 4.3.0. TEI Consortium. https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html.

TEI Consortium (eds.). 2021b. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.3.0. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Marijana Horvat. 2020. Istraživanje povijesti hrvatskoga jezika u digitalno doba. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46(2):635–643.

Marijana Horvat and Martina Kramarić. 2021. Retro-Digitization of Croatian Pre-Standard Grammars. *Athens Journal of Philology,* 8(4):297–310.

Marijana Horvat and Ermina Ramadanović. 2012. *Jezikoslovni priručnik Blaža Tadijanovića Svašta po malo iliti kratko složenje imena, riči u ilirski i njemačkijezik (1761.)*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Bartol Kašić. 2002. *Institutiones linguae Illyricae/Osnove ilirskoga jezika*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Jakov Mikalja, Marijana Horvat,and Darija Gabrić-Bagarić. 2008. *Gramatika talijanska ukratko ili kratak nauk za naučiti latinski jezik*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Sanja Perić Gavrančić. 2020. *Latinska gramatika i hrvatski jezik Marijana Lanosovića: Povijesnojezična studija i transkripcija izvornika.* Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Ignacije Szentmártony. 2014. *Uvod u nauk o horvatskome jeziku*. Institut za hrvatski jezik i jezikoslovlje, Zagreb.

Josip Voltić. 2016. *Grammatica Illirica/Ilirska gramatika. Reprint of the first edition (1803).* Institut za hrvatski jezik i jezikoslovlje, Zagreb.

The CCRU as an Attempt of Doing Philosophy in a Digital World

Tvrtko Balić

Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000, Zagreb
tvrtko.balic@gmail.com

1. Introduction

The consequences brought about by the Internet have been immense. The resulting chaos effected society at large and while natural sciences could enjoy the greater availability of information, social sciences and humanities found themselves in a new world with new problems. A new environment was created for communities to function in, and this environment was ready to be studied. But it was also an area effected by those sciences, a breeding ground for theories. The fact that theories effect realities they study has been accelerated with the emergence of the Internet. It is not clear how to act in such an environment. In 1995 at Warwick University, England, an experimental cultural theorist collective was formed called the Cybernetic Culture Research Unit (CCRU).

2. Goal of the paper

The goal of the paper is to examine the problems presented by the Internet and to look at the Cybernetic Culture Research Unit as an example of theorists (specifically those in the field of philosophy) adapting to the new medium.

3. Influences

Main influences on the CCRU were French postmodernists and it is itself a postmodern project. Sadie Plant, the feminist lecturer writing a book on "The Situationist International in a Postmodern Age" and Nick Land, the eccentric professor teaching a course on "Current French Philosophy" took their influences and led them to new levels of eccentric.

3.1. Lyotard

Jean-François Lyotard was first to introduce the term "postmodern" in philosophical context. According to him the availability of knowledge is what causes the transition from the modern to a postmodern condition. Organization of knowledge is the thing that serves to justify power in the modern world. As knowledge becomes more available, the power of old actors such as nation-states withers, new actors emerge and the nature of society changes profoundly. Scientific knowledge is not easily accessible and to bring it closer to people for the purpose of legitimation whether of itself or some political, economic, cultural or any other kind of system, it takes the form of a narrative. That is when a problem of a conflict of narratives emerges, but generally one dominates over others and becomes a metanarrative, a story which offers an explanation for the world and justifies a certain social order.

The fundamental feature of postmodernism according to Lyotard is the decay and disappearance of metanarratives. He was enthusiastic about postmodernism and wanted to fragment and break down society in order for experimentation in the social field to yield improvements.

The CCRU presents the Internet as a fertile ground for Lyotard's theories. It should be clear why. It makes knowledge even more accessible as well as the power of expression.

3.2. Derrida

The way in which Jacques Derrida is most reflected in the works of the CCRU's style. What is reflected are his hopes for philosophical writing. He was critical of the seriousness of the philosophical canon which in his day he saw dominated by Hegelian thought.

Derrida celebrates poetry, laughter and ecstasy which he sees as neglected. He sees developed two forms of writing, serious philosophy on the one hand and playful literature on another. He opposes what he calls logocentrism, perceived domination of the ideal of the spoken word and criticism of writing as in literature, something stemming all the way from antiquity with Socrates and Plato sticking out as important critics of writing.

"This-major-writing will be called writing because it *exceeds* the *logos* (of meaning, lordship, presence etc.). Within this writing – the one sought by Bataille – the same concepts, apparently unchanged in themselves, will be subject to a mutation of meaning, or rather will be struck by (even though they are apparently indifferent), the loss of sense toward which they slide, thereby ruining themselves immeasurably." (Derrida, 1990)

Derrida aims to destroy boundaries between philosophy and literature. With the CCRU the boundaries get lost in the creation of a brand-new writing style, theory-fiction. Theory-fiction could be considered a genre of its own, a surreal combination of cyberpunk and Gothic horror. Writings in this style are ambiguous and even their literary meaning is hard to distinguish yet they are filled with philosophical ideas waiting to be deciphered. One could imagine this making Derrida proud or jealous.

3.3. Deleuze and Guattari

As far as theoretical influences are concerned, the French pair that coauthored many works, the philosopher Gilles Deleuze and psychoanalyst and political activist Félix Guattari are probably the most influential. The CCRU writings aim for what Deleuze and Guattari called schizoanalysis. It is an alternative to what is typically understood as rational thinking and more in line with the spirit of the time. It embraces the kind of thinking associated with schizophrenics and people with cluster A personality disorders which are often associated with it. The similarity between a philosopher and a schizophrenic is that they both rely on abstractions and finding connections between wildly different phenomena. Schizoanalysis takes this connection and runs with it before letting it run loose. Thinking becomes chaotic yet orderly in its own way, within its own logic. Everything becomes rhizomatic.

"Let us summarize the principal characteristics of a rhizome: unlike trees or their roots, the rhizome connects any point to any other point, and its traits are not necessarily linked to traits of the same nature; it brings into play very different regimes of signs, and even nonsign states. The rhizome is reducible neither to the One nor the multiple." (Deleuze and Guattari, 2005)

A pair of terms of special note are deterritorialization and reterritorialization, the first one referring to the process by which social relations are altered, mutated or destroyed and the second one referring to the process by which new relations emerge. The CCRU was revolutionary in its accelerationist embrace of social change which meant celebrating deterritorialization, whether for its own sake, motivated by a libertarian desire for freedom, or for the sake of better alternatives emerging, maybe even new trees and new metanarratives.

3.4. Baudrillard

The last very much important figure influencing the CCRU was the French sociologist, philosopher and cultural theorist Jean Baudrillard. The key concepts for him are simulation, simulacra and hyperreality. Simulation is a process by which reality is replaced with its representation and what is left are called simulacra.

Baudrillard describes three orders of simulacra, all stemming from the original traditional symbolic order.

"In the first case, the image is a good appearance - representation is of the sacramental order. In the second, it is an evil appearance - it is of the order of maleficence. In the third, it plays at being an appearance - it is of the order of sorcery. In the fourth, it is no longer of the order of appearances, but of simulation." (Baudrillard 1994)

This fourth case, the third order of simulacra is the pure simulacra, something only ever referencing itself without any authentic reality behind it. This is how Baudrillard conceived of the postmodern world. For him the history of modernity is the history of the disappearance of the real.

However, what is left isn't the unreal or the false, it is the hyperreal. Baudrillard's writing is full of references to magic when speaking of traditional societies and to new technologies, virtual reality, explosions of information, machines conquering humanity etc. when talking about contemporary societies. This is very much the thematically relevant to the CCRU. They weren't the only ones fascinated with Baudrillard, he was so influencial that *The Matrix* is full of references to his work. However, as opposed what is depicted in *The Matrix*, in the hyperreal world there is no real to refer to, there is no exiting the simulation, no escaping the code. But for the CCRU there is hope in the Internet that from the "Desert of the Real" will emerge something new. Baudrillard is pessimistic about changes he observes and only brings up possible sollutions to problems in order to refute them, but in the CCRU there is an amor fati present even if not optimism.

4. Playful and dangerous writing group

From the name, Cybernetic Culture Research Unit and the basic knowledge of what it is about, one might suspect two things, cyberpunk and philosophy. Instead, what one finds is a surreal drug fueled collection of writing about Lovecraftian demons, numerology, ghost lemurs of Madagascar preserving the memories of psychic amphibians… And the things one might expect are so enigmatic as to be distorted beyond recognition. Two things become clear, one is the role of drugs in the CCRU and the other is that it wasn't really a philosophical or information science research group at all, but was primarily a literary club. People involved mostly had a background of philosophy and they had their independent careers, some writing in a more psychedelic style revealing their history in the CCRU and some being more "normal" and understandable. Which isn't to say that there is no philosophy to be found in the CCRU, but a lot of it is motifs and sources of inspiration arising from the chaos of collective storytelling and the authors' common interests and influences.

One important concept related to the CCRU is hyperstition. Hyperstitions are fictions that make themselves real, like how the concept of space travel caused space travel to come into reality. This explains the importance of the artistic style for some members. All ideas can be understood as hyperstition using humans as hosts that bring them into existence. The CCRU often wrong about a fictionalized version of itself. This can be understood as a sort of magic.

5. Prominent figures and their insights

From this literary group emerged strains of thought ranging from far right Nick Land to far left Mark Fisher and cyberfeminist Sadie Plant.

5.1. Sadie Plant and cyberfeminism

Plant offers a unique blend of postmodern feminism and hopes typical for the 90s and visible in films like *Hackers*. According to her, the transformative power of the Internet lies in the fact that it offers a space without physical bodies. Furthermore, computer technology and programming are inherently feminine and therefore benefit women. Finally, women are treated as machines and because of this share a connection with them emancipation of machines will bring about an emancipation of women.

In some respects, Plant proved prophetic. The Internet greatly improved the visibility of marginalized groups and made the general public more compassionate for them. In other respects, not so much, the Internet allows all kinds of opinions to prosper and that certainly includes sexist opinions. But in any case, she certainly offers food for thought about how gender identities are formed and expressed.

5.2. Mark Fisher and blogging

Fisher is most famous for writing about how hard it is to people to imagine an alternative and how capitalism is capable of coopting resistance and creating fake opposition. However, one subject where he was surprisingly optimistic was blogging. Fisher reflected on how doing serious philosophical work (for instance writing a PhD) can be difficult and depressive, but writing a blog is more relaxing, by being less serious it can trick people into doing serious philosophy and it also offers an interactivity that hasn't been seen since the days of the Greek agora. The new digital agoras have since also been assymilated into the existing system. In a way there is a contradiction in Fisher's writing, but the glimmer of hope he saw is important. If it is forgotten, we are not due for any better of a fate than Fisher who killed himself due to depression.

"I started blogging as a way of getting back into writing after the traumatic experience of doing a PhD. PhD work bullies one into the idea that you can't say anything about any subject until you've read every possible authority on it. But blogging seemed a more informal space, without that kind of pressure. Blogging was a way of tricking myself back into doing serious writing. I was able to con myself, thinking, 'it doesn't matter, it's only a blog post, it's not an academic paper'. But now I take the blog rather more seriously than writing academic papers." (Fisher, 2018)

5.3. Nick Land and neo-reaction

For better or worse the member of the CCRU who is most prominent today is Nick Land. One of the ideas which he developed was conceiving of capitalism as an artificial intelligence, but while other authors may hope for this AI to update its software and produce something new, Land seems to be content in accepting that there is no alternative. Land continues to either inspire interpretations of new phenomena on the Internet or offer new interpretations himself. A significant example of the former would be the influence by a combination of younger Land's ideas of hyperstition and older Land's right wing political attitudes in creation of the online theory of meme magick, the idea that Internet memes can influence reality and that this is why Donald Trump won the 2016 US presidential elections in a supernatural way. A significant example of the latter would be Land's philosophy of Bitcoin which isn't only economic, but metaphysical as well, using Bitcoin to explain the logical law of identity and to reaffirm the Kantian understanding of space and time.

6. Concluding remarks

The CCRU is relevant because today Internet is so ingrained in our lives that we don't even notice it any more just as fish don't notice the water they are in. It can prove useful to look at the time when this technology was new and if the future did turn out disappointing maybe we should examine yesterday's speculation of today to remind ourselves what could have been. Sometimes it happens that parts of writing prove to be oddly prophetic and in that case it is good to appreciate what we have or maybe just look at it with new eyes. And even when they seem wrong they represent a valiant attempt at doing something new.

7. References

Brent Adkins. 2015. *Deleuze and Guattari's A Thousand Plateaus: A Critical Introduction and Guide*. Edinburgh University Press, Edinburgh.
Jean Baudrillard. 1994. *Simulacra and simulation*. University of Michigan press, Michigan..
Ccru. 2015. *Ccru: Writings 1997-2003*. Time Spiral Press.
Mark Fisher and Matt Colquhoun. 2020. *Acid Communism*. Pattern Books.
Mark Fisher. 2009. *Capitalist realism: Is there no alternative?*. John Hunt Publishing.
Mark Fisher. 2018. *K-Punk: The Collected and Unpublished Writings of Mark Fisher (2004-2016)*. Repeater.
Gilles Deleuze and Felix Guattari. 2005. *A Thousand Plateaus*. University of Minnesota Press, Minneapolis
Jacques Derrida. *Writing and Difference*. Routledge, London.
Nick Land. 2011. *Fanged noumena: Collected writings 1987-2007*. MIT Press.
Jean-François Lyotard. 2015. *Libidinal economy*. Bloomsbury Publishing, London.
Jean-François Lyotard. 2005. *Postmoderno stanje: Izvještaj o znanju*. Ibis-grafika, Zagreb.
Jean-François Lyotard. 1991. *The inhuman: Reflections on time*. Stanford University Press.
Sadie Plant. 1997. *Zeros and ones: Digital women and the new technoculture*. Fourth Estate, London.

Referencing the Public by Populist and Non-Populist Parties in the Slovene Parliament

Darja Fišer*+, Tjaša Konovšek*, Andrej Pančur*

*Institute of Contemporary History
Privoz 11, SI-1000 Ljubljana
darja.fiser@inz.si
tjasa.konovsek@inz.si
andrej.pancur@inz.si

+Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana

1. Introduction

In the last two decades, political reality in many democratic countries in Europe as well as around the globe has witnessed an increase in active populist political parties and a rise in their popularity among citizens. Parallel to the spread of populism, political science and sociological analyses note a clear difference between the discourses of members of populist and non-populist parties, especially when using social and other media. However, less is known about the relationship between populist and non-populist discourses in the speeches of members of parliament (MPs) in political systems of parliamentary democracy, in which parliaments are the central representative, legislative, and controlling state institutions. This contribution aims at suggesting a model for such analysis. The proposed analysis is embedded around two key concepts. First, we use the concepts of life-world to acknowledge the existence of a specific reality of MPs in which their speech is made. Second, we draw on the existing typology of populist and non-populist parties created by political scientists and sociologists to see how MPs from two different groups of political parties, i.e. populist and non-populist, construct their view of the public. The goal of the analysis is to detect any differences between populist and non-populist discourse observed through the lens of their references to the general public.

2. Approach and methodology

To further investigate the connection between the speech of MPs, their image of the public, and their populist or non-populist origin, we combine cultural history of parliamentarianism with corpus linguistics. From a historical perspective, we draw on recent developments in political history, focusing on the cultural side of the history of parliamentarism (Aerts, 2019; Gjuričová and Zahradníček, 2018; Gašparič, 2012; Schulz and Wirsching, 2012; Ihalainen et al., 2016). For this purpose, we use the concept of life-world (or *Lebenswelt*). The concept of life-world originated in philosophy (Husserl, 1962, Habermas, 2007). The concept of life-world has been used in historiography to emphasize the circumstances in which parliamentarianism is experienced, focusing on MPs as historical actors (Gjuričová et al., 2014). The approach brings to the fore research questions about MPs' perceptions, education, and expectations; their political socialization, prior experiences, and everyday life; and the influence of collective opinions, public images, and the media on their work. In this paper, we focus on one of the aspects of MPs' life-world, namely their relationship to their counterpart, the public, through the words they choose to use, which, in turn, reveals a part of their self-understanding.

In the framework of life-world, we further distinguish between populist and non-populist parties on two axes. First, based on the contents of political parties, we draw on existing research to determine which Slovenian political parties qualify as populist. Second, on the temporal axis, we acknowledge the break of 2004 as a year that witnessed the active beginnings of modern populism in Slovene political space (Fink Hafner, 2019; Frank in Šori, 2015; Fabijan in Ribać, 2021; Campani and Pajnik, 2017; Šori, 2015; Hadalin, 2020; Hadalin, 2021; Lovec, 2019; Pajnik, 2019). We take into account the difference between modern populist parties, as they emerged in the last decade and a half, and their immediate precursors, which have existed since the early 1990s. Therefore, the analysis counts the Slovenian Democratic Party (SDS) and its predecessor, the Social Democratic Party of Slovenia (SDSS), New Slovenia (NSi) and the Slovenian National Party (Slovenska nacionalna stranka, SNS) as populist parties, while all others were classified as non-populist.

3. Analysis

The analysis is based on the *Slovenian parliamentary corpus (1990–2018) siParl 2.0* (Pančur et al., 2020). We take into account the time span from 1992 when the first term of the Slovenian parliament started until 2018 when the seventh term ended. The time frame thus includes some important events that affected the development of Slovenian political parties and their governing style, such as Slovenia's accession to the European Union in 2004 (Gašparič, 2012), the global financial crisis in 2007 and 2008, and the migrant crisis in 2015 (Moffitt, 2014). Using the typology advocated by sociologists and political scientists (see Section 2), we created subcorpora of populist and non-populist political parties for each parliamentary term, resulting in a total of 14 subcorpora. The subcorpora ranged between just under a million tokens in Term1 and to 12 million tokens in Term7 for populist parties, and between 7 million tokens in Term1 and to just under 15 million tokens in Term7 for non-populist parties.

The next step presented a challenge, as there are no pre-existing wordlists of references to the general public that we could rely on. We therefore generated frequency lists of nouns for each subcorpus and manually selected those that refer to the public in the broadest sense (e.g. *person, citizen, inhabitant*) from the 1,000 most frequent nouns in each subcorpus. We only took into account the nouns that can only refer to people (groups or individuals), disregarding those that can be used for institutions (e.g. *association*) or objects (e.g. *school*). We also checked their usage via concordance search and discarded the expressions that could potentially be used for the general public but in this specific corpus predominantly refer to the MPs, the government or their staff (e.g. *proposer*).

As can be seen in Table 1, this yielded a total of 86 unique nouns with the total absolute frequency of 359,320 and relative frequency of 7,322.53 for the populist parties and the total absolute frequency of 524,195 and relative frequency of 6,788.74 for their non-populist counterparts. Most (69) of the nouns are shared between both party groups (e.g. *human*), in addition to 10 that are unique for the populist MPs (e.g. *Croat*) and 7 that are specific to non-populist MPs (e.g. *stakeholder*).

		POPULIST1-7		NON-POPULIST1-7		
	#tokens	49,070,504		77,215,381		
	#lemmas	76		74		
	LEMMA	AF	RF	AF	RF	P:N ratio
P-ONLY	Hrvat	1,341	**27.33**	0	0.00	/
P-ONLY	žena	397	8.09	0	0.00	/
P-ONLY	Avstrijec	318	6.48	0	0.00	/
P-ONLY	diplomant	300	6.11	0	0.00	/
P-ONLY	storilec	232	4.73	0	0.00	/
P-ONLY	volilec	161	3.28	0	0.00	/
P-ONLY	delojemalec	36	0.73	0	0.00	/
P-ONLY	neslovenec	31	0.63	0	0.00	/
P-ONLY	svojec	27	0.55	0	0.00	/
P-ONLY	delavka	0	0.00	0	0.00	/
N-ONLY	deležnik	0	0.00	1,784	**23.10**	/
N-ONLY	prejemnik	0	0.00	1,191	15.42	/
N-ONLY	najemnik	0	0.00	983	12.73	/
N-ONLY	dolžnik	0	0.00	752	9.74	/
N-ONLY	vajenec	0	0.00	444	5.75	/
N-ONLY	kadilec	0	0.00	290	3.76	/
N-ONLY	krajan	0	0.00	172	2.23	/
JOINT	oče	929	18.93	329	4.26	**4.44**
JOINT	obrtnik	1,187	24.19	540	6.99	**3.46**
JOINT	davkoplačevalec	4,762	97.04	2,178	28.21	**3.44**
JOINT	migrant	2,627	53.54	1,255	16.25	**3.29**
JOINT	vlagatelj	426	8.68	260	3.37	2.58
JOINT	podjetnik	3,880	79.07	2,671	34.59	2.29
JOINT	moški	827	16.85	619	8.02	2.10
JOINT	ljudstvo	3,089	62.95	2,376	30.77	2.05
JOINT	Italijan	272	5.54	216	2.80	1.98
JOINT	Slovenka	1,432	29.18	1,143	14.80	1.97
JOINT	pacient	1,619	32.99	1,452	18.80	1.75
JOINT	zamejstvo	1,067	21.74	966	12.51	1.74
JOINT	kmet	6,839	139.37	6,739	87.28	1.60
JOINT	prijatelj	1,024	20.87	1,012	13.11	1.59
JOINT	naročnik	517	10.54	516	6.68	1.58
JOINT	Slovenec	10,103	205.89	11,090	143.62	1.43
JOINT	dijak	2,403	48.97	2,670	34.58	1.42
JOINT	kupec	1,216	24.78	1,357	17.57	1.41
JOINT	državljan	21,570	439.57	24,828	321.54	1.37
JOINT	priča	4,061	82.76	4,701	60.88	1.36
JOINT	državljanka	6,902	140.65	8,372	108.42	1.30
JOINT	narod	4,952	100.92	6,035	78.16	1.29
JOINT	žrtev	3,945	80.39	4,810	62.29	1.29
JOINT	sosed	738	15.04	928	12.02	1.25
JOINT	človek	68,517	**1,396.30**	86,824	**1,124.44**	1.24
JOINT	Rom	627	12.78	808	10.46	1.22
JOINT	bolnik	1,279	26.06	1,717	22.24	1.17
JOINT	prosilec	343	6.99	468	6.06	1.15
JOINT	javnost	16,248	331.12	22,367	289.67	1.14
JOINT	starš	5,732	116.81	7,893	102.22	1.14
JOINT	oseba	16,836	343.10	23,762	307.74	1.11
JOINT	subjekt	3,406	69.41	4,866	63.02	1.10
JOINT	družina	11,120	226.61	16,298	211.07	1.07
JOINT	otrok	18,205	371.00	26,762	346.59	1.07
JOINT	gost	966	19.69	1,438	18.62	1.06
JOINT	begunec	1,247	25.41	1,879	24.33	1.04
JOINT	mladina	1,384	28.20	2,101	27.21	1.04
JOINT	delničar	444	9.05	684	8.86	1.02
JOINT	tujec	3,169	64.58	4,908	63.56	1.02
JOINT	zavarovanec	896	18.26	1,394	18.05	**1.01**
JOINT	volivec	3,478	70.88	5,544	71.80	**0.99**
JOINT	lastnik	8,031	163.66	12,814	165.95	**0.99**
JOINT	mati	320	6.52	512	6.63	0.98
JOINT	družba	23,431	477.50	38,532	499.02	0.96
JOINT	študent	4,973	101.34	8,202	106.22	0.95
JOINT	posameznik	7,367	150.13	12,307	159.39	0.94
JOINT	zavezanec	2,437	49.66	4,096	53.05	0.94
JOINT	uporabnik	3,441	70.12	5,866	75.97	0.92
JOINT	nosilec	2,211	45.06	3,812	49.37	0.91
JOINT	občan	1,558	31.75	2,688	34.81	0.91
JOINT	prebivalec	5,318	108.37	9,404	121.79	0.89
JOINT	partner	4,580	93.34	8,312	107.65	0.87
JOINT	potrošnik	1,657	33.77	3,060	39.63	0.85
JOINT	generacija	2,279	46.44	4,215	54.59	0.85
JOINT	delavec	10,768	219.44	20,055	259.73	0.84
JOINT	invalid	3,032	61.79	5,760	74.60	0.83
JOINT	prebivalstvo	2,727	55.57	5,452	70.61	0.79
JOINT	manjšina	2,742	55.88	5,518	71.46	0.78
JOINT	učenec	1,437	29.28	3,071	39.77	0.74
JOINT	ženska	2,941	59.93	6,517	84.40	0.71
JOINT	upokojenec	3,547	72.28	8,097	104.86	0.69
JOINT	skupnost	16,208	330.30	38,163	494.24	0.67
JOINT	pripadnik	1,375	28.02	3,238	41.93	0.67
JOINT	upravičenec	1,673	34.09	4,523	58.58	0.58
JOINT	upnik	566	11.53	1,725	22.34	0.52
JOINT	podpisnik	465	9.48	1,460	18.91	0.50
JOINT	udeleženec	500	10.19	1,685	21.82	**0.47**
JOINT	porabnik	129	2.63	540	6.99	**0.38**
JOINT	populacija	480	9.78	2,179	28.22	**0.35**
	Total	359,320	7,322.53	524,195	6,788.74	1.08

Table 1: List of specific and joint public-related words identified in the subcorpora of populist and non-populist speeches with their absolute and relative frequencies as well as the usage ratio.

The list of populist-specific nouns contains words describing people according to their background (e.g. *Austrian, non-Slovenian*), family role (e.g., *relative, wife*) and employment status (e.g. *female worker, employee*). Non-populist-specific nouns contain expressions which describe the role or status of a person in an administrative or legal procedure (e.g. *stakeholder, recepient*), business transaction (e.g. *tenant, debtor*), origin (e.g. *local*), education (e.g. *apprentice*) or health status (e.g. *smoker*). Among the joint nouns, *father*, *craftsman*, *taxpayer* and *migrant* are used three times more frequently by populist MPs, whereas *beneficiary*, *participant*, *consumer* and *population* are use more than twice as frequently by non-populist MPs. *Insurance holde*r, *voter* and *owner* are used nearly identically by both groups of MPs. This might reflect a difference between the populist and non-populist parties and their focus in their political base: while the first usually rally voters from rural areas, the latter are traditionally more successful in urban areas.

	T1	T2	T3	T4	T5	T6	T7	Total
Populist #tokens	950,851	4,917,224	7,291,606	8,607,268	8,598,006	6,622,380	12,083,169	49,070,504
Populist "public" AF	6,204	27,738	49,606	68,971	57,041	48,881	100,879	359,320
Populist "public" RF	6,525	5,641	6,803	8,013	6,634	7,381	8,349	7,323
Non-populist #tokens	7,323,569	11,387,486	8,838,299	14,394,700	11,452,223	8,869,712	14,949,392	77,215,381
Non-populist "public" AF	48,446	58,100	52,118	91,254	84,878	67,310	122,089	524,195
Non-populist "public" RF	6,615	5,102	5,897	6,339	7,411	7,589	8,167	6,789
P-value	0.3059	2.54E-43	6.61E-116	0	8.25E-94	2.81E-03	2.01E-07	1.41E-269
Chi2 test	1.0482	190.4453	523.7064	2181.3538	422.1633	21.9444	27.0286	1230.5394
Statistical significance	NO	YES	YES	YES	YES	YES	YES	YES

Table 2: Absolute and relative frequency of public-related words as used by populist and non-populist MPs per parliamentary term and statistical significance tests.

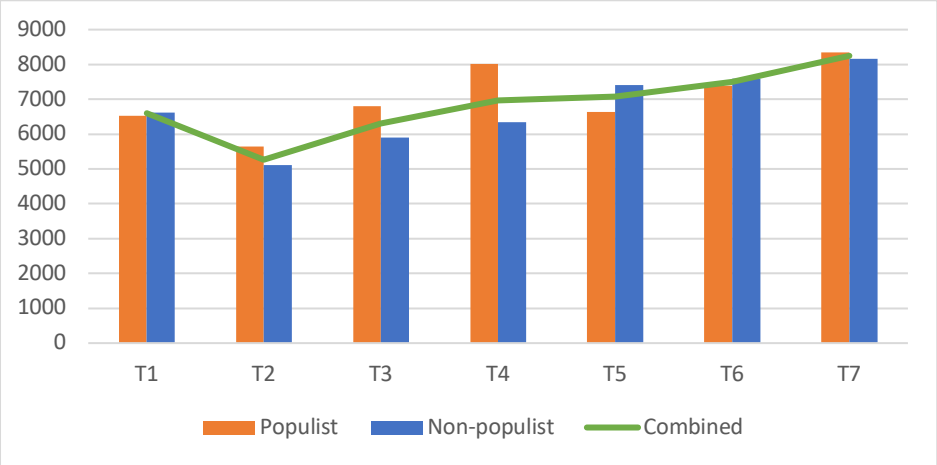


Figure 1: Relative frequency of nouns referring to the public in speeches of MPs from populist and non-populist political parties in the Slovene parliament 1992 – 2018, by parliamentary term.

As can be seen from Table 2 and Figure 1, we observe a steady general upwards trend in the use of nouns, describing the public in both populist and non-populist parties over time. For all terms combined, populist MPs refer to the public statistically significantly more frequently than their non-populist counterparts (P-value 1,41E-269, Chi2 test 1230,5394[1]), which confirms our main hypothesis. For all the MPs combined, the only, and quite substantial, drop in the frequency of references to the public can be observed from Term1 and Term2, which could be contributed to the early stages of the formation of the Slovenian political space. Especially in Term1, the MPs had to face many questions of establishing the working of the new parliament itself. It took time before a new normality of the parliamentary work was established, before the MPs began to address the public more. While early Slovene political transition exhibited a general consensus about the need to strengthen parliamentary democracy, the time after that has been much less clear, which could account to the increase of references of the public by the MPs, since they had to search for new contents of policy-making.

[1] https://www.korpus.cz/calc/

As for individual terms, populist MPs refer to the public statistically significantly more often in Terms2–4 and 7 with Term4 as the biggest outlier, while the opposite is true of Terms5–6 with Term5 as the biggest outlier. In Term1, non-populist MPs use more public-denominating expressions but the difference is not statistically significant. Terms2–3 can be interpreted as the period of formation of populist parties (1992–2004), with Term4 being the first parliamentary term working with a populist (SDS-led) government. In turn, Term7 (2014–2018) could suggest the emergence of the second-wave growing power of populist parties in the face of the crisis of the non-populist parties.

In Terms5–6, when references to the general public prevailed in what sociologists and political scientists refer to as the non-populist discourse, the Slovenian political space witnessed an emergence of numerous new political parties, many of which entered the parliament, which influenced the relation between populist and non-populist discourse. Due to the safe-guards in parliamentary procedures which ensure equal opportunity of participation for opposition MPs regardless of their number, the speeches of MPs might also be influenced by the existence of populist and non-populist led governments and the strength of the populist and non-populist parties in the parliament at the time. While party strength is usually counted by the number of seats taken in the parliament, there are many more factors that influence it and make the correlation between the number of seats, coalition and opposition roles, and party strength challenging (Sartori, 2005; Krašovec, 2000).

4. Discussion

While the results do confirm our initial hypothesis that populist parties refer to the public more, the difference between the two blocs appears to be smaller than the current findings of studies in sociology and political science suggest. Where research from these two fields mainly focuses on the speech of members of populist parties in (selected) television interviews, on social media, and other, less rigid environments, this contribution focused on taking into account all the speeches of MPs throughout the Slovenian parliament which is a highly institutionalized and regulated environment that probably allows for less differentiation between MPs of different political orientation. Our results show that the same life-world of MPs, marked by their shared experience, social forms, norms, and a shared dialogue in plenary sessions provides an environment with a strong unifying factor. Although there is little doubt that political parties themselves decisively differ from one another, the power of the institution, its rigidity and specificity as well as MPs awareness of the target audience and reach of their speeches, proved to be decisive factors in MPs speech when speaking about the public.

According to political scientists and historians, the political space in Slovenia has been increasingly polarized since 1992. Again, our results show a somewhat more nuanced picture: while a growing difference between populist and non-populist discourse can be observed in Terms2–4, the gap narrows in Terms5–7. This challenges the dominant narrative of Slovenian political space. The record high frequency of references to the public by populist MPs in Term4 coincides with SDS winning the 2004 election for the first time after 1992, which happened immediately after the party went through its populist transformation in 2003. Term5, SDS witnessed a backlash with the non-populist coalition prevailing, while one of the populist parties, the NSi, did not even reach the parliamentary threshold.

The general public as well as the media frequently refer to several of the more recent parties, such as Levica, as populist as well. While these parties do exhibit a certain populist appeal, their content, attitudes towards experts and state institutions, as well as their actions in the parliament place them in the non-populist spectrum, with Levica gravitating more towards the spectre of democratic socialism (Toplišek, 2019) than to the same category of populism as defined by Mudde (2005, 2007) which was the theoretical framework of this study. Another methodological issue is temporality: the modern populist shift is a phenomenon belonging to the 21st century; thus, the decade after 1992, included in our analysis, requires a separate interpretation and can only be understood as a preface to the later populist shift (Fuentes, 2020).

5. Acknowledgments

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 No. P6-0281: Political History, the CLARIN ERIC ParlaMint project (https://www.clarin.eu/parlamint) and the DARIAH-SI research infrastructure.

6. Bibliography

Adéla Gjuričová and Tomáš Zahradníček. 2018. *Návrat parlamentu. Česi a Slováci ve Federálním shromáždění*. Argo.

Adéla Gjuričová, Andreas Schulz, Luboš Velek, and Andreas Wirsching, eds. 2014. *Lebenswelten von Abgeordneten in Europa 1860–1990*. Droste Verlag.

Alen Toplišek. 2019. Between populism and socialism: Slovenia's Left party. In: Giorgos Katsambekis and Alexandros Kioupkiolis, eds. *The Populist Radical Left in Europe*. Routledge, Taylor & Francis Group.

Alenka Krašovec. 2000. *Moč v političnih strankah: odnosi med parlamentarnimi in centralnimi deli političnih strank*. Fakulteta za družbene vede.

Ana Frank and Iztok Šori. 2015. Normalizacija rasizma z jezikom demokracije: primer Slovenske demokratske stranke. *Časopis za kritiko znanosti,* 43(260):89–103.

Andreas Schulz and Andreas Wirsching, eds. 2012. *Parlamentarische Kulturen in Europa. Das Parlament als Kommunikationsraum*. Droste Verlag.

Benjamin Moffitt. 2015. How to Perform Crisis: A Model for Understanding the Key Role of Crisis in Contemporary Populism. *Government and Opposition,* 50(2):189–217.

Cas Mudde, ed. 2005. *Racist Extremism in Central and Eastern Europe*. Routledge.

Cas Mudde. 2007. *Populist radical right parties in Europe*. Cambridge University Press.

Danica Fink Hafner. 2019. *Populizem*. Fakulteta za družbene vede, Založba FDV.

Edmund Husserl. 1962. *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie: eine Einleitung und die phänomenologische Philosophie*. M. Nijhoff.

Emanuela Fabijan and Marko Ribać. 2021. Politični in medijski populizem v televizijskem političnem intervjuju. *Social Science Forum*, 37(98):43-68.

Giovanna Campani and Mojca Pajnik. 2017. Populism in historical perspectives. In: Gabriella Lazaridis and Giovanna Campani, eds. *Understanding the populist shift: othering in a Europe in crisis*, pages 13–30. Routledge, Taylor & Francis Group.

Giovanni Sartori. 2005. *Parties and party systems: a framework for analysis*. ECPR.

Iztok Šori. 2015. Za narodov blagor: skrajno desni populizem v diskurzu stranke Nova Slovenija. *Časopis za kritiko znanosti,* 43(260):104–117.

Juan Francisco Fuentes. 2020. Populism. *Contributions to the History of Concepts*, 15(1):47–68.

Jure Gašparič. 2012. *Državni zbor 1992–2012: o slovenskem parlamentarizmu*. Inštitut za novejšo zgodovino.

Jürgen Habermas. 2007. *The Theory of Communicative Action*. *Vol. 2, Lifeworld and system: a critique of functionalist reason.* Polity Press.

Jurij Hadalin. 2020. Straight Talk. The Slovenian National Party's Programme Orientations and Activities. *Contributions to Contemporary History,* 60(2). https://doi.org/10.51663/pnz.60.2.10.

Jurij Hadalin. 2021. What Would Henrik Tuma Say? From The Social Democratic Party of Slovenia to the Slovenian Democratic Party. *Contributions to Contemporary History,* 61(3). https://doi.org/10.51663/pnz.61.3.10.

Marko Lovec, ed. 2019. *Populism and attitudes towards the EU in Central Europe*. Ljubljana: Faculty of Social Sciences.

Mojca Pajnik. 2019. Media Populism on the Example of Right-Wing Political Parties' Communication in Slovenia. *Problems of Post-Communism,* 66(1):21–32.

Andrej Pančur, Tomaž Erjavec, Mihael Ojsteršek, Mojca Šorn, and Neja Blaj Hribar. 2020. *Slovenian parliamentary corpus (1990–2018) siParl 2.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1300.

Pasi Ihalainen, Cornelia Ilie, and Kari Palonen, eds. 2016. *Parliament and Parliamentarism. A Comparative History of a European Concept*. Berghahn.

Remieg Aerts, ed. 2019. *The ideal of parliament in Europe since 1800*. Palgrave Macmillan.

Uporaba postopkov strojnega učenja pri samodejni slovenski grafemsko-fonemski pretvorbi

Janez Križaj*, Simon Dobrišek*, Aleš Mihelič†, Jerneja Žganec Gros†

*Laboratorij za strojno inteligenco, Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška cesta 25, 1000 Ljubljana, Slovenija
janez.krizaj@fe.uni-lj.si, simon.dobrisek@fe.uni-lj.si
†Alpineon razvoj in raziskave, d. o. o., Ulica Iga Grudna 15, 1000 Ljubljana
Tržaška cesta 25, 1000 Ljubljana, Slovenija
jerneja.gros@alpineon.si, ales.mihelic@alpineon.si

1 Uvod

Grafemsko-fonemska pretvorba se nanaša na pretvarjanje izvirno črkovno zapisanih besed danega jezika v njihove fonemske zapise oziroma predstavitve. Nabor osnovnih grafemskih enot, ki se jih razume kot osnovne enote pisave in se jih upošteva pri črkovnih zapisih besed, navadno določa pravopis danega jezika, in enako velja tudi za slovenski jezik (SAZU, 1990). Osnovnim grafemskim enotam pravimo tudi grafemi, njihovim vidno zaznavnim različnim pisnim simbolnim predstavitvam, kot so velike in male črke, pa pravimo alografi. Nabor fonemov je na drugi strani določen predvsem na osnovi glasoslovnega pomensko razločevalnega slušnega kriterija. Grafemi in fonemi so kot osnovne enote do določene mere sicer povezani, a se pri pretvorbi grafemov v foneme lahko tudi več zaporednih črk v zapisani besedi preslika v posamezne foneme. Pretvarjanje grafemskih zapisov besed v njihove fonemske zapise tudi ne temelji samo na nekem manjšem številu osnovnih pravil in pri slovenskem govorjenem jeziku obstaja veliko izjem, ki se ne podrejajo osnovnim pravilom (Toporišič, 2000).

Pri razvoju jezikovnih tehnologij se postopki samodejnega računalniškega pretvarjanja grafemskih zapisov besed v njihove fonemske zapise uporabljajo tako pri izgradnji samodejnih razpoznavalnikov govora kot tudi pri sistemih za tvorjenje umetnega govora (Žganec Gros et al., 2016). V okviru razvojnega in raziskovalnega projekta Razvoj slovenščine v digitalnem okolju (RSDO, 2020) smo izvedli in ovrednotili več različnih uveljavljenih postopkov samodejne grafemsko-fonemske pretvorbe, ki so bili uporabljeni za tovrstno pretvarjanje zapisov slovenskih besed. Preizkusili in ovrednotili smo tri izbrane postopke samodejne grafemsko-fonemske pretvorbe, ki so se uveljavili v zadnjih nekaj letih in so na kratko opisani v nadaljevanju. Za preizkus in ovrednotenje izbranih postopkov smo uporabili množico besed iz slovenskega leksikona Sloleks 2.0 (Dobrovoljc et al., 2019). Množico besed smo na različne načine razdelili na učno in testno množico, ki smo ju nato uporabili za strojno učenje in preizkus izbranih samodejnih grafemsko-fonemskih pretvornikov.

2 Obravnavani postopki

V literaturi je predstavljenih mnogo različnih postopkov za samodejno grafemsko-fonemsko pretvorbo zapisov besed. Starejši postopki praviloma izvajajo pretvorbo na podlagi predhodno definiranih slovničnih pravil (Black et al., 1998). Pomanjkljivost teh postopkov je predvsem v dolgotrajnem ročnem oblikovanju pravil, ki zahtevajo znanje s področja jezikoslovja in glasoslovja in morajo vključevati tudi seznam izjem z različnimi posebnostmi pri izgovorjavah besed. Pri kasneje predlaganih postopkih se je uveljavila pretvorba z modeli skupnih zaporedij (Bisani in Ney, 2008), ki s poravnavo grafemskega zaporedja s fonemskim zaporedjem tvorijo posebne skupne enote, imenovane grafoni. Za modeliranje grafonskih zaporedij nato uporabljajo jezikovne modele n-gramov, udejanjene v obliki uteženega končnega pretvornika (angl. weighted final state transducer), ki omogočajo predvidevanja grafemsko-fonemske pretvorbe za besede, ki niso bile del učne množice.

Avtorji Novak et al. (2015) so razvoj grafemsko-fonemskega pretvornika osnovali na modelih uteženih končnih pretvornikov in predlagali postopek grafemsko-fonemske pretvorbe, ki temelji na prilagojeni metodi maksimizacije upanja za poravnavo niza grafemov z nizom fonemov in več dekodirnih postopkov, med njimi tudi jezikovni model, ki temelji na modelih rekurenčnih nevronskih omrežij (angl. recurrent neural networks).

Yolchuyeva et al. (2019) so dosegli visoko uspešnost grafemsko-fonemske pretvorbe z uporabo globokega modela, ki je poznan pod imenom *transformer*. Ti modeli imajo zgradbo vrste kodirnik-dekodirnik z dodanim mehanizmom pozornosti, ki pomaga pri strojnem učenju soodvisnosti med učnimi pari nizov grafemov in fonemov, kar se odraža tako v hitrejšem strojnem učenju kot tudi pri bolj zanesljivi pretvorbi preizkusnih nizov grafemov v ustrezne nize fonemov.

3 Kvantitativno ovrednotenje

Pri kvantitativnem ovrednotenju obravnavanih postopkov grafemsko-fonemskih pretvorb smo uporabili njihove izvedbe v prosto dostopnih računalniških programskih knjižnicah. Postopek, predlagan v (Bisani in Hermann, 2008), smo udejanjili s programskim orodjem Sequitur[1], postopek avtorjev Novak et al. (2015) je implementiran z orodjem Phonetisaurus[2], za evalvacijo metode avtorjev Yolchuyeva et al. (2019) pa smo uporabili programsko orodje Deep Phonemizer[3].

Pri tvorjenju in preizkušanju vseh obravnavanih modelov in izvajanje postopkov njihovega strojnega učenja smo uporabili ročno validirani del slovenskega leksikona Sloleks 2.0 (Dobrovoljc et al., 2019), ki poleg posameznih besed vsebuje tudi informacijo o njihovih osnovnih besednih oblikah oziroma lemah ter tudi njihove fonemske oziroma fonetične prepise. Validirani del leksikona Sloleks 2.0, ki smo ga uporabili za naše eksperimente, tako vsebuje 646.994 posameznih besed oziroma 62.729 besednih lem. Pri preizkušanju smo opazili, da so rezultati precej odvisni od tega, kako se množico razpoložljivih grafemsko-fonemsko pretvorjenih besed razdeli na učni in testni del. Pri preizkusih smo zato izvedli dve različni razdelitvi množice vseh besed v učno množico, ki je vsebovala 90 % besed iz slovarja, in testno množico, ki je vsebovala preostalih 10 % besed. Pri naključni razdelitvi, v nadaljevanju označeni z oznako "RandomSplit", smo razdelitev izvedli povsem naključno z uporabo sistemskega naključnega generatorja. Pri razdelitvi, ki je temeljila na razvrščanju besed v učno oziroma testno množico glede na njihove leme, pa smo poskrbeli, da se v testni množici ne pojavljajo besede, ki se od besed v učni množici razlikujejo le po končnicah. To namreč pogosto velja za besede z istimi lemami. Polega tega smo poskrbeli, da se leme besed v testni množici razlikujejo za vsaj tri črke glede na njim najbolj podobne leme v učni množici besed. Ta razdelitev je v nadaljevanju označena z oznako "LemmaSplit".

Pri izvajanju poskusov smo ugotovili, da je rezultat po pričakovanjih tudi precej odvisen od upoštevanega nabora fonemskih enot pri grafemsko-fonemskih pretvorbah. Pri gradnji samodejnih razpoznavalnikov govora se tako navadno ne ločuje med dolgimi in kratkimi samoglasniki oziroma med naglašenimi in nenaglašenimi samoglasniki. To ločevanje pri razpoznavalnikih govora namreč ni pomembno po pomensko razločevalnem kriteriju določanja fonemskih enot. To ločevanje pa je pomembno pri gradnji sistemov za tvorjenje umetnega govora, kjer so prozodične značilnosti umetnega govora odvisne od informacije o naglašenih in nenaglašenih samoglasnikih v besedah. V skladu s temi predpostavkami smo učno in testno množico dodatno razdelili na različna načina, glede na to, katere osnovne fonemske enote se je upoštevalo. V nadaljevanju tako oznaka ASR označuje razdelitev, ki je bila primerna za samodejne razpoznavalnike govora in temelji na upoštevanju samo 34 osnovnih fonemskih enot oziroma fonemskih različic. Oznaka TTS pa označuje razdelitev, ki je primerna za sisteme za samodejno tvorjenje umetnega govora in temelji na upoštevanju 39 osnovnih fonemskih enot. Povečanje števila fonemskih enot je posledica upoštevanja ločevanja med dolgimi in kratkimi oziroma naglašenimi in nenaglašenimi samoglasniki. V nadaljevanju predstavljeni rezultati so potrdili predvidevanja, da je pri slovenskem jeziku najteže samodejno napovedovati naglasno mesto v besedah oziroma naglašene samoglasnike. Pri naglaševanju slovenskih besed je namreč zelo veliko izjem, ki se ne podrejajo nekemu bolj splošnemu manjšemu naboru osnovnih pravil naglaševanja besed.

Rezultati uspešnosti samodejnih grafemsko-fonemskih pretvorb so v nadaljevanju podani v obliki odstotnega deleža napačno pretvorjenih besed (angl. word error rate, WER) in deleža napačno pretvorjenih fonemskih enot (angl. phoneme error rate, PER). Kot je razvidno iz tabele so se glede na različne delitve množice besed in upoštevanja ločevanja med naglašenimi in nenaglašenimi samoglasniki pri rezultatih dejansko potrdila predvidevanja. Pri naključni razdelitvi so tako rezultati bistveno boljši kot pri razdelitvi po lemah, saj se pri naključni razdelitvi v testni množici lahko pojavljajo besede, ki se od najbolj podobnih besed v učni množici

[1] https://github.com/sequitur-g2p/sequitur-g2p
[2] https://github.com/AdolfVonKleist/Phonetisaurus
[3] https://github.com/as-ideas/DeepPhonemizer

razlikujejo samo po končnici ali predponi. Rezultati pri večjem naboru osnovnih fonemskih enot, ki vključuje ločevanje med dolgimi in kratkimi samoglasniki (oznaka TTS), pa so prav tako po pričakovanjih precej slabši, kot pri manjšem naboru, ki tega ločevanja ne upošteva (oznaka ASR). To potrjuje druge že obstoječe ugotovitve, da je pri slovenskem jeziku dejansko težko samodejno napovedovati naglasno mesto v besedah (Žganec Gros et al., 2016).

Orodje	Slovar	WER [%]	PER [%]
Sequitur (Bisani in Hermann, 2008)	ASR_RandomSplit	16,5	1,9
	ASR_LemmaSplit	25,4	2,9
	TTS_RandomSplit	17,3	2,2
	TTS_LemmaSplit	50,2	7,4
Phonetisaurus (Novak et al., 2015)	ASR_RandomSplit	1,0	0,1
	ASR_LemmaSplit	14,1	1,6
	TTS_RandomSplit	2,0	0,3
	TTS_LemmaSplit	29,1	4,1
Deep Phonemizer (Yolchuyeva et al., 2019)	ASR_RandomSplit	1,1	0,1
	ASR_LemmaSplit	8,6	0,9
	TTS_RandomSplit	1,7	0,3
	TTS_LemmaSplit	16,1	2,6

Tabela 1: Uspešnost grafemsko-fonemske pretvorbe obravnavanih postopkov.

4 Zaključek

V prispevku so predstavljeni rezultati izvedb in preizkusov različnih samodejnih grafemsko-fonemskih pretvornikov za slovenski jezik. Glede na ugotovitve lahko uporabniki tovrstnih pretvornikov za izgradnjo samodejnih razpoznavalnikov govora pričakujejo približno 91% pravilno pretvorbo besed, ki niso vključene v obstoječe slovenske leksikone. Pri izgradnji sistemov za tvorjenje umetnega govora, pri katerih je pomembno pravilno določanje naglasnega mesta, pa lahko pričakujejo samo približno 84% pravilno pretvorbo.

Zahvala

Predstavljeno delo je bilo delno financirano s strani Ministrstva za kulturo in Evropskega sklada za regionalni razvoj v okviru projekta RSDO (Razvoj slovenščine v digitalnem okolju), s strani Javne agencije za raziskovalno dejavnost Republike Slovenije v okviru aplikativnega raziskovalnega projekta L7-9406 OptiLEX in s strani ARRS v okviru raziskovalnega programa Metrologija in biometrični sistemi (P2-0250).

Literatura

Maximilian Bisani in Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication,* 50(5):434–451.

Alan W. Black, Kevin Lenzo in Vincent Pagel. 1998. Issues in Building General Letter to Sound Rules. V: *Zbornik 3rd ESCA Workshop on Speech Synthesis*, str. 77–80.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Jaka Čibej, Luka Krsnik in Marko Robnik-Šikonja. 2019. Morphological lexicon Sloleks 2.0. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1230.

Josef R. Novak, Nobuaki Minematsu in Keikichi Hirose. 2015. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, 22(6):907–938.

RSDO - Razvoj slovenščine v digitalnem okolju. 2020. https://www.slovenscina.eu/.

SAZU - Slovenska akademija znanosti in umetnosti. 1990. Slovenski pravopis 1: Pravila. Državna založba Slovenije, Ljubljana.

Jože Toporišič. 2000. *Slovenska slovnica*. Založba Obzorja, Maribor.

Sevinj Yolchuyeva, Géza Németh in Bálint Gyires-Tóth. 2019. Transformer Based Grapheme-to-Phoneme Conversion. V: *Zbornik konf. Interspeech 2019*, str. 2095–2099, Gradec, Avstrija.

Jerneja Žganec Gros, Boštjan Vesnicer, Simon Rozman, Peter Holozanin Tomaž Šef. 2016. Sintetizator govora za slovenščino eBralec. V: *Zbornik konf. Jezikovne tehnologije in digitalna humanistika*, str. 180–185, Ljubljana, Slovenija.

Poravnava zvočnih posnetkov s transkripcijami narečnega govora in petja

Matija Marolt, Mark Žakelj, Alenka Kavčič, Matevž Pesek

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
matija.marolt@fri.uni-lj.si

1 Uvod

V povzetku predstavljamo sistem za poravnavo zvočnih posnetkov slovenskega govora s pripadajočimi transkripcijami na nivoju besed. Pri razvoju sistema nas je še posebej zanimala njegova uporabnost pri poravnavi narečnega govora in petja, saj avtomatska razpoznava govora v tovrstnih posnetkih deluje nezanesljivo, z veliko napakami. Natančna avtomatska poravnava posnetkov in transkripcij nam tako lahko pomaga pri analizi narečnih korpusov in pripravi novih anotiranih podatkov za učenje razpoznavalnikov. V povzetku predstavimo sistem za poravnavo in primerjamo kvaliteto poravnave nenarečnih in narečnih govorcev. Analiziramo tudi kvaliteto poravnave narečnega petja z uporabo sistema, ki je učen zgolj na govoru. Ker se petje lahko zelo razlikuje od govora (dodatna spremljava, večglasno petje, dolgi toni, ...), se v nalogi osredotočimo zgolj na enoglasno petje brez spremljave, ki je še najbolj podobno govoru.

2 Sistem za poravnavo

Sistem za poravnavo posnetkov in transkripcij je sestavljen iz treh glavnih komponent:

- segmentacija posnetka, s čimer razdelimo celoten posnetek na več krajših delov, hkrati pa odstranimo šum in tišino;
- razpoznava govora, s čimer iz avdio signala pridobimo približno tekstovno transkripcijo;
- poravnava, s čimer vsaki besedi v originalnem besedilu določimo mesto v pridobljeni transkripciji in s tem tudi čas pojavitve.

2.1 Segmentacija posnetka

Segmentacija je osnovana na Googlovem WebRTC–VAD algoritmu[1], ki je hiter, robusten in v praksi pogosto uporabljen. S tem algoritmom lahko klasificiramo posamezen časovni okvir kot govor ali ozadje. Algoritem robustne segmentacije je povzet po izvorni kodi, uporabljeni v sistemu DeepSpeech (Hilleman et al., 2018). WebRTC–vad ima nastavljiv parameter *aggresiveness*, ki lahko zasede vrednosti med 0 in 3. Parameter smo nastavili na vrednost 2, tako smo dobili dovolj kratke segmente, da proces dekodiranja pri razpoznavi govora ni trajal predolgo.

2.2 Razpoznava govora

Razpoznava govora je implementirana v dveh delih: 1) uporaba globokega akustičnega modela za pridobitev verjetnosti posameznih znakov za vsak časovni okvir in 2) dekodiranje izhoda modela za pridobitev končne transkripcije.

Podatki za učenje akustičnega modela so bili pridobljeni iz različnih virov: Gos (Zwitter et al., 2013), Gos VideoLectures (Videolectures, 2019), CommonVoice[2] , SiTEDx (Žgank et al., 2016), Sofes (Dobrišek et al., 2017) in narečni govor s portala narecja.si[3].

Akustični model je implementiran z uporabo ogrodja Nvidia NeMo, uporabili smo globoki model QuartzNet_15x5 (Kriman et al., 2019). Uporabili smo ga, ker lahko z njim kljub relativno majhnem številu parametrov (18,9 milijona) še vedno dobimo dokaj dobro natančnost razpoznave, primerljivo z večjimi modeli (več kot 100 milijonov parametrov). Primerjali smo dva modela: QuartzNet_15x5, učen zgolj na slovenskih podatkih, in QuartzNet_15x5, predučen na angleških podatkih, nato pa dodatno učen še na slovenskih podatkih. S slednjim modelom smo preverili kvaliteto prenosa znanja iz tujega jezika v slovenščino.

Za pridobitev transkripcij smo primerjali tri različne metode dekodiranja CTC: 1) požrešna metoda največjih verjetnosti (*greedy*), kjer za vsak časovni korak v CTC izberemo najbolj verjeten znak, nato združimo sosednje

[1]Webrtc google repository.
https://chromium.googlesource.com/external/webrtc/+/branch-heads/43/webrtc/common_audio/vad
[2]Mozzilla Common Voice website. https://commonvoice.mozilla.org/sl/datasets.
[3] https://narecja.si/

ponovitve; 2) iskanje v snopu z besednim jezikovnim modelom (*word*) in iskanje v snopu z znakovnim jezikovnim modelom (*char*).

Za jezikovni model smo uporabili N–gram jezikovni model KenLM (Heafield, 2011). Ker se model uporablja zgolj med dekodiranjem CTC za posamezen primer poravnave, smo za gradnjo modela uporabili kar originalno besedilo posameznega primera. Tako dobimo model, ki ni posplošen za slovenski jezik, temveč je prilagojen posamezni poravnavi. Testi so pokazali, da red jezikovnega modela ne vpliva bistveno na rezultat, na koncu smo uporabili model četrtega reda.

2.3 Poravnava in iterativno združevanje

S pomočjo razpoznavalnika govora iz posnetka pridobimo približno transkripcijo govora. Le-to moramo v zadnjem koraku poravnati z originalnim besedilom posnetka. Za osnovno poravnavo uporabimo algoritem povzet po orodju DeepSpeech. Izkaže se, da z uporabo tega algoritma ne zagotovimo poravnave vseh besed originalnega besedila. Krajše besede pogosto nimajo nujno dovolj konteksta ali pa so slabo transkribirane. Da zagotovimo poravnavo vseh besed, smo razvili algoritem iterativnega združevanja besed.

Glavna ideja algoritma je naslednja: besede, ki niso poravnane, združimo s sosednjo besedo v besedilu (odstranimo presledek in tvorimo enoten niz znakov). Osnovni algoritem poravnave ponovno poženemo, tokrat z modificiranim seznamom besed. Ta dva koraka ponavljamo, dokler niso vse besede (oziroma skupki besed) poravnani, nato lahko vsaki besedi originalnega besedila pripišemo začetni in končni čas glede na približno transkripcijo.

3 Evalvacija

Natančnost sistema smo ovrednotili na testni množici s primerjavo z ročno izdelanimi poravnavami. Za oceno kvalitete poravnave uporabljamo tri mere: povprečje (MAE) in standardni odklon (STD) absolutnih napak začetnih časov besed ter delež absolutnih napak, manjših od 0,5 sekunde (< 0,5s).

3.1 Testna množica

Testno množico sestavlja 26 primerov: 7 primerov nenarečnega govora, 13 primerov narečnega govora in 6 primerov narečnega enoglasnega petja brez spremljave. Najkrajši posnetek je dolg 21 sekund, najdaljši 219, povprečna dolžina posnetkov je 89 sekund. Primeri so pridobljeni iz naslednjih virov: Slovenske ljudske pesmi V (Kaučič et al., 2007), portal narecja.si, terenski posnetki GNI ZRC SAZU. Pravilne poravnave so bile narejene ročno z orodjem Praat.

Tip posnetka	Število besed	Dolžina (min)
narečni govor	2428	18,7
nenarečni govor	1394	11,0
narečno petje	508	8,7
skupaj	4330	38,4

Tabela 1: Testna množica.

3.2 Primerjava modelov in metod dekodiranja

Primerjali smo osnovni akustični model (*base*), ki je grajen zgolj na slovenskih podatkih, ter model, ki je učen na angleških podatkih, nato pa doučen na slovenskih (*transfer*). Ob tem smo primerjali tri metode dekodiranja: požrešna metoda (*greedy*), iskanje v snopu z jezikovnim modelom na nivoju znakov (*char*), iskanje v snopu z jezikovnim modelom na nivoju besed (*word*). Primerjavo smo opravili za vsak tip testnih podatkov posebej. Rezultati so podani v Tabeli 2.

Iz tabele je razvidno, da pri nenarečnem govoru ne glede na metodo uporaba modela *transfer* prinese manjšo povprečno napako. Razlika je sicer majhna (0,06 do 0,07 sekunde), vendar je približno enaka za različne metode. Pri uporabi požrešne metode ima *transfer* sicer večji standardni odklon in manjši delež napak pod 0,5s, vendar je razlika minimalna. Različne metode dajejo zelo podobne rezultate. Kombinacija modela *transfer* in metode *word* da najboljši rezultat s povprečno napako 0,12s, standardnim odkonom 0,10s in 99,4% deležem napak pod 0,5s.

Tudi v primeru narečnega govora uporaba modela *transfer* izboljša rezultate. Razlika v povprečnih napakah je majhna (0,04 do 0,09 sekunde), vendar je med akustičnima modeloma opazna razlika tudi v standardnem odklonu in deležu napak manjših od 0,5s. Z uporabo modela *transfer* so rezultati za različne metode poravnave zelo podobni, pri čemer se metoda *word* izkaže za najbolj robustno, saj ima najmanšo napako in standardni odklon pri obeh modelih. Pri modelu *transfer* ima metoda *greedy* sicer nekoliko večji delež napak pod 0,5s,

vendar je razlika majhna (0,4%). Kombinacija modela *transfer* in metode *word* da najboljši rezultat s povprečno napako 0,14s, standardnim odkonom 0,24s in 97,3% deležem napak pod 0,5s. V primerjavi z najboljšim rezultatom nenarečnega govora se povprečna napaka poveča za 0,02s, standardna deviacija za 0,13s, delež napak pod 0,5s se zmanjša za 2,1%. Razlika ni velika in je približno podobna za ostale kombinacije metod in modelov.

tip testnih podatkov	metoda	model	MAE	STD	< 0,5s
Nenarečni govor	greedy	base	0,20	0,13	99,1%
		transfer	0,14	0,15	98,5%
	char	base	0,21	0,09	99,0%
		transfer	0,14	0,10	98,9%
	word	base	0,19	0,10	98,6%
		transfer	**0,12**	0,11	99,4%
Narečni govor	greedy	base	0,22	0,39	94,9%
		transfer	0,15	0,27	97,7%
	char	base	0,21	0,32	95,7%
		transfer	0,15	0,28	97,1%
	word	base	0,18	0,28	97,2%
		transfer	**0,14**	0,24	97,3%
Narečno petje	greedy	base	0,59	0,82	70,2%
		transfer	1,28	2,49	63,9%
	char	base	0,82	1,66	66,7%
		transfer	0,44	0,41	73,4%
	word	base	0,48	0,58	73,4%
		transfer	**0,37**	0,30	79,9%

Tabela 2: Rezultati

Pri narečnem petju je napaka poravnave opazno večja. Pri metodah *word* in *char* akustični model *transfer* deluje bolje. Z metodo *char* je povprečna napaka prepolovljena, standardni odklon je štirikrat manjši, delež napak pod 0,5s se izboljša za 6,7%. Z metodo *transfer* je povprečna napaka za 0,11s manjša, standardni odklon za 0,28s, delež napak pod 0,5s se izboljša za 6,5%. Pri metodi *greedy* je boljši model *base*, kar je edini tak primer v rezultatih. Rezultati različnih metod dekodiranja med seboj niso podobni. Pri obeh modelih metoda *word* bistveno izboljša rezultat. Kombinacija modela *transfer* in metode *word* da najboljši rezultat s povprečno napako 0,37s, standardnim odkonom 0,30s in 79,9% deležem napak pod 0,5s. V primerjavi z najboljšim rezultatom nenarečnega govora se povprečna absolutna napaka poveča za 0,25s, standardna deviacija za 0,19s in delež napak pod 0,5s se zmanjša za 19,5%. Razlika je velika in je vidna tudi pri ostalih kombinacijah metod in modelov. Povprečna absolutna napaka se poveča za faktor vsaj 2,5, standardni odklon za faktor vsaj 2,7 in delež napak pod 0,5s se zmanjša za vsaj 19,5%.

3.3 Ugotovitve

Kvaliteta poravnav na nenarečnem govoru se izkaže za dobro in je primerljiva s podobno delujočimi sistemi, npr. (Malfrère et al, 2003). Tudi pri narečnem govoru je kvaliteta poravnav dobra. Napaka je nekoliko večja kot pri nenarečnem govoru, kar je pričakovano, saj je večina učnih podatkov za akustični model nenarečnih. V splošnem ocenjujemo, da sistem dobro deluje na slovenskem govoru in je zato uporaben za večino aplikacij. Vredno je omeniti, da v primeru kratkih posnetkov in popolnih transkripcij za učenje akustičnih modelov obstajajo potencialno boljše tehnike poravnave (Brognaux in Drugman, 2015).

Kvaliteta poravnav enoglasnega petja brez spremljave je v primerjavi z govorom opazno slabša, kar smo tudi pričakovali, saj je v splošnem poravnava petja in besedila težji problem. V primerjavi z nenarečnim govorom je povprečna napaka približno trikrat večja in veliko več je napak večjih od pol sekunde. Povprečna napaka je sicer primerljiva s podobno delujočim sistemom za poravnavo petja (Stoller et al., 2019), vendar naši testni podatki ne vključujejo večglasnega petja ali petja s spremljavo, zato ta primerjava ne pove veliko. Domnevamo, da bi se kvaliteta poravnav bistveno izboljšala, če bi učna množica akustičnega modela vsebovala petje.

V veliki večini primerov se akustični model *transfer* izkaže bolje od modela *base*. Edini obraten primer je v primeru petja in metode *greedy*, kjer model *base* doseže boljši rezultat, vendar ker ta kombinacija metode in modela ne da najboljšega rezultata pri petju, ni bistvena za oceno kakovosti. Na podlagi rezultatov potrjujemo domnevo, da prenos znanja z modelom *transfer* pozitivno vpliva na kvaliteto poravnave tako pri govoru kot pri petju.

Čeprav je v primeru govora najboljša metoda za dekodiranje *word*, ostali dve metodi nimata bistveno večjih napak. V primeru nenarečnega govora z modelom *transfer* je povprečna napaka z metodo *word* manjša za 0,02s, v primeru narečnega govora pa za 0,01s. V aplikacijah, ko zelo natančna poravnava govora ni ključna, je pa pomemben čas računanja, je bolj smiselno uporabiti metodo *greedy,* saj le-ta ne zahteva iskanja v snopu ter uporabe jezikovnega modela in je zato bistveno hitrejša. Pri petju metoda *greedy* da bistveno slabše rezultate od metode *word*, zato je smiselno uporabiti slednjo.

Zahvala

Raziskave, opisane v prispevku, so bile opravljene v okviru temeljnega raziskovalnega projekta »Misliti folkloro: folkloristične, etnološke in računske perspektive in pristopi k narečju« (J7-9426, 2018-2022), programske skupine »Digitalna humanistika: viri, orodja in metode« (P6-0436, 2022-2027), oba financira ARRS, in raziskovalne infrastrukture DARIAH-SI.

Literatura

Sandrine Brognaux in Thomas Drugman. *Hmm-based speech segmentation: Improvements of fully automatic approaches*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24:1–1, 01 2015.

Simon Dobrišek, Jerneja Žganec Gros, Janez Žibert, France Mihelič, in Nikola Pavešić. *Speech database of spoken flight information enquiries SOFES 1.0*, 2017. Slovenian language resource repository CLARIN.SI.

Kenneth Heafield. *KenLM: Faster and smaller language model queries*. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 187–197, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

Ryan Hilleman, Tilman Kamp in Tobisas Bjornsson. *Dsalign*. https://github.com/mozilla/DSAlign, 2018.

Marjetka Golež Kaučič, Marija Klobčar, Zmaga Kumer, Urša Šivic, and Marko Terseglav. *Slovenske ljudske pesmi V*. 2007.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li in Yang Zhang. Quartznet: *Deep automatic speech recognition with 1d time-channel separable convolutions*, 2019.

F. Malfrère, O. Deroo, T. Dutoit, in C. Ris. *Phonetic alignment: speech synthesis-based vs. viterbi-based*. Speech Communication, 40(4):503–515, 2003.

Daniel Stoller, Simon Durand, in Sebastian Ewert. *End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model*, 2019.

VideoLectures.NET. *Spoken corpus gos VideoLectures 4.0 (audio)*, 2019. Slovenian language resource repository CLARIN.SI.

Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej in Tomaž Erjavec. *Spoken corpus gos 1.0*, 2013. Slovenian language resource repository CLARIN.SI.

Andrej Žgank, Mirjam Sepesy Maučec in Darinka Verdonik. *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), pages 4670–4673, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

A Parallel Corpus of the New Testament:

Digital Philology and Teaching the Classical Languages in Croatia

Petra Matović,[*] Katarina Radić[†]

[*] Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
pmatovic@ffzg.hr

[†] Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
katarina.radic1@gmail.com

1. Introduction

Corpus linguistics has been one of the liveliest disciplines in Croatian linguistics, and parallel corpora have been established by Croatian scholars since the 1960s (Tadić 1997, 2001; Simeon 2002). These corpora normally include Croatian and another living language, while corpora consisting of texts in Croatian and at least one of the so-called "dead" languages are still underrepresented, although there are corpora including languages like Ancient Greek, Latin, Sanskrit, Arabic, Persian and Akkadian, to be found on the World Wide Web (The Alpheios Project 2019; Palladino et al., 2021). The Department of Classical Philology at the University of Zagreb can already boast one of the earliest online (monolingual) corpora of Latin texts, the CroALa database, built and curated by Neven Jovanović (CroALa, 2014). In the last few years the said department has been steadily building small parallel corpora, and this paper aims to describe one of them, the Greek-Croatian parallel corpus of the New Testament, currently in the making, and furthermore discuss its educational uses in teaching Ancient Greek.

2. Goal of the paper

Building parallel copora has been garnering more and more attention in the field of classical philology. The Department of Classical Philology at the University of Zagreb has been building smaller corpora, both as a part of several small-scale projects lead by Neven Jovanović and courses on Greek and Latin language (e.g. Soldo and Šoštarić 2019). Since 2021, several professors and students at the department have been working on project titled "A Linguistic Analysis of Selected Early Christian Writings", lead by Petra Matović. Within the scope of the project we have started building a parallel corpus of the New Testament, so far comprising the Gospel of Mark and a part of the Apocalypse. The texts are aligned using the Alpheios tool for text alignment at the Perseids environment (The Alpheios Project, 2019; The Perseids Project, 2017). Alpheios enables the user to align words or word combinations in the source text with corresponding parts of its translation (The Alpheios Project, 2019). In this poster we firstly aim to explain the principles of alignment we followed while building the corpus, and, secondly, discuss some peculiarities in aligning Ancient Greek with Croatian. Finally we aim to look at the corpus from an educational point of view and discuss its possible uses in teaching Ancient Greek today.

Text aligment was done by 4 students (Mateo Cader, Ružarijo Lukas, Katarina Radić, Luka Šop) and supervised by Petra Matović. The editions of the texts were Nestle-Aland 28 (Greek New Testament) and the so-called Zagreb Bible (https://biblija.ks.hr/). Initially, the main principle of alignment was to align units (words or word combinations) in the Greek text with their Croatian counterparts; these units had to be as small as possible. Full stops and commas were aligned, too. After the initial period it became clear that additional rules were necessary. While the students did not struggle with the meaning of the Greek text, they were sometimes unsure how to align the Greek with the Croatian. These uncertainties typically arose in the following situations due to specific linguistic features of the two languages:

- the use of the article (exist in Greek, but not in Croatian: ὁ Ναζαρηνός = Nazarećanin, Mark 10,47)
- commas can be aligned with conjunctions
- participles (extensively used in Greek, not common in Croatian: ἀκούσας = kad je čuo, Mark 10,47)
- particles (Greek is rich in particles, Croatian often lacks equivalents: the particle δέ is translated as "ali" in Mark 13,5, but left untranslated in Mark 13,13)
- features of Hellenistic Greek (The New Testament was written in this later variety of the Greek language, which is often different from the Classical, 5th century BC Attic dialect of Greek which is mainly taught in schools and universities; one of theseis the preterite form ἤμην διδάσκων = „naučavah" Mark 14,49).

There was also one unexpected problem: students often struggled with aligning prepositions, for example in Mark 1,6: ἐνδεδυμένος τρίχας καμήλου καὶ ζώνην δερματίνην περὶ τὴν ὀσφὺν, the preposition "s" was left unaligned in the Croatian translation ("odjeven u devinu dlaku, s kožnatim pojasom oko bokova").
Consequently, the following set of rules was formed:
1. The article is aligned together with the corresponding nouns, unless translated separately.
2. Conjunctions should be aligned either with conjunctions, particles or punctuation.
3. Punctuation should be aligned whenever possible.
4. Participles should be aligned with the corresponding word combination, even if it is an entire sentence.
5. If something is left out in the translation, the Greek original is left unaligned and *vice versa*, for example the verb "to be".
6. Prepositions should never be left unaligned. Whenever possible, they should be aligned with a corresponding Greek preposition. In the case where a preposition is added in Croatian, together with its noun it should be aligned with the corresponding noun in Greek.

The work done on this corpus highlights several problems in teaching not only Ancient Greek, but also Croatian. Students are unsure of the uses of certain parts of speech, usually those parts of speech that do not have an equivalent in their mother tongue. They are also unaware of the nature of the comma, which can connect (or divide) two words just like a conjunction. Prepositions are often an obstacle because their meaning can be incorporated into a nominal form in Greek and does not have to be expressed separately. These problems probably arise because the school curriculum for Croatian is different from the curricula for Greek and Latin: the curricula for the classical languages pay more attention to grammar, while Croatian has to include both language and literature. Hopefully, projects like this one can highlight specific problems that can then be resolved either by adapting the school curricula or the teaching of classical languages on university level.

3. References

The Alpheios Project. 2019. https://alpheios.net/.

CroALa (Croatiae Auctores Latini). 2014. http://croala.ffzg.unizg.hr.

Chiara Palladino, Maryam Foradi, and Tariq Yousef. 2021. Translation Alignment for Historical Language Learning: a Case Study. *Digital Humanities Quarterly*, 15(3). https://www.proquest.com/openview/e048d32e8e991c67282c3fbda5c1f0d4/1?pq-origsite=gscholar&cbl=5124193.

The Perseids Project. 2017. https://www.perseids.org/.

Ivana Simeon. 2002. Paralelni korpusi i višejezični rječnici. *Filologija*, 38-39: 209–15.

Petar Soldo and Petra Šoštarić. 2018. Treebanking Lucian in Arethusa: Experiences, Problems and Considerations. *Studia UBB Digitalia* 63(2):7–18.

Marko Tadić. 1998. Raspon, opseg i sastav korpusa suvremenoga hrvatskoga jezika. *Filologija* (30-31):337–47.

Marko Tadić. 2001. Procedures in Building the Croatian-English Parallel Corpus. *International Journal of Corpus Linguistics*, *Special issue, pages* 1–17.

Novum Testamentum Graece, Neste-Aland 28. https://www.academic-bible.com/en/online-bibles/novum-testamentum-graece-na-28/read-the-bible-text/bibel/text/lesen/stelle/51/10001/19999/ch/418f354347a79b322324823db62504dc/.

The Zagreb Bible https://biblija.ks.hr/.

Pre-Processing Terms in Bulgarian from Various Social Sciences and Humanities (SSH) Domains: Status and Challenges

Petya Osenova*, Kiril Simov*, Yura Konstantinova†

*Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Acad. G. Bonchev bl. 2, 1113 Sofia
{petya, kivs}@bultreebank.org

†Institute of Balkan Studies and Centre of Tracology, Bulgarian Academy of Sciences
Moskovska St 45, 1000 Sofia
yura.konstantinova@balkanstudies.bg

Abstract

1. Introduction

There exists a great number of focused initiatives, projects and conferences that tackle deeply various topics related to terminology construction, understanding, processing and usage. We will mention only a small part of them here rather as initiatives than as distinct publications. These are, among others, ENeL COST Action on e-Lexicography,[1] related activities in the NexusLinguarum COST Action,[2] related activities in the ELEXIS project,[3] globaLEX organization. There is also ongoing work on providing language technology help to colleagues in SSH within CLARIN-ERIC and DARIAH.[4,5]

Within the CLaDA-BG infrastructure,[6] which combines the goals of CLARIN and DARIAH in Bulgaria, there are two types of partners – technological ones and colleagues also from SSH. The latter are historians, ethnographers, specialists in the deeds and lives of Cyril and Methodius, museum and library workers. This combination of complementary partners allows us to construct the necessary resources and immediately to verify their utility for SSH partners.

In the task of creating the Bulgarian-centric Knowledge Graph (BGKG) within CLaDA-BG (Simov and Osenova, 2020) we requested data from our SSH partners in order to perform linguistic pre-processing and to enhance the creation of terminological dictionaries that cover the SSH subdomains based on these data.

The size of the corpus is nearly half a million – 484,815 tokens. The selected words and phrases for pre-processing and creation of entries towards terminological dictionaries were about 5,000 within nearly 26,000 usages annotated within the corpus. From them the rejected candidates, or the false positives, were 542 candidate phrases. Out of them 328 candidates were completely rejected either because they were named entities or free compositional phrases.

Thus, our colleagues from SSH would facilitate their own work with only checking and validating the previously pre-processed data. The data consists of selected texts from various sources such as: scientific texts authored by our SSH colleagues and related to Bulgarian history and society; Linked Open Data like Wikipedia; available textbooks, specialised dictionaries etc.

Here we give a brief outline of our pre-processing strategy towards handling the data-driven terminology in these domains.

2. The Task Overview

The work flow that is discussed here is related to the SSH data (publications, autobiographies, archive documents, newspaper articles from past periods, descriptions of artefacts, etc.) that were collected from partners, and annotated within the INCEpTION platform[7] with named entities, events and roles. Thus, while annotating linguistically the texts, the annotators were additionally asked to mark candidate terms with the label *term*. This task was set in the view of the subsequent creation of specialised terminological dictionaries

[1] https://www.cost.eu/actions/IS1305/
[2] https://nexuslinguarum.eu/
[3] https://elex.is/
[4] https://www.dariah.eu/
[5] https://www.clarin.eu/
[6] https://clada-bg.eu/en/
[7] https://inception-project.github.io/

in each participating SSH domain – history, ethnography, biographical studies, etc. The annotators were instructed to view as candidate terms the keywords that are specific for the domain.

Later on, these candidate terms were extracted and transferred to a huge excel table in Google Drive. The table consists of three main areas: a) the candidate term, b) the term in its context of occurrence, and c) the source that delimits the domain of usage. In Figure 1 an excerpt from the excel view is presented:

	Зимно време конете, заедно с другите хайва́ни, държат в айр или в да́ма за работния добитък /инф. 6/. На @@@ *ездитния кон* @@@ на гърба слагали само едно черджѐ , седло́ не е имало . Конския юла́р е бил от въжета, но в града ги правели – мешинови /инф. 6/; „моят баща е	
ездитния кон	запомнил в селото 60 коня“ /инф. 6/.	Etnographs-text01-04
ездитен кон	кон, който се използва за езда	етно

Figure 1: An example from the excel table.

In the first row the following information is given: the term as it occurred in the text (riding-the horse, ‘the riding horse’), the text excerpt with the term placed among the symbols @@@, and the name of the source text. In the second row the following information is given: the normalised term (*riding horse*), the definition *(a horse that is used for riding*) and the domain – ethnography.

All the one-word terms got initial definitions from the digitised version of the Explanatory dictionary of Bulgarian (Popov et al., 1994). This step was performed automatically through a rule-base matching method. First, the word forms in the texts were lemmatized with our in-house Inflectional Bulgarian dictionary. Then the coinciding lemmas in the dictionary and in the texts were matched. The terms with more than one meaning also received all the possible definitions automatically. Afterwards, these candidate terms were processed manually by the team that previously worked on the event and roles annotations. The core team engaged with the terminology pre-processing consisted of 4 members as a subpart of the whole annotating team that consisted of 8 people.

The tasks related to the terminology processing were organized as follows: one person (outside the 4 working colleagues) performed the automatic construction of the table and the assignment of the existing definitions and sources. Initially the candidate terms were assigned in an alphabetical order to workers, i.e. each colleague was responsible for the candidate terms that began with certain letters. However, after having completed some letters, a decision was taken to go by domain source instead. This approach allowed us to observe the terms in their domain contexts and interrelations. Then, once more the terms were checked in their alphabetical appearance.

The workflow was generally divided into two phases that respects the competences of the experts. In *Phase 1* the corpus linguists (who were also annotators) pre-processed the candidate terms while in *Phase 2* the specialists in SSH areas are supposed to check and validate these terms against their own area.

3. The Workflow

The respective annotated data was uploaded in advance including the annotated candidate term. The workflow consisted of the following steps:

3.1 Deciding which candidate terms are true terms

Here the main task of the corpus linguists was to try to reduce the list of the obvious non-terms or the common words and expressions from the specialised terms. Sometimes the boundaries were not very clear, especially with respect to the multiword expressions (MWE) and the nested terms. See more about this issue in point 3 below. The annotators had at their disposal three options to select from: *a sure term*, *a maybe term* and *a non-term*.

3.2 Checking the availability of the definition and its relevance

In case there was a definition, the annotator had to: accept it as it is, reject it or modify it. If there was no definition, the annotator had to create one. When the term was one-word, the task was to check the definition that came from the Explanatory dictionary of Bulgarian. In case of lexical ambiguity the annotator had to select the correct definition among the available ones, or again - to provide their own, if no appropriate is present. Then the selected definition was marked as the right one. Please note that the other definitions were not deleted for the sake of completeness and future addition into BTB-Wordnet.

3.3 Handling multiword expressions

Here the prevailing part of terms consisted of a head noun and a pre-positioned modifier. For example, *демокрация* (democracy) and *пряка демокрация* (direct democracy). The problems might go into two directions: to accept a MWE as a domain term or not, and to provide a definition of it since it is usually not available in the consulted sources. We decided to be inclusive in accepting what had to be a term. This means that all the expressions that were considered specific for the domain, were approved. The annotator could also add the definitions about the parts of compositional MWEs. For example, *невалиден глас* (invalid vote) can have a definition as a phrase, while its two elements *невалиден* (invalid) and *глас* (vote) might also be added below with their own definitions.

3.4 Re-checking the domain/genre.

This step relies on the domain/genre classification that has already been used. Thus, an initial pre-defined schema was explored that in the process of work was further expanded and hierarchized. At the moment the list of the applied domains amounts to 76 categories (for example, architecture with a subdomain of construction; geography with a subdomain of geology; philosophy with subdomains of ethics, rhetorics and logic) and the registers to 15 (for example, dialectal, metaphorical, colloquial, etc.). The initial schema came from the classifications used in the Explanatory dictionary and had 36 domains and 4 registers. At the beginning, we tried to keep the terms in separate groups that do not overlap: history, ethnography, etc. These areas however are highly interdisciplinary and they inter-cross with each other. For that reason this approach was abandoned at a very early stage in our work. In this way one and the same term could be put in more than one domain with the same or different meaning.

Other tasks that were part of the workflow – although with a lower priority were:

3.5 Adding other senses of the lemma of the term, and

3.6 Adding examples to these additional senses.

The idea behind tasks 5 and 6 was that we aim at reaching better coverage also in other language resources like BTB-Wordnet (Osenova and Simov, 2018) and at compiling a sense corpus per lemma and usage.

The results of this preparatory work was a classification of the initially selected about 5000 candidate terms and keywords with respect to the hierarchy of domains. This allows the further processing to be done by different experts in the corresponding domains. Their tasks are the following:

Final Sorting of lexical items within true terms and keywords.

As it was mentioned above, the examples annotated within the domain documents were classified within two main categories: general lexica and compositional phrases, on the one hand, and terms, on the other. The second group sometimes contains keywords that happen to be true terms in the respective domain. Thus, the first task for the experts was to sort out the true terms.

Addition of missing terms.

Despite the wide range of documents selected for annotation, they do not contain all the relevant terms in the domain. For example, from the set of all genres of Old Bulgarian literature, only three were identified within the annotated documents. The rest of terms for these identified genres were added by the experts of Old Bulgarian literature. Thus, by completing the missing slots, we expect that each domain will have a relatively complete list of terms.

Extension of the definitions.

In the original list of candidate terms we had to also add definitions from online sources or to construct our own definitions. Since the pre-processing group included linguists, but not experts in the domains, very often the definitions were not complete and/or precise enough. Thus, the domain experts extended the definitions with encyclopaedic information. In some cases also appropriate images were added. The resulting encyclopaedic entries were cross linked on the basis of the included terms. Here is an example of such an entry from the area of architecture:

АЖУР техника при *резбарското*, *златарското*, *плетаческото* и други изкуства, при която между декоративните елементи има отвори	

Figure 2: One example from the terminology lexicon in the area of architecture. On the left, the term *openwork* and the definition (ornamental work such as embroidery or latticework having a pattern of openings) are given, and on the right there is an image illustrating it. The links to other terms are represented via italicising the corresponding words/phrases in the definition. This example is only illustrative. The actual entries might contain longer texts, references to relevant literature, more images and links to external resources.

The resulting terminological lexicons are further processed by the team working on the Bulgarian Bultreebank Wordnet (BTB-WN). This work has been done in cooperation with the domain experts. Such alignment of the terminological lexicons and wordnet allows for a joint usage of both lexical resources for the main use cases – explanation of the specific knowledge in the domains and indexing of various types of domain documents. Figure 3 depicts a part of the hierarchy of Bulgarian folk units of measurement. They are linked with a hyponymy relation to the concept for *Bulgarian folk units* and the concept for *linear units*.

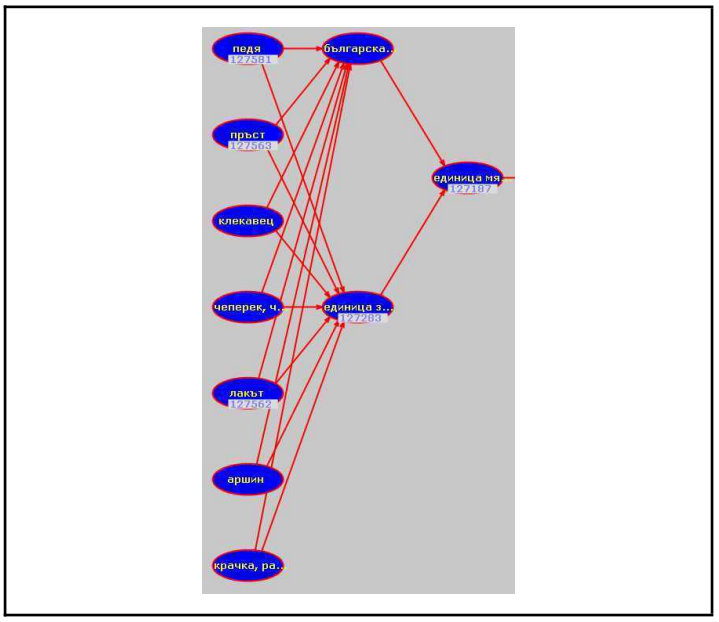


Figure 3**:** In this figure a graphical view on Bulgarian folk units of measurement is presented. Each term is classified into two ways - as a unit of measurement for distance (*linear units*) and that its domain is Bulgarian folk units. The hierarchy of terms could interleave with synsets that are not terms in the domain.
The mapping to synsets in the English WordNet are given with identification (IDs) at the lower part of the graphical representation of each Bulgarian synset. Here measures are given such as *педя* (span), *пръст* (finger), *лакът* (elbow), etc.

Our idea is BTB-WN to be the main resource within CLaDA-BG for representation of lexical data related to general language, terminology and to be aligned to the ontologies on which BGKG is constructed.[8] In this way we hope to be able to provide access to these data by different types of users with different knowledge about the domains, with different goals in mind, etc.

In addition to the standard wordnet relations (hypernymy, meronymy, etc.), we envisage other semantic relations that represent various aspects of knowledge within the corresponding domains. In this way, we will ensure the representation of encyclopaedic information and will facilitate the representation of Named Entities (NEs) classified with respect to the corresponding concepts. This approach relies on specially created templates based on the domain relations as well as their domain and range restrictions. We already defined about 20 such templates for main classes of NEs like geopolitical entities, historical events (wars, uprisings, etc.), artefacts (icons, stamps, ect), political parties and regimes, and so on.

4. Conclusions

In this extended abstract we described the main steps that were followed in the creation of terminological lexicons in a bottom-up approach starting from real texts within SSH domains. After the domain texts were annotated with named entities, events, roles and candidate terms, a concordance of the latter from different documents was performed where they were grouped together and linguistically processed. As a result, they had the representation of the term in its basic form, listings of related words for MWEs, the existing potential

[8] This approach is similar to the lexeme assignment in Wikidata.
https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation

senses from different sources (available locally to the annotators and on the web). The appropriate senses for the given context were selected or created. Then the result was further processed by the domain experts in order to make the definitions more precise and complete. Also, an addition of missing terms was performed. Then the terminological lexicons were aligned to the BTB-WN in order to be used for navigation, annotation of more documents (manually or automatically) and to establish links to the necessary ontologies.

The main challenges can be divided as either technical or theoretical. In the first group we can mention the insufficient context, lack of enough sources for terms related to previous historical times; approaching the task in the best way - alphabetically or by source, etc. In the second group we can outline: the difficulty to differentiate between a term and a non-term; aiming at the most informative definition when there are too many found in the sources; finding and/or constructing a definition when it lacks or is wrong with the help of other available resources; handling close definitions for some lemma; construction of a definition for multiword terms; handling multi-domain inclusion of terms.

5. References

Petya Osenova and Kiril Simov. 2018. The data-driven Bulgarian WordNet: BTBWN. In: *Cognitive Studies | Études cognitives*, vol. 18, https://doi.org/10.11649/ cs.1713.

(freely available at: https://ispan.waw.pl/journals/index.php/cs-ec/article/view/cs.1713/4458)

Kiril Simov and Petya Osenova. 2020. Integrated Language and Knowledge Resources for CLaDA-BG. In: *Selected Papers from the CLARIN Annual Conference 2019*, 172 (2020), LiU Electronic Press: Linköping Electronic Conference Proceedings 172 (2020), 2020.

Dimitar Popov et al. 1994. D. Popov, L. Andreychin, L. Georgiev, St. Ilchev, N. Kostov, Iv. Lekov, St. Stoykov and Tsv. Todorov 1994 . *Bulgarian Explanatory Dictionary*. Nauka i izkustvo. Sofia. (in Bulgarian)

An Approach to Computational Crisis Narrative Analysis: A Case-study of Social Media Narratives Around the COVID-19 Crisis in India

Henna Paakki*, Faeze Ghorbanpour*, Nitin Sawhney*

* Department of Computer Science, Aalto University
P.O.Box 15400, FI-00076 AALTO, Espoo, Finland
henna.paakki@aalto.fi

1. Introduction

Societal crises create an empty narrative space and a need for explanation about the crisis, related risks and required mitigation actions (Sellnow et al., 2019). Crises are socially constructed through discourses and have the potential to change social structures and perceptions (Walby, 2015). Crisis narratives also have an important role in attributing blame and structuring crisis responses and recovery plans (Walby, 2015, p. 14). The role of social media has increased significantly as a forum for seeking information about crises, as well as for discursive sense-making. People use discourses and narratives related to a crisis to construct the world socially and epistemologically and to explain the impending crisis (Joffe, 2003; Bednarek et al., 2022), which makes it important for authorities, experts and crisis regulators to understand various discourses around the crisis. This paper examines the possibilities for analyzing social media discourses using a novel computational approach, using a discourse act classifier based on zero-shot learning (Yin et al., 2019) to categorize discourse types into narrative function groups (Labov, 1972). Such tools can help support other means of inquiry and crisis preparedness. Our empirical case study examines discourses around the COVID-19 pandemic in the context of English-language social media in India. This abstract describes an ongoing research project.

2. Goal of the paper

As crisis discourses on social media encompass a large set of data, there is a need for computational methods that can support close readings. Although some methods have been developed for computational discourse and narrative analysis (Piper et al., 2021), this line of research needs more tools. Lakoff and Narayanan have proposed that computational narrative analysis could be approached by focusing on the structural building blocks of narratives (Lakoff and Narayanan, 2010), which have been outlined in linguistics and social sciences (Labov 1972; Labov and Waletzky, 1967; van Dijk, 1976). Such rules can aid computational models.

Narratives encase human motivations, goals, emotions, actions, events, and outcomes, elements that have been considered essential for computational models to understand (Lakoff and Narayanan, 2010). We posit that sense-making in crisis is action (Joffe, 2003), at the surface-level formulated as discursive actions (Edwards and Potter, 1993; Schegloff, 2007). Thus, for capturing social media narratives, we explore the validity of using a widely used and well-established narrative functions theory from linguistics (Labov, 1972; Labov and Waletzky, 1967) to categorize social media comments based on their functions. These functions have already been used to computationally analyze more traditional narratives like personal histories or short stories (see e.g., Li et al., 2017). We explore the possibilities for further extending their use to analyzing changes in social media discourses around crises.

Many narrative theories agree that a sequence of events that forms a narrative whole includes first 1.) an **orientation** to the story or situation (identifying the time, place, persons, and situation of the narrative), some type of 2.) **complication** or disruption (the core event that creates tension in the narrative), 3.) an **evaluation** (clarification of why or how the events are important), and finally 4.) a **resolution** (how the story ends or how the core problematic event is resolved) (Lakoff and Narayanan 2010; Labov, 1972; Labov and Waletzky, 1967; Todorov 1971; Van Dijk, 1976). Conflict in communication is central in the narrative space surrounding a crisis and needs to be managed for successful crisis mitigation (Sellnow et al., 2019). Central to crisis discourses are critical events that have transformative power: they mobilize discourses and transform perspectives on the crisis through conflict (Jørgensen & Phillips, 2002). Thus, we might expect crisis narratives to involve a significant complication phase that needs to be followed by a resolution phase.

We maintain that by analyzing the functional categories of orientation, complication, evaluation, and resolution, it is possible to understand shifts in perspectives to the ongoing crisis, ones that contribute to the narrativization of the crisis. Furthermore, we expect that it is possible to identify points of discursive struggle within crisis discourses, points where critical understandings of the crisis are negotiated to achieve a consensus or to legitimize a selected narrative (Jørgensen & Phillips, 2002; Sellnow et al., 2019). This is central in understanding how a consensus on crisis resolution is achieved. We seek to investigate the validity and utility of computationally categorizing social media crisis discourses based on their functions. We ask:

1. Can narrative functions be applied to analyzing online crisis discourses using a computational model? Are these functions operationalizable through discursive actions?
2. Do social media comments grouped by their actions correspond well enough to the functions of orientation, complication, evaluation, and resolution?
3. By using these function-based groupings, is it possible to find patterns of narrativization in online crisis discourses? Do comments have different functions at different points in time during the crisis?

3. Data sources and sampling

Crisis news reporting has a significant impact on citizen perspectives on the crisis (Kasperson et al., 1988). We are thus interested in this relationship between the evolution of crisis news discourses and how citizen discourses develop during a long-lasting crisis. YouTube news channels' crisis news videos and their comments offer an opportunity for investigating this interaction over time. We examine viewer comments to crisis news videos on English-language NDTV news' YouTube channel during the Covid-19 crisis in India, in conjunction with news reports and contextual insights on the pandemic. The data were collected using a scraper and the YouTube API. They involve the beginning of the crisis (1/2020-8/2020), acute vaccination phase of the crisis (02/2021-08/2021), and a later prolonged phase of crisis (11/2021-02/2022). Channel selection criteria included that the channel should be among the most followed English news providers in the country, one of the most trusted (Newman et al., 2021), that the channel allows viewer comments, has a wide viewership and is politically as close to the centre as possible. The Indian context is of interest as trust in news has been reported to be low (Newman et al., 2021), and as the Global South perspectives have not been sufficiently represented in research.

4. Methods

Our approach is mixed, utilizing computational modeling to analyze a large set of data to achieve reproducibility and quantifiability, but also employing qualitative close reading.

To operationalize the narrative functions theory, we posit that this can be approached through the pragmatic items of discursive actions, as these are often used to analyze accountability, agency, position, and intention in conversation (Edwards and Potter, 1993; Schegloff, 2007). We expect that in our social media data, the function of informing statements is to mostly **orient** to the crisis and to express beliefs; questions, accusations and challenges most often express a **complication** or problematize some aspect of the crisis; evaluations and appreciations mostly attempt to elaborate and **evaluate** the situation; and requests and proposals aim at a **resolution** of some aspect of the crisis (Couper-Kuhlen and Selting, 2017; Turowetz and Maynard, 2010). Thus, we argue that what a comment does can be used to conclude what function it has within the larger crisis narrative. The selection of actions is based on frameworks of core actions in social interaction (Clark and Schaefer, 1989), ones found relevant across different contexts (Stivers et al., 2010) and computer-mediated communication (Paakki et al., 2021).

We manually annotated a set of 438 social media crisis news comments with actions. First, two annotators independently annotated the same set of comments, then compared and negotiated their annotations and resolved all conflicts, analyzing especially the difficult cases. Then annotators resumed annotation work, and finally an inter-annotator score was calculated using Krippendorf's alpha. We achieved a score of 0.75, which indicates a good degree of agreement. Using the hand-labelled comments, we trained a classifier using few-shot learning (Yan et al., 2018), achieving an f1 score of 0.50. We also ran a zero-shot NLI classifier (Yin et al., 2019), which at the present time achieved better results (f1 0.61) and was thus used for labeling all comments. The labeling followed carefully prepared annotation guidelines based on the descriptions of actions in the literature (e.g. Couper-Kuhlen and Selting, 2017; Schegloff, 2007). The whole action annotation scheme involved 13 classes following research on which actions are relevant and common in computer-mediated communication (Paakki et al., 2021). It involved responsive actions (e.g. apology, acceptance) that were not included in the function groups. At this stage, we concentrate on the 8 actions described above.

We further sorted comments into groups based on their action label by using a python script. We proceeded to validate our approach by 1.) qualitatively analyzing the functions (per Labov, 1972) of a set of hand-labeled comments from a time-period from 17th-25th August 2021 (125 comments excluding doubles), based on their content, comparing this analysis to our action-based computational classification, and 2) using time-series analysis to investigate the emergence of function groups at different times during the crisis. We calculated a threshold to identify significant peaks in function group values ($1.5 \times \mathrm{SD}\ over\ group\ mean$). We suspected that if the narrative functions were applicable to analyzing social media crisis discourses, there should be significant changes in which function groups are most common in crisis comments at different times.

5. Results

Our validation step 1 shows that the computational classification of comments gives us similar results as our manual analysis. The time-period chosen involved an especially high amount of complication actions in both the manually annotated set as well as the computationally annotated set. These are mostly related to criticism or mistrust in authorities and the COVID-19 vaccine, comments about negative symptoms from the vaccine, confusion about who to trust and what to do, but also some arguments that support the authorities. The qualitative analysis of the functions of comments corresponds sufficiently well to the action-based computational categorization: most statements and announcements had an orientation function also in our qualitative analysis; most questions, accusations and challenges served a complication function; evaluations and appreciations corresponded well with the evaluation function; and requests and proposals mostly aimed at some type of a resolution. However, 10% of comments did not fall into the assumed function group based on action type. In some cases actions had another function than expected: informing statements sometimes provided a complication in a few cases where negative effects from vaccines were described, evaluations sometimes had an orientation function, and some long comments involved more than one significant function.

Secondly, our preliminary results from the time-series analysis show that there are significant changes in which functions crisis news comments have at different points of the crisis timeline. Within the NDTV crisis news comments during the early phase of the Corona crisis, there are more significant peaks in orientation or resolution oriented discourses. During the acute mid-phase of the crisis, the frequency of comments that have a complication function is significantly higher. At the last phase, functions become dispersed, i.e. none of the function groups come above the threshold. The time-series analysis is still a work in progress, but the results so far show that the crisis narrative achieves its most conflictive point at the acute mid-phase of crisis where COVID-19 vaccinations have become relevant.

6. Conclusion

Our results so far show that the Labovian narrative theory is to some extent applicable to analyzing crisis discourses on social media. The applied model allows us to analyze how the functions of discourses shift along the crisis timeline, and to identify significant points of discursive struggle. The operationalization of functions through actions seems to work sufficiently well, as it allows a justifiable and pragmatic frame for annotation, rooted in a well-researched field.

Based on our results, the action-based categorization has some limitations that need consideration, as the actions used do not always correspond to the expected function. However, the narrative function categories are highly abstract and thus difficult to classify as such, as we found in some earlier experiments, and thus for a computational model we consider an action-based labeling scheme to be a more pragmatic approach. Social media discourses did not exactly follow the Labovian narrative structure in our empirical case: although complication-oriented discourses seemed to occur during the second phase similarly to the narrative theory, the early phase already involved significant crisis resolution discourses. The dataset for our third phase of crisis should be extended in later research to gain further insights into whether discourses related to some function group might emerge as significant. Further research also needs to investigate if similar patterns of narrativization can be found in different cultural contexts and crises, and whether social media discourses follow their own pattern of narrative structure as compared to Labov's theory (1972). Also, our few-shot classification also needs more work to achieve higher accuracy in action classification. Action classification for social media comments is not an easy task, for example because comments might often involve several actions, and as deciding what action a comment represents sometimes requires interpretation that is hard to define clearly for each case in annotation guidelines. Thus, action classification in this area requires more work.

This research advances the development of the growing line of computational narrative analysis methods, elaborating on the possibilities for using narrative functions to understand the narrativization of crisis discourses. We argue that such tools are needed for supporting other means of research into crisis communication, for a multi-sided understanding of perspectives on crisis and social media engagement. Further, as social media is a site used to influence public opinion and to spread disinformation, the various discursive conflicts taking place in this arena are essential for crisis communicators to both understand and manage.

7. References

Leiming Yan, Yuhui Zheng, and Jie Cao. 2018. Few-shot learning for short text classification. *Multimedia Tools and Applications,* 77(22):29799–29810.

Monika Bednarek, Andrew Ross, Olga Boichak, Y.J. Doran, Georgia Carr, Eduardo Altmann, and Tristram Alexander. 2022. Winning the discursive struggle? The impact of a significant environmental crisis event on dominant climate discourses on Twitter. *Discourse, Context & Media,* 45:100564.

Herbert Clark and Edward F. Schaefer. 1989. Contributing to Discourse. *Cognitive Science*, 13(2):259–294.

Derek Edwards and Jonathan Potter. 1993. Language and causation: A discursive action model of description and attribution. *Psychological review*, 100(1):23–41.

Elizabeth Couper-Kuhlen and Margret Selting. 2017. *Interactional linguistics: Studying language in social interaction.* Cambridge University Press.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020. Task-Aware Representation of Sentences for Generic Text Classification. In: *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3202–3213, Barcelona, Spain. International Committee on Computational Linguistics.

Hélène Joffe. 2003. Risk: From Perception to Social Representation. *British Journal of Social Psychology,* 42(1): 55–73.

Marianne Jørgensen and Louise Phillips. 2002. *Discourse analysis as theory and method.* Sage, London.

Robert Kasperson, Ortwin Renn, Paul Slovic, Halina Brown, Jacque Emel, Robert Goble, Jeanne Kasperson, and Samuel Ratick. 1988. The social amplification of risk: A conceptual framework. *Risk analysis*, 8(2):177–187.

William Labov. 1972. *Language in the Inner City.* Philadelphia: University of Pennsylvania Press.

William Labov and Joshua Waletzky. 1967. Narrative analysis: oral versions of personal experience. In: J. Helms, ed., *Essays in the Verbal and Visual Arts*, pages 12–44. University of Washington Press, Seattle.

George Lakoff and Srini Narayanan. 2010. Toward a computational model of narrative. In: *2010 AAAI Fall Symposium Series*, pages 21–28, Menlo Park, California. https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2323

Nic Newman, Richard Fletcher, Anne Schulz, Simge Andi, Craig Robertson, and Rasmus Nielsen. 2021. *Reuters institute digital news report 2021*. Reuters Institute for the Study of Journalism, Oxford.

Henna Paakki, Heidi Vepsäläinen, and Antti Salovaara. 2021. Disruptive online communication: How asymmetric trolling-like response strategies steer conversation off the track. *Computer Supported Cooperative Work*, 30(3):425–461.

Andrew Piper, Richard So, and David Bamman. 2021. Narrative Theory for Computational Narrative Understanding. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. 10.18653/v1/2021.emnlp-main.26

Emanuel Schegloff. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge; New York: Cambridge University Press.

Timothy Sellnow, Deanna Sellnow, Emily Helsel, Jason Martin and Jason Parker. 2019. Risk and crisis communication narratives in response to rapidly emerging diseases. *Journal of Risk Research*, 22(7):897–908.

Tanya Stivers, Nick Enfield, and Stephen Levinson 2010. Question–Response Sequences in Conversation Across Ten Languages: an Introduction. *Journal of Pragmatics*, 42(10):2615–2619.

Tzvetan Todorov. 1971. The Two Principles of Narrative. *Diacritics*, 1(1):37–44.

Teun A Van Dijk. 1976. Philosophy of action and theory of narrative. *Poetics*, 5(4):287–338.

Jason Turowetz and Douglas Maynard. 2010. Morality in the social interactional and discursive world of everyday life. In: Hitlin S. and Vaisey S., eds., *Handbook of the Sociology of Morality*, pages 503–526, Springer, New York.

Sylvia Walby. 2015. *Crisis*. Polity Press, Cambridge.

Gradnja Korpusa študentskih besedil KOŠ

Tadeja Rozman,* Špela Arhar Holdt‡†

* Fakulteta za upravo, Univerza v Ljubljani
Gosarjeva ulica 5, 1000 Ljubljana
tadeja.rozman@fu.uni-lj.si
‡Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva ulica 2, 1000 Ljubljana
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
spela.arharholdt@ff.uni-lj.si

1 Uvod

Korpusi avtentičnih besedil šolajoče se populacije so tako v svetu kot pri nas pomemben vir informacij o jezikovni zmožnosti oseb, ki v procesu izobraževanja svojo jezikovno zmožnost še razvijajo, hkrati pa so tudi pokazatelj jezikovnih in didaktičnih praks v izobraževalnih okoljih. Ti viri so zato pomembni za jezikovno didaktiko, pripravo k uporabnikom usmerjenih jezikovnih priročnikov in gradiv, kot tudi za razvoj različnih jezikovnotehnoloških orodij. Korpusno jezikoslovje v svetovnem merilu sicer večjo pozornost namenja razvoju in analizi korpusov usvajanja tujih jezikov,[1] v slovenskem prostoru pa imamo po vzoru tovrstnih korpusov zgrajen tudi *Korpus šolskih pisnih izdelkov Šolar* (Rozman et al., 2012) oziroma razširjeno verzijo *Šolar 2.0* (Kosem et al., 2016). Vsebuje besedila, napisana pri pouku v tretjem triletju osnovnih šol in v srednjih šolah, del korpusa pa tudi avtentične učiteljske popravke, ki so s hierarhično zasnovanim sistemom oznak (Arhar Holdt et al., 2018) kategorizirani glede na vrsto jezikovnega problema. Slovenščina je tako eden redkih jezikov, ki ima tovrstne podatke za prvi jezik, a le na omejeni šolski populaciji, zato v okviru projekta *Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti (ARRS, J7-3159)*[2] pripravljamo širitev korpusa s študentskimi besedili, na začetku v obliki pilotnega korpusa študentskih besedil.

2 Namen korpusa

Gradnja Korpusa študentskih besedil KOŠ je v prvi vrsti namenjena pridobivanju empiričnih podatkov o pisni jezikovni zmožnosti študentske populacije, pa tudi analitičnemu vpogledu v procese razvoja strokovnega pisanja. Temeljna jezikovna (normativna, besedilna, pragmatična) znanja naj bi dijaki sicer usvojili že do konca srednje šole, na fakultetah pa naj bi se to znanje nadgrajevalo z usvajanjem terminologije in stilističnih značilnosti strokovnih besedil. Vsaj na nejezikoslovnih študijskih smereh, kjer jezikovna izobrazba ni cilj, ampak je dobro jezikovno znanje le temelj za uspešno profesionalno delovanje, nadaljnje razvijanje jezikovne zmožnosti načeloma poteka hkrati ob usvajanju strokovnega znanja, torej ob recepciji strokovnih del ter s pisanjem npr. seminarskih nalog, esejev, raziskovalnih poročil ter pripravo govornih nastopov, sodelovanjem v strokovnih debatah ipd. Ob tem naj bi študenti uzaveščali procese razumevanja in pisanja, se ukvarjali z razumljivostjo in sprejemljivostjo besedil ter rabo strokovnega besedišča, po potrebi pa tudi odpravljali pravopisne in slovnične pomanjkljivosti. Vendar pedagogi opažamo, da obstajajo velike razlike med jezikovnimi zmožnostmi študentov, profesorji stroke pa se z reševanjem jezikovnih težav lahko ukvarjajo le v omejenem obsegu. Zdi se, da so tudi pristopi pedagogov k ozaveščanju o jezikovnih izbirah različni, ne samo zaradi različnega jezikovnega znanja, ampak tudi pogledov na smiselnost tovrstne povratne informacije, pisnih akademskih praks ipd., pa tudi pomanjkanja didaktičnih usmeritev.

Potreba po razvoju sporazumevalne zmožnosti v slovenskem strokovnem jeziku je bila prepoznana že pri pripravi *Resolucije o Nacionalnem programu za jezikovno politiko 2014–2018*,[3] tedaj določeni jezikovnonačrtovalni cilji jezikovne ureditve visokega šolstva in znanosti pa se v aktualni *Resoluciji o Nacionalnem programu za jezikovno politiko 2021–2025*[4] niso bistveno spremenili. Dokument tako določa, da je na visokošolski strokovni ravni treba omogočati učenje strokovne slovenščine ter na podlagi raziskav in analiz

[1] Več o korpusih usvajanja tujega jezika in gradnji korpusa usvajanja slovenščine kot tujega jezika gl. npr. Stritar Kučuk (2020).
[2] https://www.cjvt.si/prop/
[3] https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina/2013-01-2475?sop=2013-01-2475
[4] https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina/2021-01-1999?sop=2021-01-1999

strokovno-znanstvenega pisanja na visokošolski ravni izdelati učni načrt za strokovno-znanstveno pisanje za uvodni predmet v prvem letniku prvostopenjskih programov. Na podlagi teh določil, zapisanih že v predhodni resoluciji, je bil z namenom pridobivanja empiričnih podatkov o strokovno-znanstvenem pisanju leta 2019 izdelan *Korpus akademske slovenščine KAS*, tj. korpus diplomskih, magistrskih in doktorskih del (Erjavec et al., 2021), objavljenih na Nacionalnem portalu odprte znanosti.[5] V korpusu so torej zbrana strokovna študentska besedila po zaključenih stopnjah visokošolskega in univerzitetnega študija, mentorirana in v veliki meri tudi lektorirana, tako da je s stališča analiz pisne jezikovni zmožnosti študentske populacije in za analizo procesa razvoja strokovnega pisanja le deloma uporaben. Korpus KOŠ bi v perspektivi lahko odpravil vrzel korpusnih podatkov med Šolarjem in KAS-om ter ponudil osnovo za raziskave, katera temeljna znanja je potrebno (bolje) nasloviti na predhodnih stopnjah in katera na terciarni stopnji, kjer se razvoj pisnih jezikovnih zmožnosti nadaljuje na kompleksnejših besediloslovnih ravneh. Širša slika razvojnega loka bi omogočila, da opismenjevanje bolje usmerimo proti končnemu cilju, ki je polnomočno in samostojno (čeprav v skladu s sodobnimi praksami tehnološko in podatkovno podprto) pisanje različnih vrst besedil za različne sporazumevalne namene, kar je pomembno tudi za uspešno poklicno delovanje.

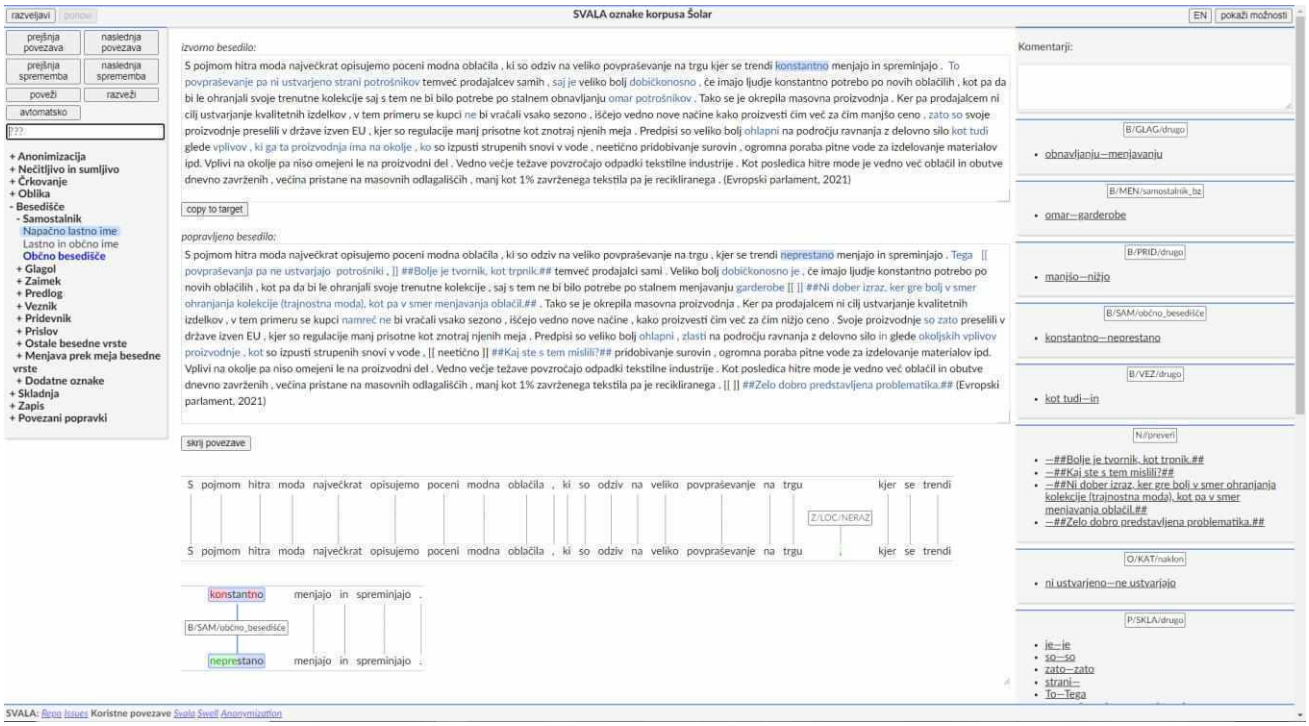
3 Zasnova korpusa

V okviru projekta je predvidena priprava pilotnega korpusa, ki bo objavljen kot podatkovna baza na repozitoriju CLARIN.SI. Gradnja korpusa poteka v študijskem letu 2021/22 in se bo predvidoma končala jeseni 2022, besedila pa bodo zbrana po metodologiji priprave korpusa Šolar, ki vključuje: pravno ureditev odprtega dostopa do rezultatov (priprava in podpis pogodb o prenosu pravic in dovoljenja za uporabo pravic), beleženje vseh relevantnih metapodatkov (program, letnik, področje študija, tip besedila, morebitno večavtorstvo, ob oddanih več verzijah istega besedila tudi oznaka prvotne in spremenjenih verzij), vsaj delna vključitev profesorskih jezikovnih popravkov, zapis v združljivem formatu in strojno označevanje.

Jezikovne popravke bomo v korpus beležili z orodjem *Svala* (Wirén, 2019), ki omogoča pregledno sopostavitev izvornega ter popravljenega besedila, psevdonimizacijo tistih delov besedila, ki bi lahko razkrili avtorstvo ali kake druge občutljive osebne informacije, ter označevanje in vsebinsko kategorizacijo jezikovnih popravkov. Orodje je bilo na projektu *Razvoj slovenščine v digitalnem okolju*[6] prilagojeno za delo s slovenskima korpusoma KOST (Stritar Kučuk, 2020) in Šolar in kot tako podpira označevanje s sistemom oznak korpusa Šolar (Arhar Holdt et al., 2018). Te oznake bomo uporabili tudi za korpus KOŠ (gl. sliko 1), predvideno pa je, da bo za študentska besedila sistem označevanja treba deloma prilagoditi. Pričakovati namreč je (in do sedaj zbrano gradivo to potrjuje), da so zaradi žanrske specifike študentskih besedil, ki jih pregledujejo profesorji nejezikoslovci, popravki redko tudi konkretni predlogi pravilnih jezikovnih izbir, ampak da gre bolj za usmeritve profesorjev, ki v svojih komentarjih študente le na splošno opozarjajo na jezikovne napake in se v večji meri posvečajo stilistiki strokovnih besedil, ustrezni rabi terminologije, citiranju, razumljivosti pisanja, argumentaciji ipd. Vsa korpusna besedila (z označenimi popravki in brez) bomo nato strojno označili na ravneh stavčne segmentacije, tokenizacije, lematizacije, oblikoskladnje, skladnje in imenskih entitet z označevalnikom CLASSLA StanfordNLP (Ljubešić in Dobrovoljc, 2019), ki se v času pisanja povzetka prav tako razvija na omenjenem projektu.

Besedila zbiramo na prvostopenjskih študijskih programih na dveh fakultetah, za vključitev v korpus pa so potencialno relevantna vsa besedila, ki so jih študenti oddali pedagogom v študijskem procesu na fakulteti in niso napisana na roko. Besedila zato zbiramo prek učiteljev, saj bomo tako z večjo gotovostjo prejeli avtentična besedila, ki se realno pišejo v študijskem okolju, predvidoma pretežno seminarske naloge, eseje, poročila, povzetke strokovnih člankov, daljše (esejske) odgovore na vprašanja, morda pa tudi dispozicije in osnutke diplomskih del. Besedila, povezana s pripravo zaključnih del, so s stališča ugotavljanja zmožnosti oblikovanja daljšega strokovnega besedila po končanem izobraževanju zelo dragocena, tudi zaradi vpogleda v mentorske komentarje in popravke, a se trenutno zdi vključitev teh besedil v korpus problematična s stališča anonimizacije, saj so zaključna dela praviloma prosto objavljena na spletu in zlahka povezljiva z osnutki, avtorji in mentorji.

[5] https://openscience.si/
[6] https://www.slovenscina.eu/



Slika1: Preizkus metodologije vpisovanja popravkov v testni različici lokaliziranega programa Svala.

4 Nadaljnji koraki

Na projektu *Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti* želimo zagotoviti pilotni korpus v obsegu 200.000 pojavnic, ob gradnji pa pripraviti tudi oceno prenosljivosti metodologije Šolar na pisno produkcijo študentov in specifikacije nadaljnjega razvoja korpusa študentskega pisanja, tj. opredelitev želenega obsega, strukture glede na regionalno zastopanost, vrsto in področje izobraževanja ter tipologije popravkov. V tem okviru pripravljamo tudi krajši anketni vprašalnik za univerzitetne pedagoge, s katerim želimo pridobiti dodatne podatke o tem, kakšne so prakse podajanja povratnih informacij študentom, ter tako čim učinkoviteje zasnovati zbiranje in beleženje tega gradiva. Do sedaj zbrano gradivo po pričakovanjih nakazuje, da so prakse precej raznolike in da se v mnogočem razlikujejo od podajanja informacij profesorjev slovenščine, ki so zabeležene v korpusu Šolar.

Pod okriljem projekta bomo sicer zbrano korpusno gradivo uporabili za pilotne kvantitativne in kvalitativne jezikoslovne analize študentskega pisanja. Analize se bodo osredotočile na tipične težave pisanja in trende opozarjanja na jezikovno neustrezne ali manj ustrezne ubeseditve, kar vključuje podajanje povratne informacije z vnosom rešitve, opisna priporočila ali grafično nakazovanje mesta težave, kot morebitne druge načine. Rezultate bomo primerjali s frekvenčno urejenimi seznami jezikovnih zadreg v korpusu Šolar. Izsledki bodo predvidoma že nakazali obrise razvoja pisne jezikovne zmožnosti na prehodu iz srednješolskega v univerzitetno izobraževanje, morebitne primanjkljaje temeljnega jezikovnega znanja ter kako je mogoče učni proces z empiričnimi podatki najbolje podpreti.

Zahvala

Projekt *Empirična podlaga za digitalno podprt razvoj pisne jezikovne zmožnosti (J7-3159)* in program *Jezikovni viri in tehnologije za slovenski jezik (P6-0411)* sofinancira Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Špela Arhar Holdt, Polona Lavrič, Rebeka Roblek in Teja Goli. 2018. Kategorizacija učiteljskih popravkov: Smernice za označevanje korpusa Šolar 2.0. V: *1.0. Kazalnik projekta Nadgradnja korpusa Šolar*. https://solar.trojina.si/wp-content/uploads/2022/05/Smernice-za-oznacevanje-korpusa-Solar-2.0-v1.0.pdf

Tomaž Erjavec, Darja Fišer in Nikola Ljubešić. 2021. The KAS corpus of Slovenian academic writing. V: *Lang Resources & Evaluation 55*, 551–583. https://doi.org/10.1007/s10579-020-09506-4

Iztok Kosem, Tadeja Rozman, Špela Arhar Holdt, Polonca Kocjančič in Cyprian Adam Laskowski. 2016. Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov. V: *Zbornik konference Jezikovne tehnologije in digitalna*

humanistika, 95–100. Znanstvena založba Filozofske fakultete, Ljubljana. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Kosem-et-al_Solar-2-0-nadgradnja-korpusa-solskih-pisnih-izdelkov.pdf

Nikola Ljubešić in Kaja Dobrovoljc. 2019. 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, BSNLP@ACL 2019*, pages 29–34. https://aclanthology.org/W19-3704.pdf

Tadeja Rozman, Mojca Stritar Kučuk in Iztok Kosem. 2012. Šolar – korpus šolskih pisnih izdelkov. V: T. Rozman, ur., I. Krapš Vodopivec, M. Stritar, I. Kosem: *Empirični pogled na pouk slovenskega jezika*, 15–35. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Mojca Stritar Kučuk. 2020. Modul Leto plus – prvi korak do korpusa slovenščine kot tujega jezika. V: *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2020*, pages 131–135. Inštitut za novejšo zgodovino, Ljubljana. http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_StritarKucuk_Modul-Leto-plus%e2%80%93prvi-korak-do-korpusa-slovenscine-kot-tujega-jezika.pdf

Mats Wirén, Arild Matsson, Dan Rosén in Elena Volodina. 2019. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. V: *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8–10 October 2018*, pages 227–239. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=159&Article_No=23

Korpusni pristopi za identifikacijo metafore in metonimije: primer metonimije v korpusu g-KOMET

Špela Antloga

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška cesta 46, 2000 Maribor
s.antloga@um.si

Povzetek

Prepoznavanje vrednosti in razširjenosti metaforičnih in metonimičnih izrazov v jeziku je v zadnjih dvajsetih letih vodilo k povečanemu zanimanju za sistematično identifikacijo in luščenje tovrstnih figurativnih izrazov v korpusih posameznih jezikov. Izraze, pri katerih potekajo konceptualne preslikave, ki sodelujejo pri metaforičnih in metonimičnih procesih, je namreč težko izluščiti iz korpusa, ki niso posebej označeni za namene raziskovanja figurativnega jezika. V članku predstavim najpogostejše metode luščenja metaforičnih in metonimičnih izrazov iz jezikovnih korpusov ter na primeru korpusa g-KOMET, ki je ročno označen za metaforične in metonimične izraze v slovenskem govorjenem jeziku, ponazarjam poskus sistematizacije metonimičnih prenosov.

Corpus approaches to metaphor and metonymy identification: The case of metonymy in g-KOMET

Recognizing te value of metaphorical anod metonymic expressions in language has in the last two decades led to increased interest in the systematic identification and extraction of figurative expressions in various language corpora. Expressions in which conceptual mappings that participate in metaphorical and metonymic processes take place are difficult to extract from a corpus that is not specifically annotated for the purposes of figurative language research. We describe prevailing methods of searching for metaphorical and metonymic expressions in language corpora. Using the manually annotated corpus for metaphorical and metonymic expressions in the Slovene spoken language g-KOMET, we try to systemize some of the prevailing annotated metonymical mappings.

1. Uvod

Jezik in mišljenje sta tesno povezana. Naše mišljenje je tako zapleteno, da z jezikom nismo vedno zmožni vsega »neposredno« izraziti, zato za razlago sveta uporabljamo različne jezikovno-kognitivne postopke, med drugim metafore in metonimije. Korpusnih raziskav metafore in metonimije ter tudi drugih oblik figurativnega jezika v slovenščini je malo. Čeprav so v zadnjem desetletju korpusne metode raziskovanja slovenščine postale uveljavljena empirična paradigma v jezikoslovju predvsem na področjih, povezanih z leksikologijo in slovnico ter jezikovno rabo, področje figurativnega jezika, ki je sicer na teoretski ravni dobilo zagon z razmahom teorije konceptualne metafore in metonimije (Lakoff in Johnson, 1980; Lakoff in Turner, 1989; Lakoff, 1993), pri tem trendu nekoliko zaostaja (Bedkowska-Kopczyk, 2016; Antloga, 2020c). Eden od možnih razlogov je pomanjkanje enotne in uspešne metode za sistematično identifikacijo metaforičnih in metonimičnih izrazov v že obstoječih korpusih, ki niso posebej označeni za konceptualne preslikave. Posledično so se za sistematično analizo konceptualnih struktur v jeziku jezikoslovci zatekli k izgradnji korpusov z označenimi potencialnimi metaforičnimi in metonimičnimi izrazi, ki pa so časovno zamudni in zahtevajo veliko prilagoditev označevalnih shem ciljnemu jeziku raziskovanja.

V prispevku bodo opisane različne bolj ali manj uveljavljene metode identifikacije metaforičnih in metonimičnih izrazov v obstoječih (splošnih) korpusih besedil z vsemi prednostmi in slabostmi. Kot eden od virov za sistematično analizo metaforičnih in metonimičnih izrazov v slovenskem govorjenem jeziku bo predstavljen korpus G-KOMET, ki je nastal v okviru razpisa CLARIN 2021. Na primeru korpusa g-KOMET bo predstavljen poskus sistematizacije in klasifikacije najpogostejših označenih metonimičnih prenosov v slovenskem govorjenem jeziku.

2. Opredelitev metafore in metonimije v kognitivnem jezikoslovju

Ena od ključnih ugotovitev sodobnega pogleda na metaforo in metonimijo je, da metafor in metonimij ne uporabljamo zgolj za jezikovno sporazumevanje, temveč da v metaforah in metonimijah tudi mislimo. V tem duhu konceptualno teorijo metafore zanimajo zlasti načini mentalne organizacije konceptov, s pomočjo katerih človek osmišlja stvarnost, ki ga obdaja, in družbo, v kateri živi (Bratož, 2010). Za jezikoslovce so bila tovrstna vprašanja sprva svojevrsten izziv, saj zahtevajo pogled preko meja področja jezikoslovja na druge discipline, kot so psihologija, nevroznanost, filozofija in druge vede, ter s tem predpostavljajo interdisciplinaren način dela. Konec sedemdesetih let prejšnjega stoletja se je tako zgodil t. i. kognitivni preobrat, ki je metaforo in metonimijo iz jezikovne ravni prenesel na konceptualno, miselno raven. Metaforo in metonimijo so začeli obravnavati kot konceptualni mehanizem, s pomočjo katerega se védenje o konkretnih pojavih in izkušnjah projicira na številne abstraktne domene. Na primer čas običajno konceptualiziramo kot prostor, čustva kot naravne sile, organizacije kot organizme ali stroje (Bratož, prav tam).

2.1. Metafora

Po sodobni definiciji metafore torej niso samo jezikovni izraz, ampak v njih tudi razmišljamo. Med različnimi teoretičnimi pristopi, ki so se posvečali preučevanju metafore, je danes pri raziskovanju metafore in metaforičnostih izrazov ena najvidnejših teorija konceptualne metafore, ki so jo razvili George Lakoff in njegovi sodelavci (Lakoff in Johnson, 1980; Lakoff in Turner, 1989). Po omenjenem teoretičnem modelu so metafore bistven element človekovega spoznavanja in sredstvo, ki nam omogoča, da razumemo in doživljamo eno izkušenjsko področje ali domeno (*domain*) s pomočjo (v okviru) drugega. Prenos poteka s t. i. medpodročnimi preslikavami (*cross-domain mappings*) med izhodiščnim področjem (*source domain*), ki je običajno konkretnejše, in ciljnim področjem (*target domain*), ki je bolj abstraktno.

2.2. Metonimija

Tradicionalna retorika je metonimijo obravnavala predvsem kot retorično figuro, torej je o njej razmišljala kot o jezikovnem pojavu, kot o predmetu figurativnega jezika (Radden in Kövecses, 1999). Tudi Aristotel ni povsem prepoznal značilnosti metonimije in jo je pojmoval kot podtip metafore (Bernjak in Fabčič, 2018). Podobno definicijo metonimije zasledimo tudi v sodobnih slovarjih, npr. v Slovarju slovenskega knjižnega jezika.[1] Jakobson (1956) je poudaril inherentnost metonimije v jeziku in izpostavil pojem bližine kot temeljni princip metonimije. Kognitivni jezikoslovci se opirajo na ta in podobna stališča in razširijo fenomen metonimije na pojmovno-pomenski mehanizem, ki omogoča strukturiranje jezika ter mišljenja, torej deluje kot centralno sredstvo v procesu konceptualizacije. Lakoff in Johnson (1980: 46–52) metonimijo definirata na ravni konceptualizacije kot pojmovno operacijo ali kognitivni proces, v katerem eno, izhodiščno entiteto uporabimo zato, da nam omogoča mentalni dostop do druge, ciljne entitete znotraj določene pojmovne domene. Torej metonimijo obravnavata kot pojmovno-pomenski mehanizem, ki strukturira ne samo jezik, ampak tudi naše mišljenje. Če pri metafori prihaja do preslikave z enega konceptualnega področja na drugo, metonimija vključuje samo eno domeno, saj do preslikave med dvema elementoma prihaja v okviru ene same domene. Lakoff in Johnson (1980) poudarjata, da je tako kot metafora tudi metonimija konceptualne narave in da gre za fenomen, ki igra osrednjo vlogo pri strukturiranju našega védenja o svetu. Kövecses (2002) pravi, da je metonimija kognitivni proces, v katerem do določene konceptualne entitete (cilja) pridemo s pomočjo druge konceptualne entitete (sredstva). Z drugimi besedami, ena konceptualna entiteta je referenčna točka, ki omogoča mentalni dostop do druge konceptualne entitete.

Bolj shematično primerjavo konceptualne metafore in metonimije predstavlja Tabela 1.

	metafora	metonimija
funkcija konceptualnega razmerja	sklepanje na podlagi podobnosti	referencialnost
narava konceptualnega razmerja	podobnost	(logična) povezava

Tabela 1: Razlikovanje med metaforo in metonimijo. Povzeto po Feyaerts (2012).

3. Metode luščenja metaforičnih (in metonimičnih) izrazov v korpusih

V povezavi z metaforo in metonimijo sta zaradi pomanjkanja ustrezne metodologije problematična predvsem (sistematična) identifikacija in luščenje ustreznih podatkov iz splošnega jezikovnega korpusa. Konceptualne preslikave, ki sodelujejo pri metaforičnih in metonimičnih procesih, namreč niso neposredno povezane s posameznimi jezikovnimi oblikami in jih je težko izluščiti iz korpusa, ki niso posebej označeni za namene raziskovanja figurativnega jezika. S kombinacijo avtomatskega in ročnega luščenja podatkov iz splošnih korpusov so se v drugih jezikih izoblikovale naslednje metode identifikacije metaforičnih (in metonimičnih) izrazov (Stefanowitsch, 2006):

Ročno luščenje metaforičnih besed iz korpusa se je uveljavilo zaradi potrebe po (bolj) sistematični analizi konceptualne metafore in metonimije, tako da je branju besedila v korpusu sledilo sistematično izpisovanje metaforičnih in metonimičnih izrazov (Semino in Masci, 1996). Seveda je bilo delo zamudno in obsegovno omejeno, predvsem pa neizkoriščeno z vidika količine podatkov v korpusu, a vsekakor bolj sistematično kot zanašanje na sporadične primere ali primere, ki niso izhajali iz dejanske jezikovne rabe. Kljub temu so kognitivistom očitali subjektivnost, neempiričnost in nekonsistentnost pri prepoznavanju (iskanju) in razlagi konceptualnih metafor in metonimij (npr. Tummers et al., 2005; Wasow in Arnold, 2005).

Metaforični in metonimični izrazi so v izhodiščni domeni preslikave vedno povezani z neprenesenimi (nefigurativnimi) leksikalnimi enotami. Zato je bila kot odziv na kritike naslednja stopnja korpusnega pristopa k figurativnemu jeziku **iskanje izhodiščne domene po ključnih besedah** oziroma identifikacija metafor na podlagi potencialnih izhodiščnih domen (pomensko polje, za katerega se predpostavlja oziroma je bilo že ugotovljeno, da sodeluje pri metaforičnih preslikavah, kot so na primer *srce*, *ogenj*, *boj*, *potovanje* ipd.). Iskanje lahko poteka preko posameznih besed v konceptualni strukturi ali preko skupine besed, ki so pomensko povezane (na primer *ogenj*, *plamen*, *vročina*, *pogoreti*, *zgoreti*, *plamteti*, *vzplamteti* ipd.). Z ročnim pregledovanjem rezultatov je bila določena potencialna metaforičnost izraza in nato ciljna domena metaforične preslikave (npr. LJUBEZEN, JEZA ipd.). Postopoma so se začeli izoblikovati seznami ključnih besed izhodiščnih domen za identifikacijo metafor v posameznih jezikih. Jezikoslovci so nato na podlagi seznamov raziskovali metafore v različnih jezikih, kontekstih in diskurzih (Hanks, 2004; Koller, 2006).

Postopna uveljavitev identifikacije metaforičnih in metonimičnih izrazov v korpusih z iskanjem po ključnih

[1] »metonimija-e ž lit. besedna figura, za katero je značilno poimenovanje določenega pojma z izrazom za kak drug predmetno, količinsko povezan pojem«.

besedah izhodiščne domene je vodila k zanimanju za raziskovanje figurativnega jezika v konkretnejših, bolj specifičnih domenah, npr. v političnem diskurzu, v ekonomiji, športu ipd. V teh primerih pristop, usmerjen v izhodiščno domeno, ni bil učinkovit, saj bi zahteval predhodno poznavanje vira preslikave (izhodiščne domene), ki bi lahko bil potencialno najden v ciljni domeni. Zato se je uveljavila metoda **iskanja ciljne domene s seznamom ključnih besed izhodiščnih domen.** Za učinkovito identifikacijo metaforičnih in metonimičnih izrazov s ključnimi besedami ciljne domene je potrebna velika količina reprezentativnih in enotematskih besedil, ki so povezana z iskano ciljno domeno. To je relativno enostavno pri »konkretnih« ciljnih domenah, kot so zgoraj naštete POLITIKA, EKONOMIJA, ŠPORT, težje pa bi bilo iskanje metaforičnih in metonimičnih izrazov s ciljnimi domenami, kot so na primer ČUSTVOVANJE, UMSKA AKTIVNOST, ZAZNAVANJE ipd. (nekaj rešitev ponuja Tissari, 2003). Drugi problem, povezan s tovrstno identifikacijo metafor v korpusu, pa je, da bi identificirali le tiste izhodiščne domene, ki so povezane z izrazi, katerih pogostnost je v ciljni domeni tako visoka, da so se uvrstili na seznam ključnih besed ciljnih domen. Analiza metaforičnih prenosov torej ne bo celovita in sistematična.

Z združitvijo obeh predhodno navedenih metod se je uveljavila metoda **iskanja stavkov, ki vsebujejo ključne besede tako izhodiščne kot ciljne domene**, predvsem v obliki avtomatskega luščenja metaforičnih izrazov. Kljub temu metoda še vedno zahteva poglobljen ročni pregled izluščenih podatkov zaradi možnih enakopisnic ali neprenesenega pomena obeh izrazov v stavku. Problem je tudi, da je za tako iskanje potreben zelo izčrpen seznam besed z obeh domen, saj je sicer iskanje nepopolno. Poleg tega je ta metoda bolj uporabna za raziskovanje že poznanih konceptualnih struktur, metafor in metonimij, manj pa za sistematično identifikacijo (novih oziroma vseh) konceptualnih struktur.

Nekaj poskusov identifikacije metaforičnih izrazov je potekalo tudi s t. i. **kazalniki metaforičnosti**, to so metajezikovni izrazi, ki napovedujejo oziroma signalizirajo metaforično rabo. Goatly (1997) kot metaforične signalizatorje navaja izraze, kot so *metaphorically/figuratively speaking* (metaforično/ figurativno rečeno, v prenesenem pomenu), *so to speak* (tako rekoč/če tako rečem), intenzifikatorje *literally* (dobesedno), *actually* (pravzaprav) ali celo ortografska znamenja, kot so narekovaji, poševni tisk ipd. S to metodo lahko sicer izluščimo relativno malo metaforičnih izrazov, vendar lahko po drugi strani opazujemo jezikovne okoliščine, ko je metaforična raba v besedilu namerno (ali nenamerno) eksplicitno signalizirana (Skorczynska in Ahrens, 2015).

Ena od zadnjih uveljavljenih metod je **iskanje po korpusu, označenem s konceptualnimi preslikavami**. Prvi korpus, označen s konceptualnimi preslikavami v obliki indirektne, direktne in implicitne metaforične besede[2] v štirih besedilnih tipih (časopisna besedila, strokovna besedila, literarna besedila in konverzacijska besedila) za angleški jezik[3] je leta 2012 razvila skupina raziskovalcev, ki se je poimenovala Praglejazz. Ob tem je razvila postopek za ugotavljanje metaforičnih besed v besedilu, poimenovan MIPVU (Steen et al., 2010), da bi omogočila objektivnejšo, natančnejšo in bolj sistematično (jezikoslovno) analizo metaforičnih izrazov v različnih besedilih. Temeljno izhodišče za označevanje metaforičnih besed pri tem postopku je ugotavljanje razmerja med osnovnim in kontekstualnim pomenom besede. Pri tem je treba za vsako leksikalno enoto ugotoviti, ali se njen konkretni kontekstualni pomen razlikuje od njenega osnovnega pomena. Postopek je s prilagoditvami značilnostim posameznih jezikov sprožil zanimanje za identifikacijo metaforičnih izrazov in metafor v češčini (Pavlas et al., 2018), litovščini (Urbonaitė, 2016), madžarščini (Babarzy in Bencze, 2010), poljščini (Risinski in Mahula, 2015), srbščini (Bogetić, 2019) ter za izdelavo korpusov metafor v ruščini (Badryzlova in Lyashevskaya, 2017), hrvaščini (Despot et al., 2019) in kitajščini (Lu in Wang, 2017). Eden od poskusov oblikovanja korpusa metafor v slovenščini, ki bi omogočal jezikoslovno analizo metaforičnih izrazov in metafor v različnih besedilih ter ponujal možnost za prepoznavanje kulturnospecifičnega pomena metafor, je korpus metafor KOMET 1.0 (Antloga, 2020a) in njegovo nadaljevanje z dodanimi transkripcijami govorjenega jezika korpusa g-KOMET (Antloga in Donaj, 2022).

4. Korpus g-KOMET

Korpus g-KOMET[4] (korpus metaforičnih in metonimičnih izrazov v govorjenem jeziku) je nadgradnja pisnega korpusa metaforičnih izrazov in metafor KOMET 1.0 s transkripcijami (po)govora v obsegu 52.529 besed. Nadgradnja vključuje tudi definiranje in ročno dodajanje novih oznak v primerjavi s korpusom KOMET 1.0, in sicer oznak za idiome in metonimije. Besedilo za korpus je bilo izluščeno iz korpusa GOS. Glede na želeno velikost našega korpusa smo iz vsake datoteke korpusa GOS izbrali 5 % besedila. Pri tem smo naključno izbrali začetno izjavo[5] govora in dodajali zaporedne izjave govora, dokler nismo dosegli želene velikosti. Če smo velikost dosegli sredi izjave, smo dodali tudi vse preostale besede v njej. S tem smo dosegli končno velikost korpusa 52.529 besed z enako uravnovešenostjo besedila, kot je prisotna v korpusu GOS. Korpus torej vključuje uravnotežen nabor transkripcij informativnega, izobraževalnega, razvedrilnega, zasebnega (telefonski pogovor, osebni stik) in nezasebnega (telefonski pogovor, osebni stik) diskurza. Če je bila beseda zapisana tako v pogovorni kot normalizirani obliki, smo prevzeli normalizirano obliko. Pri tem se nekatere pogovorne besede zapišejo kot dve besedi v normalizirani obliki, npr. »nemo« v »ne bomo«. Pri izluščanju besedila smo odstranili časovne oznake in oznake za menjavo govornih

[2] Ne gre za označevanje metafor, ampak besed, ki se potencialno lahko realizirajo kot metafore.

[3] Gl. http://www.vismet.org/metcor/search/showPage.php?page=start.

[4] Projekt izdelave korpusa je bil financiran v okviru projekta CLARIN.si 2021. Korpus je dostopen na naslovu http://hdl.handle.net/11356/1293.

[5,6] Izjavo in govorno vlogo razumemo, kot sta opredeljeni v specifikacijah za transkribiranje GOS, gl. Zwitter Vitez et al., 2009.

vlog,[6] saj začetki in konci izluščenega dela besedila niso hkrati začetki in konci govornih vlog. Ohranili pa smo druge oznake, npr. smeh, hrup, in prekinjene besede oz. napačne začetke. Za označevanje je bilo uporabljeno orodje Q-CAT (Brank, 2019).

4.1. Označevanje metaforičnih besed (oznake MRWi, MRWd, MFlag in WIDLI)

Označevanje metaforičnih besed je temeljilo na postopku za identifikacijo metafor MIPVU (Steen et al., 2010),[7] ki omogoča sistematično identifikacijo jezikovne metafore. Identificirani so bili jezikovni izrazi, ki imajo potencial, da jih ljudje realiziramo kot metafore. Za vsako leksikalno enoto v besedilu je bil določen njen osnovni pomen (po SSKJ) in njen pomen v kontekstu. Če se je kontekstualni pomen razlikoval od osnovnega pomena te besede, je bila beseda označena kot metaforična beseda (MRW). Označenim metaforičnim besedam je bila nato pripisana informacija o tem, ali gre za (1) indirektno metaforo (MRWi), (2) direktno metaforo (MRWd) ali (3) mejni primer (WIDLI). Označeni so bili tudi (4) metaforični signalizatorji (MFlag).[8] Korpus je označevala ena oseba.

4.2. Označevanje stalnih besednih zvez (oznaka idiom)

Označene so bile večbesedne enote, katerih pomen je različen od pomena posameznih sestavin večbesedne enote. Vsaj ena sestavina v označeni stalni besedni zvezi je bila torej rabljena metaforično.

4.3. Uvrščanje v pomensko polje metaforičnega prenosa (oznaka *frame*)

Označeni metaforični izrazi in stalne besedne zveze so bili uvrščeni v pomenska polja, ki funkcionirajo kot sistem kategorij, ki so strukturirane glede na določen kontekst, ki jih motivira. Pomensko polje omogoča, da znotraj določene pomenske kategorije (npr. naravni pojavi, čas, prostorska orientacija, družina, premikanje itd.) poiščemo metaforične izraze, ki so lahko potencialno uresničitev neke konceptualne strukture. V korpusu g-KOMET je bilo označenim metaforičnim besedam in stalnim besednim zvezam določenih 65 pomenskih polj.

4.4. Označevanje metonimij

Če se pri metaforah dogaja preslikava z enega izkušenjskega področja na neko drugo izkušenjsko področje, se pri metonimijah preslikava dogaja znotraj enega področja, pri čemer ugotavljamo razmerje med obema entitetama preslikave. Ugotovljenim metonimičnim izrazom je bilo določenih 45 tipov metonimične preslikave.

Označeni elementi	Število označenih besed (odstotek); ∑ = 52.529 besed
metaforične besede	728 (1,38 %)
idiomi	256 (0,49 %)
metonimije	744 (1,42 %)
pomenska polja	65

Tabela 2: Označeni figurativni elementi v korpusu g-KOMET.

5. Analiza in klasifikacija označenih metonimičnih izrazov v korpusu g-KOMET

Čeprav sta bili od samih začetkov kognitivnega jezikoslovja predmet zanimanja kognitivne semantike tako metafora kot metonimija, je bila pozornost vseskozi usmerjena zlasti na metaforo. Še danes je raziskovanje metonimije v primerjavi z metaforo zelo marginalno, čeprav številni jezikoslovci prepoznavajo ključni pomen metonimije v vsakdanjem jeziku in poudarjajo raznovrstne metonimične relacije kot načine organizacije konceptualne strukture (Bratož, 2010). V korpusu g-KOMET je bilo označenih 744 metonimičnih izrazov, ki jim je bila dodana ena od 54 oznak za različne metonimične prenose.

Tip metonimičnega prenosa	Odstotek glede na vse označene metonimične izraze v korpusu g-KOMET
splošno za specifično	16,8 %
institucija za osebo (skupino)	9,7 %
del za celoto	7,1 %
rezultat dejanja za dejanje	6,4 %
ime za delo	6,3 %
lastnost za osebo	6 %
smer za cilj	5,6 %
celota za del	3,6 %
predmet za aktivnost	3,6 %
kraj za osebo (skupino)	3,5 %
last za aktivnost	2,1 %
del telesa za osebo (skupino)	1,6 %
sredstvo dejanja za rezultat dejanja	1,3 %
ideologija za osebo (skupino)	1,3 %
dejanje za rezultat dejanja	1,2 %
stavba za institucijo	1,2 %
podjetje za delavca (skupino)	1,2 %
kraj za dogodek	1,2 %

Tabela 3: Najpogostejši označeni metonimični prenosi v odstotkih glede na vse označene metonimične izraze v korpusu g-KOMET.

Namesto tradicionalne opredelitve tipov metonimije glede na metonimični prenos (gl. zgoraj) navajam še alternativno, vsebinsko delitev metonimije, kot izhaja iz

[7] Metaphor Identification Procedure Vrije Universiteit (MIPVU).

[8] Za podrobnejšo razlago metodoloških izhodišč za definiranje označevalne sheme glej Antloga 2020a.

označenega korpusa g-KOMET. Delitev izhaja iz predpostavke, da lahko metonimije kategoriziramo glede na vrsto pojmovne vsebine, do katere se dostopa preko metonimije. Konceptualna metonimija je tako klasificirana glede na to, katero konceptualno vsebino aktivira v metonimičnem prenosu. Navedeni so zgolj tipi metonimije, ki so najpogostejši v korpusu g-KOMET. Za smiselne zaključke o vlogi metonimije v govorjenem jeziku/konverzaciji bi bila nujna primerjava z zastopanostjo in vlogo metonimije tudi v negovorjenih besedilih.

5.1. Metonimija STVAR ZA X

Metonimije STVAR ZA X so metonimije, katerih cilj (predvideni referent) je STVAR, do katere se dostopa s pomočjo referenčne vsebine, ki je z njo povezana v istem idealiziranem kognitivnem modelu. Metonimije STVAR ZA X lahko razdelimo v podkategorije glede na konceptualno izhodišče metonimičnega prenosa:

STVAR ZA STVAR
Metonimični prenos omogoča neposredni mentalni dostop do stvari preko neke druge stvari ali njene vloge ali funkcije v situaciji, ki jo ta stvar opravlja.

(…), da *kozica* vre 20 do 25 minut (…) →
POSODA (kozica) NAMESTO VSEBINE (vode v kozici)

STVAR ZA ČLOVEKA (SKUPINO)
(…) so samo še *bobni* igrali (…)
INŠTRUMENT NAMESTO GLASBENIKA, KI IGRA TA INŠTRUMENT

STVAR ZA LASTNOST
(…) vidi *mercedesa* ko se pogleda v ogledalo (…)
AVTO NAMESTO VRLINE/POMANJKLJIVOSTI

STVAR ZA DODODEK
(…) na *rdeči preprogi* (…) znova zablestela (…)
SVAR NA DOGODKU NAMESTO CELOTNEGA DOGODKA

5.2. Metonimija LASTNOST ZA X

Pri metonimijah LASTNOST ZA X je za cilj (predvideni referent je LASTNOST) prenosa pomembno, da je posameznik ali skupina znotraj kategorije »idealnih članov« te kategorije, kar je pogojeno z bližino posameznika ali skupine idealu, ki ga postavlja standardni referent, npr. stereotipna lastnost (ki nadomesti preostale lastnosti), vidna lastnost (ki nadomesti čustvene lastnosti) ipd. Glede na konceptualno izhodišče metonimičnega prenosa v korpusu g-KOMET jih lahko razdelimo v dve skupini:

LASTNOST ZA SKUPINO
(…) taka mesta ki so (…) pa *črni* so tam (…)
ČRNA BARVA NAMESTO TEMNOPOLTIH LJUDI

LASTNOST ZA OSEBO
(…) *najlepša* danes (..)
IZGLED OSEBE NAMESTO OSEBE

5.3. Metonimija OSEBA ZA X

Metonimije OSEBA ZA X so pogoste metonimije, pri katerih prihaja do prenosa človekove dejavnosti, rezultatov dejavnosti, prostora dejavnosti ipd. na osebo, ki opravlja to dejavnost. Razdelimo jih lahko v naslednje podkategorije:

OSEBA ZA AKTIVNOST
(…) zadnjič gledal *nogometaše* (…)
OSEBA, VKLJUČENA V AKTIVNOST, NAMESTO AKTIVNOSTI

OSEBA ZA TEORIJO
(…) vsi citirajo *Žižka* (…)
PREDSTAVNIK TEORETIČNEGA PRISTOPA NAMESTO IZHODIŠČ TEGA PRISTOPA

OSEBA ZA LOKACIJO
(…) pa pri *zdravniku* sto let čakala (…)
OSEBA, KI OPRAVLJA DEJAVNOST, NAMESTO PROSTORA, KJER SE OPRAVLJA DEJAVNOST

5.4. Metonimija LOKACIJA ZA X

Pri metonimijah LOKACIJA ZA X je LOKACIJA uporabljena za priklic ene ali več entitet, ki so na tej lokaciji. Ker sta lokacija in to, kar se nahaja na lokaciji, v nekakšni prostorski relaciji, bi lahko tovrstne metonimije opredelili tudi kot DEL ZA CELOTO. Metonimije LOKACIJA ZA X lahko razdelimo v podkategorije:

LOKACIJA ZA DOGODEK
(…) to mi je ostalo od *Otočca* (…)
KRAJ, KJER JE POTEKAL DOGODEK, NAMESTO DOGODKA

LOKACIJA ZA INSTITUCIJO
(…) se zmenijo na *Čufarjevi* (…)
IME ULICE NAMESTO STAVBE NA TEJ ULICI

LOKACIJA ZA STVAR
(…) da sem kar *McDonald's* prinesla domov (…)
RESTAVRACIJA NAMESTO JEDI V RESTACRACIJI

LOKACIJA ZA OSEBO (SKUPINO)
(…) *gostilna* pa vse čisto tiho (…)
PROSTOR, KJER SE ZADRUŽUJE OSEBA (SKUPINA), NAMESTO OSEBE (SKUPINE) V TEM PROSOTRU

Metonimije lahko opazujemo tudi glede na vidik, ki določa izhodišče/sredstvo (vehikel) metonimičnega prenosa. Pogled izhaja iz predpostavke kognitivnega jezikoslovja, da ima konceptualna metonimija izkustvene in spoznavne temelje, njene jezikovne uresničitve pa so samo ena od možnih oblik, skozi katere se izraža. Zato kognitivizem uporabi pojem *idealiziranih kognitivnih modelov* (IKM), ki predstavljajo abstrakcijo človekovih izkustev. Delujejo kot abstrahirane sheme, ki delno zajemajo naše vedenje o svetu. Za kognitivne pristope je primarno vprašanje, zakaj izberemo prav določeno konceptualno entiteto za metonimični izraz, in ne neke druge. Na tej podlagi (razširjeno po Radden in Kövecses, 1999) lahko označene metonimične izraze opazujemo tudi:

- z vidika povezave med pogostnostjo metonimičnega prenosa in **človekovim izkustvom** (npr. metonimični prenosi v korpusu g-KOMET *splošno za specifično* (125) : *specifično za splošno* (3); *konkretno za abstraktno* (7) : *abstraktno za konkretno* (3); *definirano za*

nedefinirano (2) : *nedefinirano za definirano* (0)). Zaradi lažjega razumevanja je bolj verjetno, da bodo metonimični prenosi potekali s splošnega na specifično, s konkretnega na abstraktno ipd. Povezanost z visoko pogostnostjo označenih tovrstnih metonimičnih prenosov v korpusu je ena od bistvenih (najpogostejših) funkcij metonimije, tj. referencialna funkcija, ki je nekakšna bližnjica za označevanje kompleksnega in abstraktnega pojava z enostavnejšim, konkretnejšim in razumljivejšim pojavom (izrazom);

- z vidika povezave med pogostnostjo metonimičnega prenosa in **kulturno preferenco** (v korpusu g-KOMET lahko opazujemo različne kulturnospecifične metonimične prenose *lastnost za osebo* (45), *lastnost za stvar* (9), *lastnost za institucijo* (1), *posameznik za skupino* (4), *ideologija za človeka (skupino)* (11), *ustanova za človeka (skupino)* (72) ipd.). Te pojmovne sheme združujejo posamezne elemente, povezane z našim kulturnospecifičnim védenjem o svetu, družbi, konvencijah in običajih. V konkretni jezikovni situaciji pogosto kontekst in izkustvo določata, kateri segment enciklopedičnega védenja se bo profiliral kot pomemben in se jezikovno realiziral.

6. Nadaljnje delo

Identifikacija in analiza metonimičnih in metaforičnih izrazov s korpusnega vidika imata v slovenščini pred sabo še dolgo pot. Čeprav so nekatere metaforične preslikave in metonimični prenosi univerzalni oziroma prisotni v več jezikih, je unikatna njihova frekvenca pojavljanja v posameznih jezikih, njihova realizacija in vpetost v kulturnospecifične elemente jezikovnega prostora. Za nadaljnjo analizo metaforičnih izrazov v slovenskem jeziku bo zanimiva primerjava korpusa KOMET 1.0, v katerem so označene metaforične besede v zapisanem jeziku, in korpusa g-KOMET, ki vsebuje govorjena besedila v obliki transkripcij. Ker so bile v korpus g-KOMET dodane tudi oznake za metonimične prenose, je eden od naslednjih ciljev tudi sistematična analiza metonimije v govorjenem jeziku.

7. Literatura

Špela Antloga. 2020a. *Korpus metafor KOMET 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1293.

Špela Antloga. 2020b. Korpus metafor KOMET 1.0. V: *Jezikovne tehnologije in digitalna humanistika [elektronski vir]: zbornik konference: 24.–25. september 2020*, str. 176–170. Ljubljana: Inštitut za novejšo zgodovino.

Špela Antloga. 2020c. Vloga metafor in metaforičnih izrazov v medijskem diskurzu: analiza konceptualizacije boja. V: J. Vogel, ur., *Slovenščina – diskurzi, zvrsti in jeziki med identiteto in funkcijo*, str. 27–34. Ljubljana: Znanstvena založba Filozofske fakultete.

Špela Antloga in Gregor Donaj. 2022. *Korpus g-KOMET*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1490.

Yulia Badryzlova in Olga Lyashevskaya. 2017. Metaphor Shifts in Constructions: the Russian Metaphor Corpus. V: *Computational construction grammar and natural language understanding: Papers from the 2017 AAAI Spring Symposium*. The AAAI Press.

Anna Babarczy in Idiko Bencze. 2010. The automatic identification of conceptual metaphors in Hungarian texts: A corpus-based analysis. V: *LREC 2010 Workshop on Methods for the Automatic Acquisition of Language Resources: Proceedings*, str. 31–36.

Elizabeta Bernjak in Melanija Fabčič. 2018. Metonimija kot konceptualni in jezikovni pomen. *Anali PAZU HD* 4/1-2: 11–23. Združenje Pomurska akademsko znanstvena unija.

Agnieszka Bedkowska-Kopczyk. 2016. Začutiti in občutiti: kognitivna analiza pomensko-skladenjskih lastnosti dveh predponskih tvorjenk iz glagola čutiti. V: E. Kržišnik in M. Hladnik, ur., *Toporišičeva obdobja*, str. 41–48. Ljubljana: Znanstvena založba Filozofske fakultete.

Ksnenija Bogetić. 2019. Linguistic metaphor identification in Serbian. V: S. Nacey in T. Krennmayr, ur., *MIPVU in Multiple Languages*, str. 203–226. Amsterdam: John Benjamins.

Janez Brank. 2019. *Q-CAT Corpus Annotation Tool*. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1262.

Silva Bratož. 2010. Metafore našega časa. Fakulteta za management, Koper.

Janez Brank. 2019. Q-CAT Corpus Annotation Tool, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1262.

Kristina Despot, Mirjana Tonković, Mario Brdar, Benedikt Perak, Ana Ostroški Anić, Bruno Nahod in Ivan Pandžić. 2019. MetaNet.HR: Croatian Metaphor Repository. V: *Metaphor and Metonymy in the Digital Age. Theory and Methods for Building Repositories of Figurative Language*, str. 123–146. Amsterdam: John Benjamins.

Kurt Feyaerts. 2012. Refining the Inheritance Hypothesis: Interaction between metaphoric and metonymie hierarchies. V: A. Barcelona, ur., *Metaphor and Metonymy at the Crossroads: A Cognitive Perspective*, str. 59–78. Berlin: De Gruyter Mouton.

Raymond W. Gibbs. 1999. Researching Metaphor. V: *Researching and applying metaphor*, str. 29–47. Cambridge: Cambridge University Press.

Stefan Gries in Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9/1: 97–129.

Adrew Goatly. 1997. *The Language of Metaphors*. London & New York: Routledge.

Patrick Hanks. 2004. The syntamatics of metaphor and idiom. *International Journal of Lexicography* 17/3: 245–274.

Roman Jakobson. 1956. The Metaphoric and Metonymic Poles. V: *Metaphor and Metonymy in Comparison and Contrast*, str. 41–47. Berlin/New York: Mouton de Gruyter.

Veronika Koller. 2006. Of critical importance: Using electronic text corpora to study metaphor in business media discourse. V: A. Stefanowitsch in S. Gries, ur.,

Corpus-Based Approaches to Metaphor and Metonymy, str. 237–266. Berlin: De Gruyter Mouton.

Zoltan Kövecses. 2002. *Metaphor: A practical Introduction*. Oxford/New York: Oxford University Press.

George Lakoff. 1993. The contemporary theory of metaphor. V: Andrew Ortony, ur., *Metaphor and thought*, str. 202–251. Cambridge: Cambridge University Press.

George Lakoff in Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

George Lakoff in Mark Turner. 1989. *More than Cool Reason: A Field Guide to Poetic Metaphor*. The University of Chicago Press.

Xiaofei Lu in Ben Pin Yun Wang. 2017. Towards a metaphor-annotated corpus of Mandarin Chinese. *Language Resources and Evaluation* 51/3: 663–694.

Klaus-Uwe Panther in Günter Radden. 1999. The potentiality for actuality metonymy in English and Hungarian V: K. U. Panther in G. Radden, ur., *Metonymy in Language and Thought*, str. 333–357. Amsterdam: John Benjamins.

Dalibor Pavlas, Ondřej Vrabeľ in Jiří Kozmér. 2018. Applying MIPVU Metaphor Identification Procedure on Czech. V: *Proceedings of the Workshop on Annotation in Digital Humanities co-located with ESSLLI 2018*, str. 41–46. Sofia, Bulgaria.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol* 22 (1): 1–39.

Günter Radden in Zoltan Kövecses. 1999. Toward a theory of metonymy. V: K.-U. Panther in G. Radden, ur., *Metonymy in language and thought*, str. 17–60. Amsterdam: John Benjamins.

Maciej Rosiński in Joanna Marhula. 2015. MIPVU in Polish: On Translating the Method. *RaAM Seminar 2015*.

Elena Semino in Michela Masci. 1996. Politics is football: metaphor in the discourse of Silvio Berlusconi in Italy. *Discourse and Society* 7/2: 243–269.

Elena Semino. 2017. Corpus linguistics and metaphor. V: *The Cambridge Handbook of Cognitive Linguistics*, str. 463–476. Cambridge: Cambridge University Press.

Hanna Skorczynska in Kathleen Ahrens. 2015. A corpus-based study of metaphor signaling variation in three genres. Text & Talk. *An Interdisciplinary Journal of Language Discourse Communication Studies* 35(3): 359–381.

Slovar slovenskega knjižnega jezika, druga, dopolnjena in deloma prenovljena izdaja. www.fran.si.

David Stallar. 1993. Two Kinds Of Metonymy. V: *31st Annual Meeting of the Association for Computational Linguistics*, str. 87–94. Association for Computational Linguistics: Columbus, Ohio.

Gerard J. Steen, Aletta G. Dorst, Berenike J. Herrmann, Anna A. Kall, Tina Krennmayr in Tryntje Pasma. 2010. *A method for linguistic metaphor identification. From MIP to MIPVU*. Amsterdam: John Benjamins.

Anatol Stefanowitsch. 2006. Corpus-based approaches to metaphor and metonymy. V: A. Stefanowitsch in S. Th. Gries, ur., *Corpus-Based Approaches to Metaphor and Metonymy*, str. 1–17. Berlin: De Gruyter Mouton.

Elen Tissari. 2003. *LOVEscapes: Changes in Prototypical Senses and Cognitive Metaphors Since 1500*. Societe Neophilologique.

Jose Tummers, Kris Heylen in Dirk Geeraerts. 2005. Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1(2): 225–261.

Justina Urbonaitė. 2016. Metaphor identification procedure MIPVU: an attempt to apply it to Lithuanian. *Taikomoji kalbotyra [Applied Linguistics]* 7: 1–25.

Thomas Wasow in Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115: 1481–1496.

Beatrice Warren. 2002. An alternative account of the interpretation of referential metonymy and metaphor. V: R. Dirven in R. Pörings, ur., *Metaphor and Metonymy in Comparison and Contrast*, str. 113–133. Berlin: De Gruyter Mouton.

Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Marko Stabej in Simon Krek. 2009. Načela transkribiranja in označevanja posnetkov v referenčnem govornem korpusu slovenščine. V: M. Stabej, ur., *Infrastruktura slovenščine in slovenistike*, str. 437–442. Ljubljana: Znanstvena založba Filozofske fakultete.

Neural Translation Model Specialized in Translating English TED Talks into Slovene

Eva Boneš*, Teja Hadalin†, Meta Jazbinšek†, Sara Sever†, Erika Stanković*

* Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, 1000 Ljubljana
{eb1690,es6317}@student.uni-lj.si

† Faculty of Arts
University of Ljubljana
Aškerčeva 2, 1000 Ljubljana
{th3112,mj6953,ss6483}@student.uni-lj.si

Abstract

In this paper, we present our work on a neural translation model specialized in translating English TED Talks into Slovene. The aim is to provide transcriptions of the speeches in Slovene to make them available to a wider audience, possibly with the option of automatic subtitling. First, we trained a transformer model on general data, a collection of corpora from the Opus site, and then fine-tuned it on a specific domain which was a corpus of TED Talks. To see the functionality of the model, we carried out an evaluation of the pretrained, general, and domain versions of the model. We evaluated the translations with automatic metrics and manual methods – the adequacy/fluency and the end-user feedback criterion. The analysis of the results showed that our translation model did not produce the expected results and it can not be used to translate speeches in real life. However, in the TED talks addressing more everyday issues and using simple vocabulary, the translations successfully conveyed the main message of the speech. Any further research should consider improvements, such as including more specialized data covering only one specific topic.

1. Introduction

In this paper, we trained a transformer model from scratch on a large general corpus, which we then fine-tuned on a corpus consisting of TED Talks in order to make a model specialized for the translation of transcribed speeches. We also found a pretrained model for the baseline to which we were able to compare our translation models. We then automatically and manually evaluated all three models on the validation datasets constructed from TED Talks. Finally, we evaluated the general translation model on the validation dataset constructed from the large general corpus.

In Section 3, we first describe the data we used. In the subsequent Section, we describe all the methods for both training and evaluating the models. Later on, in Sections 5 and 6, we present the results and discuss them.

1.1. Goal of the paper

The main goal of this project is to provide a useful and effective tool for translating and subtitling speeches from English to Slovene, and this way granting access to a wide range of talks and other speeches to the Slovene-speaking audience. This paper focuses on translating TED Talks, a form of learning and entertainment that has gained popularity in recent years. Since TED Talks are currently subtitled by volunteer translators, enabling automatic subtitles would facilitate this process. Machine translation (MT) has been researched since the 1950s, but only recently, with the rise of deep learning, did it prove to be solvable, although the possibility of achieving fully automatic machine translations of high quality is still being questioned. This project was our attempt at machine translation of spoken language, which, if efficient, could also be used for automatic subtitling in general.

2. Related work

There are three main approaches to solving the MT problem, all with their own advantages and shortcomings. The rule-based machine translation (RBMT) is the oldest of the bunch and it requires expert knowledge of both the source and the target language in order to develop syntactic, semantic, and morphological rules. Another approach, which gained popularity in the 1990s, uses statistical models based on the analysis of bilingual text corpora. The idea behind statistical machine translation (SMT) as proposed in (Brown et al., 1990) is, if given a sentence in the target language, we seek the original sentence from which the translator produced it. Today, as with many computer science fields, the current state-of-the-art approaches for machine translation are based on neural networks. The biggest challenge when building a successful English to Slovene (or vice-versa) automatic translator is obtaining a sufficiently large bilingual corpus. Like all deep learning approaches, having a large and quality dataset is crucial for the success of the model. To deal with this exact problem, a lot of approaches to pre-training a network on monolingual data (that can be obtained easily) have been proposed.

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) uses two strategies to deal with the problem, namely masked language modeling (MLM) and next sentence prediction (NSP). By using these two strategies in our models, we generally achieve bigger datasets and a model with more context-awareness.

In 2020, the mRASP (Lin et al., 2021) was introduced. Its authors built a pretrained NMT model that can be fine-tuned for any language pair. They used 197M sentence pairs, which is considerably more than we could obtain for only English-Slovene translations.

Although these methods have proven to be successful, one of the largest currently available databases of pretrained translation models was trained using just a standard transformer model and it still achieved great results. The Tatoeba Translation Challenge (Tiedemann, 2020) aims to provide data and tools for creating state-of-the-art translation models. The focus is on low-resource languages to push their coverage and translation quality. It currently includes data for 2,963 language pairs covering 555 languages. Along with the data, pretrained translation models for multiple languages were also released and are being regularly updated.

3. Dataset

3.1. General translation model

The datasets for the general translation model are the eight biggest corpora from the Opus site (`https://opus.nlpl.eu` (Tiedemann, 2012)) for the Slovene-English language pair. The corpora were chosen based on the quantity of the data, so the general translation model would contain a large amount of diverse information. After a brief look at the contents of each one, we can see that some datasets are of higher quality and more reliable because of the source of the original texts and their translations For example, the corpora from European institutions, such as **Europarl**, which is a parallel corpus extracted from the proceedings of the European Parliament from 1996–2011, and the **DGT** corpus, which is a collection of translation memories from the European Commission's Directorate-General for Translation. The other corpora are a collection of translations from different Internet sources, which makes them less reliable, however, they are still very valuable because they ensure a large quantity of the data. These include the **CCAligned** corpus consisting of parallel or comparable web-document pairs in 137 languages aligned with English, the **MultiCCAligned v1** multi-parallel corpus, the **OpenSubtitles** corpus compiled from an extensive database of movie and TV subtitles, the **Tilde MODEL** corpus consisting of over 10M segments of multilingual open data for publication on the META-SHARE repository, the **WikiMatrix v1**, a parallel corpus from Wikimedia compiled by Facebook Research, the **Wikimedia v20210402** corpus, and the **XLEnt v1** corpus created by mining CCAligned, CCMatrix, and WikiMatrix parallel sentences. The exact size of each one, complete with the number of tokens, links, sentence pairs, and words, is noted in Table 1.

3.2. Domain translation model

Our domain translation model is specialized in translating TED Talks.

For the domain-specific machine training, we opted for the two TED Talk corpora accessible on the Opus website – the TED2013 and TED2020 corpus. The included texts are mainly transcripts of speeches on various topics and their Slovene translations. Both datasets add up to 1.8 million words (MOSES format) and 2.1 million tokens, which is enough to form a well-rounded base for machine learning. For more information about the domain-specific corpora see Table 2.

We expanded the datasets by manually aligning 15 TED Talks from 2018 and 2019 that are available on the TED website (https://www.ted.com/talks).

4. Methods

4.1. Pretrained model

As a baseline for evaluating our models, we found an already trained model, available in HuggingFace (Tiedemann, 2020). It is a transformer-based multilingual model that includes all the South Slavic languages. The framework provides both the South-Slavic to English model and the English to South-Slavic model. On the Tatoeba test dataset for Slovene, the English to South-Slavic (en-zls) model has achieved 18.0 BLEU score and 0.350 chr-F score.

The model in question was trained using Marian-NMT (Junczys-Dowmunt et al., 2018). The authors applied a common setup with 6 self-attentive layers in both, the encoder and decoder network using 8 attention heads in each layer. SentencePiece (Kudo and Richardson, 2018) was used for the segmentation into subword units.

The translation model can be loaded through the *transformers* library in Python and for translation into Slovene, we must add the Slovene language label at the beginning of each sentence (>>slv<<).

4.2. Training from scratch

There exist several different frameworks to use with natural language processing tasks, each with their own advantages and shortcomings. One of them is fairseq (Ott et al., 2019) – a sequence modeling toolkit written in PyTorch for training models for translation, summarization, and other tasks. It provides different neural network architectures, namely convolutional neural networks (CNN), Long-Short-Term Memory (LSTM) networks, and Transformer (self-attention) networks. The architectures can be configured to specific needs and many implementations for different tasks have been proposed since the fairseq's introduction in 2019. In addition to different architectures, they also provide pretrained models and preprocessed test sets for different tasks, but sadly none of them is in Slovene.

For training our model from scratch, we have decided to use an extension of fairseq (stevezheng23, 2020) that has additional data augmentation methods. We have trained our general model on a corpus described in Subsection 3.1.

4.2.1. Preprocessing

Before training the model, we had to preprocess the data. The datasets were already formatted as raw text with one sentence per line and with lines aligned in English and Slovene datasets. We first normalized the punctuations, removed non-printing characters, and tokenized both corpora with Moses tokenizer (Koehn et al., 2007). We removed all the sentences that were too short (2 tokens or less) or

CORPUS	Tokens	Links	Sentence pairs (MOSES format)	Words (MOSES format)
Europarl.en-sl	31.5 M	0.6 M	624,803	27.56 M
CCAligned.en-sl	131.3M	4.4 M	4,366,555	110.08 M
DGT.en-sl	215.8M	5.2 M	5,125,455	162.58 M
MultiCCAligned.en-sl	5.6 G	4.4 M	4,366,542	110.01 M
OpenSubtitles.en-sl	178.0 M	2.0 M	19,641,477	213.00 M
TildeMODEL.en-sl	2305.4 M	21.1 M	2,048,216	79.90 M
WikiMatrix.en-sl	1.1 G	0.9 M	318,028	11.99 M
wikimedia.en-sl	350.6 M	31.8 K	31,756	1.50 M
XLEnt.en-sl	200.7 M	0.9 M	861,509	4.53 M

Table 1: Size of datasets for the general translation model.

CORPUS	Tokens	Links	Sentence pairs (MOSES format)	Words (MOSES format)
TED2013	0.5 M	15.2 k	14,960	0.45 M
TED2020	1.6M	43.9 k	44,340	1.35 M
Extras	23005	/	983	/

Table 2: Size of datasets for the domain translation model.

too long (250 tokens or more), and the ones where the ratio of lengths was too big because there is a good chance that these kinds of sentences are not translated properly. We then applied Byte pair encoding (BPE) (Sennrich et al., 2016) to the dataset. The algorithm learns the most frequent subwords to compress the data and thus induces some tokens that can help recognize less frequent and unknown words.

With this preprocessed data, we then built the vocabularies that we used for training and binarized the training data. Cleaned and preprocessed training data has $\approx 16M$ sentences with $\approx 345M$ tokens in English and $\approx 341M$ in Slovene. Both of the vocabularies have around 45,000 types. In the end, we split the data into a training and validation set.

4.2.2. Training

We trained a transformer (Vaswani et al., 2017) model with 5 encoder and 5 decoder layers in the fairseq framework. We used Adam optimizer, an inverse square root learning rate scheduler with an initial learning rate of $7e^{-4}$ and dropout. We also used the proposed augmentation with a cut-off augmentation schema that randomly masks words and this way produces more training data and a more robust translator.

We trained our model for 8 epochs with the mentioned initial learning rate, after which the minimum loss scale (0.0001) was reached, meaning that our loss was probably exploding. We tried training one more epoch with a lower initial learning rate and obtained an even worse performance with the minimum loss scale reached again. That is why we decided to stop the training at 8 epochs. Results of all the epochs are shown in Chapter 5.

4.3. Fine-tuning on TED talks

We preprocessed the TED data in the same way as the general, only this time we used the same dictionary as before and we did not build a new one. Less than 0.1% of tokens in training and validation sets were replaced with *unknown* tokens, so our original dictionary was evidently large enough. We used the best performing epoch from our general translation model (according to the loss on our validation set) for fine-tuning it on our domain data. We trained three different models with three slightly different configurations – one with the same augmentation parameters as the general model, one with increased masking probability and decreased dropout and initial learning rate, and one without augmentation. We trained all of the models for 100 epochs and we are presenting the results of the best epoch for each of them.

4.4. Evaluation

In order to test the performance of the pretrained and general translation model, and the fine-tuned translation model for TED Talks we had to evaluate the translations.

The automatic evaluation was carried out on two validation sets. First, the general translation model was evaluated on a subset of the general data, which was split in the preprocessing step (hereinafter referred to as the general validation set). All three models were evaluated on a subset of the domain data (hereinafter referred to as the domain validation set). The manual evaluation was only performed on a subset of the domain validation set, as described in Subsection 4.4.2.

4.4.1. Automatic evaluation

Since the manual evaluation of the translations is very time-consuming, it is very difficult to evaluate a sufficient amount of sentences this way. In cases like this, automatic evaluation metrics are often used. Natural language is quite subjective. Hence, the perfect measure does not exist, but by evaluating our results with different techniques, we were able to assess the performance of our translation model and compare it with other models. We used automatic metrics most often used in NLP tasks – namely BLEU, chr-F, GLEU, METEOR, NIST, and WER.

4.4.2. Manual evaluation

The translations were also evaluated manually, namely by the fluency-adequacy criterion first described by Church

(Church, 1993). For this part of the evaluation, the Excel format was used. We extracted 6 paragraphs containing 10 consecutive segments from each speech to ensure that the context was clear. Three evaluators (the translators from our group) were assigned 20 segments each. To determine the adequacy of the translation, the evaluator marks how much of the meaning expressed in the source text is also expressed in the target translation. To determine the fluency of the translation, the evaluator marks whether the translation is grammatically well-formed, contains the correct spelling, is intuitively acceptable, and can be sensibly interpreted by a native speaker. To test the adequacy, the evaluator compares both, the source text and the translation, whereas, in the process of the fluency evaluation, the focus is merely on the translation. The evaluators had to provide the scores on a scale from 1 to 4. We chose this evaluation technique because it clearly and simply summarizes and presents the quality of the translations. Since we evaluated three different translation models (pretrained, general, and domain), we had to evaluate the same segments of texts three times. Evaluating one text multiple times by the same person is not recommended, therefore, the translations were exchanged between the three evaluators at the beginning of the evaluation of each translation model.

4.4.3. End user comprehensibility questionnaire

Finally, we evaluated the domain machine-translated texts from the end-user's point of view. Evaluators, who were not familiar with the content of this project, were given the translated texts from the domain model and a questionnaire formed by the translation team of this project. The objective of this questionnaire was to examine whether the end-users understand the information given in the translation, meaning it tested the functionality of the text. The questionnaire was given to nine persons, each evaluating 20 segments from two different speeches - the segments were identical to segments used in the manual evaluation. In the end, we obtained three evaluations for each text (6 speeches altogether). The questionnaire included the following questions:

1. How comprehensible is the text?
2. To what degree does the text seem like it was produced by a native speaker of Slovene?
3. How would you grade the text as a whole?
4. What is the main message of the text?
5. What do you consider as the most problematic part of the text?

For the first and second question, the end-users answered on a scale from 1 to 4, with 1 meaning 'not at all' and 4 meaning 'very much'. The third answer had to be a score from 1 to 4. The fourth question had to be answered with one sentence, and for the fifth question, they had to choose between the following answers: 'unknown words', 'too little context', 'wrong syntax', and 'other'. We chose this evaluation technique because it shows whether the translation is, in fact, functional and useful to the end-user.

5. Results

For the training of our models, we used the Slovenian national supercomputing network that provides access to cluster-based computing capacities. We used the Arnes cluster which is equipped with 48 NVIDIA Tesla V100S PCIe 32GB graphic cards. When training on two of them, one epoch took approximately 4 hours for the general translation model and one minute for fine-tuning on the TED data.

5.1. Automatic evaluation results

In Table 5, we present the quantitative results of the automatic evaluation for the pretrained, general, and domain models.

5.2. Manual evaluation results

Along with the automatic evaluation metrics, we also performed a manual evaluation which provided a valuable human insight into the final product and a better understanding of the typology of the mistakes that occurred in the translations. Each validation set was assessed by two evaluators at all three stages of the model development. The results presented in Table 4 represent the average value of the fluency and adequacy scores for the pretrained, general, and domain models, respectively.

MODEL	Fluency	Adequacy
Pretrained	2.99	3.09
General	2.83	2.9
Domain	2.71	2.9

Table 3: Manual evaluation results on the TED validation set.

5.3. End-user comprehensibility questionnaire results

We received feedback from the end-users based on the questionnaire for the texts from the domain translation model. The average score of the answers that could be interpreted numerically is presented in Table 4. According to the answers to the question 'What is the main message of the text?', the users have, for the most part, understood the text to the degree where they could sufficiently summarize the content. The most frequent answer to the last question (What do you consider as the most problematic part of the text?) was 'wrong syntax', followed by 'lack of context' and 'unknown words'. The participants also pointed out that the general structure of the text was rather confusing.

Text	Question 1	Question 2	Question 3
1	1.33	1	1
2	2	1.33	1.33
3	3	2	2.33
4	1.66	1	1.33
5	2	1.66	1.66
6	2.33	1.66	2
All	2.05	1.44	1.61

Table 4: End-user feedback results from the questionnaire with average scores on a scale from 1 to 4.

Dataset	Metric	Pretrained	General (epochs)								Domain		
			1	2	3	4	5	6	7	8	Configuration 1	Configuration 2	Configuration 3
General	BLEU	-	0.387	0.398	0.405	0.409	0.411	0.417	0.417	**0.420**	-	-	-
	chr-F	-	0.606	0.616	0.619	0.624	0.625	0.629	0.629	**0.629**	-	-	-
	GLEU	-	0.391	0.401	0.407	0.411	0.413	0.417	0.417	**0.420**	-	-	-
	METEOR	-	0.545	0.556	0.560	0.565	0.566	0.569	0.569	**0.571**	-	-	-
	NIST	-	8.752	8.922	8.987	9.063	9.096	9.144	9.114	**9.177**	-	-	-
	WER	-	0.518	0.508	0.503	0.501	0.496	0.497	0.498	**0.494**	-	-	-
Domain	BLEU	**0.192**	0.155	0.167	0.168	0.171	0.175	0.175	0.168	0.179	***0.182***	0.173	0.114
	chr-F	**0.514**	0.487	0.496	0.495	0.497	0.500	0.498	0.500	***0.505***	0.503	0.497	0.440
	GLEU	**0.230**	0.201	0.211	0.212	0.214	0.217	0.218	0.213	0.222	***0.224***	0.216	0.167
	METEOR	0.420	0.398	0.407	0.409	0.409	0.414	0.412	0.416	0.420	**0.426**	0.416	0.346
	NIST	**5.481**	4.877	5.067	5.105	5.132	5.151	5.179	5.074	5.230	***5.344***	5.209	4.228
	WER	**0.659**	0.711	0.696	0.694	0.690	0.689	0.689	0.698	0.685	***0.667***	0.680	0.756

Table 5: Evaluation scores for all models and all validation datasets. The best scores for each dataset and each metric are shown in bold. If the best score was the pretrained model, the second best score is shown in bold and italic to showcase our best score.

6. Discussion

Looking at the results in Table 5, we can first see that on the general validation set, the final epoch of our general model performs the best according to most metrics. This is expected, as the general validation set is comprised of the texts from the corpora that we used for training, so our model may be overfitted on this dataset.

Connected to this, all of the results in the domain validation set are considerably worse than in the general dataset. We can account this to the fact that the domain validation set is truly different from the main training data. As to why the pretrained model in most aspects performs better than our fine-tuned model, we assume that our domain data is not specific enough. Therefore, we could not really fine-tune our model to any specific styles or words, nor were we able to do that in the validation set. The pretrained model performs better because it is trained on a larger dataset than our domain model is fine-tuned on – the TED corpus is relatively small even though we included some additional texts.

Similarly, the results of the manual evaluation showed that the pretrained model produced the most fluent translations with an average score of 2.99 out of 4. This model also achieved the highest score in the adequacy criterion. If we take a closer look at the results of the other two models, it can be seen that both models faced similar difficulties in translating phrasal verbs, terminology, word order, and other lexical structures. The manual evaluation results are relatively low: the general and the domain model received an average of less than 3 points, in both fluency and adequacy. The following examples show the discrepancies between the pretrained model and the other two models on the syntactic, semantic, and morphological levels:

Original: *So then, what is our gut good for?*
Pretrained: *Torej, za kaj je naš občutek dober?*
General: *Torej, kaj je naš črevo dobro za?*
Domain: *Kaj je torej naš črevesje dobro?*

Original: *And I was not only heartbroken, but I was kind of embarrassed that I couldn't rebound from what other people seemed to recover from so regularly.*
Pretrained: *Ne samo, da me je zlomilo srce, ampak me je bilo sram, da se nisem mogel odvrniti od tega, kar so si drugi ljudje zdelo, da si je opomoglo tako redno.*
General: *In nisem bil samo zlom srca, ampak sem bil neprijetno, da se nisem mogel odvrniti od tega, kar se je zdelo, da se drugi ljudje tako redno opomorejo.*
Domain: *In nisem bil le srčni utrip, ampak sem bil neprijetno, da nisem mogel vrniti od tega, kar se je zdelo, da se drugi ljudje tako redno opomorejo.*

However, a quick analysis of the evaluation rates showed that the lowest ratings for the domain model appeared in segments with specialized vocabulary, for example: *"Ampak ko gre za res velike stvari, kot bo naša kariera ali kdo se bo poročil, zakaj bi morali domnevati, da so naše intuicije bolje kalibrirane za te kot počasne, pravilne analize?"* vs the original: *"But when it comes to the really big stuff, like what's our career path going to be or who should we marry, why should we assume that our intuitions are better calibrated for these than slow, proper analysis?"*, and in segments with a higher register, for example, the eloquent text on immigrants: *"Ta vprašanja so protipriseljenska in nativistična v svojem jedru, zgrajena okoli neke vrste hierarhične delitve notranjih in zunanjih oseb, nas in njih, v katerih smo pomembni le in ne." vs the original: "These questions are anti-immigrant and nativist at their core, built around a kind of hierarchical division of insiders and outsiders, us and them, in which only we matter, and they don't."*. In both cases, the rate was never lower than 2.8. The highest rated segments (with the score above 3) included short and simple sentences with everyday vocabulary, such as *"In rekla mi je: Samo dihajte."* or *"Na srečo kriminalci podcenjujejo moč prstnih odtisov."*. Based on the evaluation results, it appears that our domain model would be more valuable in translating general texts with a neutral style and vocabulary.

The group members that evaluated these segments had been participating in this project from the very beginning, so it was crucial to obtain a more objective assessment of our models. Looking at the results from Table 5, the gathered feedback from the questionnaire revealed that overall, the end-users thought that the texts are relatively comprehensible, but are not at all seen as being produced by a native speaker of Slovene. For the first two questions, for which the answers were chosen on a scale from 1–4 (1='not at all'/2='little'/3='good'/4='very much'), only two texts received a score lower than 2 in terms of comprehensibility. When grading the texts, the highest average score for a specific text was 2.33, while the lowest is 1. This variation occurs because not all of the chosen texts were equally complex. For the highest graded text, we received similar responses to the question asking what the main message of the text was: *Opisovanje prstnih odtisov./Puščanje prstnih odtisov./Prstni odtisi poleg vizualne sledi pustijo tudi sled na molekularnem nivoju.* There were only two out of eighteen answers stating that the message was not clear and where the end-users could not summarize the main message, i.e. in texts 1 and 5. The fact that the end-users were in almost all cases able to summarize the main message in one sentence shows that comprehension of the text was still possible despite a large number of significant mistakes (wrong syntax, unknown words, lack of context, changing genders, etc.).

The following examples, segments from text 2, text 3, and text 6, which have also been scored above average in manual evaluation, support this claim:

Original: *And you need something else as well: you have to be willing to let go, to accept that it's over.*
Domain: *Potrebujete tudi nekaj drugega : biti morate pripravljeni pustiti, da sprejmete, da je konec.*

Original: *I'm talking about an entire world of information hiding in a small, often invisible thing.*
Domain: *Govorim o celotnem svetu informacij, ki se skrivajo v majhni, pogosto nevidni stvari.*

Original: *Five years ago, I stood on the TED stage, and I spoke about my work.*
Domain: *Pred petimi leti sem stal na odru TED in govoril o svojem delu.*

Unfortunately, the final version of a machine translator did not meet our expectations regarding the quality of the translations. Some of the major flaws that appeared in the translations were wrong syntax, untranslated words, incomprehensible grammatical structures, wrong use of terminology, and wrong translations of polysemes. While we expected the machine translator to be inappropriate for translating complex sentences, we were surprised that it did not perform well when translating even basic grammatical structures. Here are two examples:

Original: *So then, what is our gut good for?*
Domain: *Kaj je torej naš črevesje dobro?*

Original: *I later found out that when the gate was opened on a garden, a wild stag stampeded along the path and ran straight into me.*
Domain: *Kasneje sem ugotovil, da ko so vrata odprta na vrtu, je divji stag žigosanih po poti in tekel naravnost v mene.*

Original: *And for two years, we tried to sort ourselves out, and then for five and on and off for 10.*
Domain: *Dve leti smo se poskušali razvrstiti, nato pa pet let in več.*

The reasons for the poor functioning of the machine translations could be numerous. It is possible that we have not collected enough data or that the chosen data might not have been the most suitable for this project. We estimate that the main factor that impacted the final results the most is the wide range of different topics covered in TED Talks. This means that our domain translation model did not focus on just one domain and, essentially, there was not enough specific data from which it could train. What is more, the initial data consisted of transcriptions of English spoken discourse and their Slovene translations in the form of subtitles. It is important to keep in mind that neither spoken discourse nor subtitles have characteristics typical for standard text types. Finally, not all of the chosen texts were equally complex and they had different syntactic, morphological, and lexical features. Therefore, some of the texts in the data were essentially too difficult to translate.

7. Conclusion

The main purpose of this project was to develop a tool that would automatically provide Slovene transcriptions or subtitles for English TED Talks. Our domain translation model provides translations that convey the main message of the texts, is based on the appropriate methodology, and built with all the necessary tools. Even more, the results of automatic metrics showed that it is comparable to other neural machine translation models. On the other hand, the lack of a uniform training dataset resulted in poor and incomprehensible translations. However, we believe that acknowledging all of the discussed shortcomings in future research could significantly improve the development of speech-to-text and translation technologies for Slovene language users. Neural machine translation is still relatively new and will develop in the following years because it is useful for translators and the general public. Our project contributed to the advancement of the field and could provide valuable information for similar work in the future.

Acknowledgments

We would like to thank our mentors, Slavko Žitnik, Špela Vintar, and Mojca Brglez, for helping us with the project. We would also like to thank the nine evaluators who provided end-user feedback by filling out our questionnaire.

We would also like to thank SLING for giving us access to powerful graphic cards to successfully finish our training, as we would still be training our general model without them. Special thanks to Barbara Krašovec from Arnes support who helped us with our numerous problems when trying to connect to their cluster.

8. References

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85.

Kenneth Church. 1993. Good applications for crummy machine translation. *Machine Translation*, 8:239–258, 12.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In: *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2021. Pre-training multilingual neural machine translation by leveraging alignment information.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In: *Proceedings of NAACL-HLT 2019: Demonstrations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

stevezheng23. 2020. fairseq_extension.

Jörg Tiedemann. 2012. Parallel data, tools and inter- faces in opus. In: *Proceedings of the Eight Interna- tional Conference on Language Resources and Evalua- tion (LREC'12)*, Istanbul, Turkey, may. European Lan- guage Resources Association (ELRA).

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In: *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online, November. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Govoriš nevronsko? Kako ljudje razumemo jezik sodobnih strojnih prevajalnikov

David Bordon
Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
david.bordon@ff.uni-lj.si

Povzetek

Namen prispevka je predstaviti raziskavo preverjanja razumljivosti nerevidiranih strojno prevedenih spletnih besedil. Primarni udeleženci v raziskavi so bili splošni bralci in ne izurjeni prevajalci ali popravljalci strojnih prevodov. Gre za prvo tovrstno raziskavo, ki je bila izvedena za slovenski jezik. Cilj raziskave je bil preveriti, v kolikšni meri so nerevidirani strojni prevodi razumljivi splošnemu bralstvu, pri čemer sem se posvetil tudi vplivu besedilnega in slikovnega konteksta. Preverjal sem prevode prevajalnikov Google Translate in eTranslation. Raziskava je bila izvedena z anketo, v kateri so udeleženci odgovarjali na vprašanja, ki so preverjala razumevanje spremljajočega besedilnega segmenta, v katerem je bila napaka. Rezultati nudijo vpogled v trenutno stopnjo razvoja strojnih prevajalnikov, ne z vidika storilnosti pri njihovem popravljanju, ampak z vidika, koliko jih razume ciljno bralstvo.

Do you Speak Neuralese?

The aim of this paper is to present a study on the comprehensibility of unedited machine-translated web texts. The primary participants in the study were general readers, not trained translators or post-editors, and it is the first study of its kind to be conducted for the Slovene language. The aim of the study was to examine the extent to which unedited machine translations are comprehensible to general readers, while giving focus to the influence of textual and pictorial context. The translations were obtained from Google Translate and eTranslation. The survey was conducted by means of a questionnaire, in which participants answered questions that tested their understanding of a text segment that included an error. The results provide an insight into the current state of development of machine translation engines, not from the point of view of PEMT, but from the point of view of how well machine translations are understood by the target readership.

1. Uvod

Članek obravnava raziskavo razumljivosti strojno prevedenih spletnih besedil pri bralcih, ki ne vedo, da prebirajo strojne prevode. Uporabil sem naključno izbrana angleška spletna besedila, slovenske prevode pa sem pridobil z nevronskima strojnima prevajalnikoma Google Translate in eTranslation. Prevodi niso bili revidirani, saj sem želel replicirati okoliščine, v katerih bi jih dejansko lahko našli – na spletu, kjer so zaradi (za nekatere) dovolj visoke kakovosti in cenovne nepremagljivosti (namreč so brezplačni) vedno bolj pogosta, kar velja tudi za prevajalske vtičnike, ki so vgrajeni v sodobne brskalnike in aplikacije.

Vprašanje razumljivosti v taki obliki je postalo aktualno samo v zadnjem času, saj so starejši, statistični modeli prevajalnikov slovnično nekonsistentni in jezikovno okorni, sodobni nevronski prevajalniki pa proizvajajo tekoča besedila, ki so težje ločljiva od človeških, hkrati pa je že profesionalnim pregledovalcem prevodov težje ugotoviti, kje so storili napako (Donaj in Sepesy Maučec, 2018).

Te napake nastanejo predvsem zaradi težav pri razdvoumljanju večpomenskih besed in prevajanju besed, ki jih ni v podatkovni zbirki, s katero smo prevajalnik urili (Thi-Vinh et al. 2019, 207; Koehn in Knowles 2017, 28, 31–33; Sennrich et al. 2016, 3). Kljub morebitnim posamičnim napačno prevedenim besedam pa lahko ljudje pomen razberemo iz sobesedila. Pri preverjanju razumljivosti sem v vseh primerih vključil še kontekst, saj se v stvarnosti bralci nikoli ne srečujejo z izoliranimi besedami, ampak z zaključenimi besedili, ker pa se osredotočam na spletno okolje, sem besedilnemu kontekstu dodal še slikovnega, ki je inherentna lastnost sodobnega spleta.

2. Namen članka

Namen članka je predstaviti grobo oceno razumljivosti prevodov NMT-sistemov (ang. *Neural machine translation*) v času, ko so taka besedila na spletu vedno bolj pogosta, pri čemer me zanima predvsem, kako slikovno gradivo v besedilnem kontekstu vpliva na rezultate. Tovrstna raziskava za slovenščino še ni bila izvedena.

2.1. Sorodne raziskave

Raziskav na področju razumevanja nerevidiranih strojnih prevodov pri naključnem splošnem bralstvu je razmeroma malo, saj je z vidika omejenosti na stroko in gospodarstvu bolj zanimive analize storilnosti pri popravljanju prevodov veliko več raziskav osredotočenih zgolj na prevodno prebivalstvo.

Razširjenost prakse popravljanja strojnih prevodov lahko opazimo že v zapisih o najboljših praksah pri popravljanju prevodov, ki so zapisani v blogih večjih ponudnikov jezikovnih rešitev, kot so denimo MemoQ (Lelner, 2022), Crowdin (Voroniak, 2022) in Memsource (Zdarek, 2020).

Na Univerzi v Gentu je bila v sklopu projekta ArisToCAT izvedena raziskava o razumevanju izmišljenih besed in samostalniških besednih zvez (Macken et al. 2019). Primeri, ki so bili iz angleščine v nizozemščino prevedeni s strojnima prevajalnikoma Google Translate in DeepL, so bili predstavljeni samostojno ali v kontekstu povedi, pri tem pa udeleženci niso imeli dostopa do izvirnega besedila. V povprečju je bilo 60 % odgovorov napačnih; rezultati so bili boljši, če je bil primer predstavljen v kontekstu povedi.

V sklopu istega projekta je bila izvedena še analiza bralnega razumevanja človeškega prevoda na eni in nepopravljenega strojnega prevoda na drugi strani.

Človeški prevodi so bili ocenjeni bolje z vidika jasnosti podajanja informacij, z vidika končnega razumevanja pa je bila razlika manjša (Macken in Ghyselen, 2018).

Castilho in Guerberof Arenas (2018) sta izvedli primerjalno analizo bralnega razumevanja za statistični in nevronski model strojnega prevajalnika v primerjavi s človeškim izvirnikom. Glede na omejen vzorec (6 udeležencev) in nedoslednost rezultatov je končna ugotovitev, da sistemi-NMT izkazujejo najboljše rezultate, občasno še boljše kot angleški izvirnik, nedokončna.

Martindale in Carpuat (2018) sta v raziskavi obravnavali odziv bralcev na tekočnost in natančnost nevronskih strojnih prevodov, ob tem pa sta preverjali stopnjo zaupanja informacijam v besedilu. Ugotovili sta, da bralce zelo zmotijo prevodi, ki niso tekoči, medtem ko se ob samo natančnost informacij obregne veliko manjši delež bralstva.

Izsledke potrjuje tudi Popović (2020). V njenem eksperimentu so bralci v 30 % primerov zaradi zavajajoče tekočnosti sprejeli popolnoma napačno informacijo, še 25 % dodatnih primerov pa je bilo skoraj popolnoma (narobe) razumljivih.

Na tem mestu velja omeniti, da so se nedavno začele pojavljati bolj eksperimentalne metode prevajanja, katerih značilnost je upoštevanje multimedijskega konteksta, denimo zvočnega ali slikovnega. Lala in Specia (2018) sta razvila model multimedijskega leksikalnega prevajanja, katerega namen je prevajanje dvoumnih večpomenskih besed s pomočjo slikovnega konteksta. Sulubacak et al. (2020) so predstavili sorodne raziskave, uporabne podatkovne zbirke in metode raziskovanja na področju multimedijskega strojnega prevajanja, ki so vezane na prevajanje z zvokom, sliko in videom. Med novejšimi raziskavami Liu (2021) ponuja nevronski model vizualno-tekstovnega enkodiranja in dekodiranja.

Pričakujemo lahko, da se bo to področje v bodoče še hitreje razvijalo, predvsem zaradi tehnološkega napredka v drugih panogah (prepoznavanje slik, sinteza govora, avtomatsko podnaslavljanje ipd.).

3. Metoda

Raziskava je bila zasnovana okrog vprašalnika, ki je vseboval primere štirih vrst napak v slovenskih strojnih prevodih splošnih angleških spletnih besedil. Preverjal sem prevajalnika Google Translate in eTranslation, pri čemer je bil vsak zastopan z 12 vprašanji. Poseben pomen sem posvetil slikovnemu gradivu v sobesedilu.

3.1. Izbor besedil

Besedila sem zbiral glede na verjetnost, da bi se bralci z njimi lahko dejansko srečali na spletu. Analiza prevajalskega trga je pokazala, da večje prevajalske agencije popolnoma obvladujejo sektorje, ki nudijo največ dobička in hkrati zahtevajo človeško revizijo (tehnika, zdravstvo, pravo, finance ipd.) (Evropska komisija, 2020). V manj dobičkonosnih sektorjih, kjer človeška revizija ni tako bitna, obstaja večja verjetnost objave nerevidiranih strojnih prevodov.

Pregleda tržnega deleža spletnih iskalnikov, ki jih uporabljamo v Sloveniji je pokazal, da 96 % vseh uporabnikov spleta uporablja iskalnik Google.[1] Na osnovi najbolj iskanih pojmov v brskalniku[2] sem izločil spletišča, ki nimajo prevodnega potenciala (družbena omrežja, spletni portali v slovenščini, slovenski mediji). S tem sem prišel do končnega izbora besedilnih področij: spletno nakupovanje, turizem, elektronika, multimedija in videoigre, luksuzne storitve, moda, osebno zdravje (telesna vadba in prehrana).

3.2. Prevodi besedil

Pri preizkušanju strojnih prevajalnikov se je izkazalo, da Googlov prevajalnik nudi drugačne prevodne rešitve glede na to, kako besedilo naložimo v obdelavo. Če besedilo prevajamo v pogovornem oknu vmesnika ali v brskalniku prevedemo spletno stran kot celoto, so rezultati boljši kot tisti, ki jih dobimo s funkcijo prevajanja dokumenta. Od štirih različnih specializiranih domen, ki jih nudi eTranslation, je najboljše rezultate nudil prevajalnik za splošna besedila (General Text). Uporabil sem najboljše možne prevode – omenjeno domeno v eTranslation, v Googlu pa sem prevajal v pogovornem oknu.

Prevod iz vnosnega polja oz. samodejni prevod strani	Prevod, pridobljen s funkcijo »prevedi dokument«	Izvirnik
Naj bo topla - mikrovalovna pečica ohranja hrano, kot so zelenjava, juhe, jedi, graviža, omake in sladice, topla in okusna v pečici, dokler niso pripravljene za postrežbo.	Naj bo toplo funkcijo - Mikrovalovna ohranja živila, kot so zelenjava , juhe , nerazporejenega d'oeuvres , gravies , omake in sladice toplo in okusno v pečice , dokler oni propravljeni , da služijo.	Keep Warm Feature Maintains Food Temperature Keeps foods like vegetables, soups, hors d'oeuvres, gravies, sauces and desserts warm and delicious in the oven until they're ready to serve.

Tabela 1: Razlike v prevodih glede na način obdelave; Google Translate.

Prevod modela »General Text« prevajalnika eTranslation
Ohraniti toplo funkcijo - Microwave ohranja hrano, kot so zelenjava, juhe, predjed d'oeuvres, omake, omake in sladice tople in okusne v pečici, dokler niso pripravljeni za postrežbo

Tabela 2: Prevod enakega segmenta; eTranslation.

[1] https://gs.statcounter.com/search-engine-market-share/all/slovenia

[2] https://ahrefs.com/keyword-generator

3.3. Kategorizacija napak

Prevode sem analiziral in določil štiri kategorije najpogostejših napak, ki niso vezane na jezikovni sistem oz.predpis.

- **Neprevedena beseda;** v prevodu se pojavlja beseda v enaki obliki kot v izvirniku. Dopustil sem možnost spremembe začetnih ali končnih morfemov, če je prevajalnik besedo samo preoblikoval[3].
- **Napaka pri razdvoumljanju večpomenske besede;** denotativni pomen večpomenske besede ali besedne zveze ne ustreza pomenu v izvirniku.
- **Hujša pomenska napaka;** napaka, ki otežuje razumevanje celotnega besedila.
- **Izmišljena beseda;** prevajalnik si izmisli novo besedo, ki je na prvi pogled videti slovenska, a ne spada v slovensko besedišče – t. i. »nevronščina«.

3.4. Kontekst

Izbranim besedilom sem glede na inherentne lastnosti spletne pojavitve dodal kontekst. Kontekst je lahko bil več vrst:

- izključno besedilni,
- besedilni in slikovni; slika ne vpliva na razumevanje,
- besedilni in slikovni; slika vpliva na razumevanje,
- izbor ene izmed več predlaganih slik glede na to, kaj piše v besedilu.

Slikovni kontekst sem vključil pri besedilih, ob katerih so se na spletu pojavljale fotografije, ki so pri nekaterih primerih bile zgolj vizualni dodatek, pri drugih pa je bilo pravilno razumevanje besedila vezano na prepoznavanje pravilnega vizualnega elementa.

V svoji raziskavi besed nisem nikoli predstavil v izolaciji, kot so to denimo storili v raziskavi Macken in drugi (2019), saj to niso realne okoliščine – napake v objavljenih strojnih prevodih bodo vedno del nekega besedila. Besedil nisem popravljal, anketirancem so bila predstavljena vključujoč vse slovnične in pomenske napake, take, kot bi jih našli v divjini.

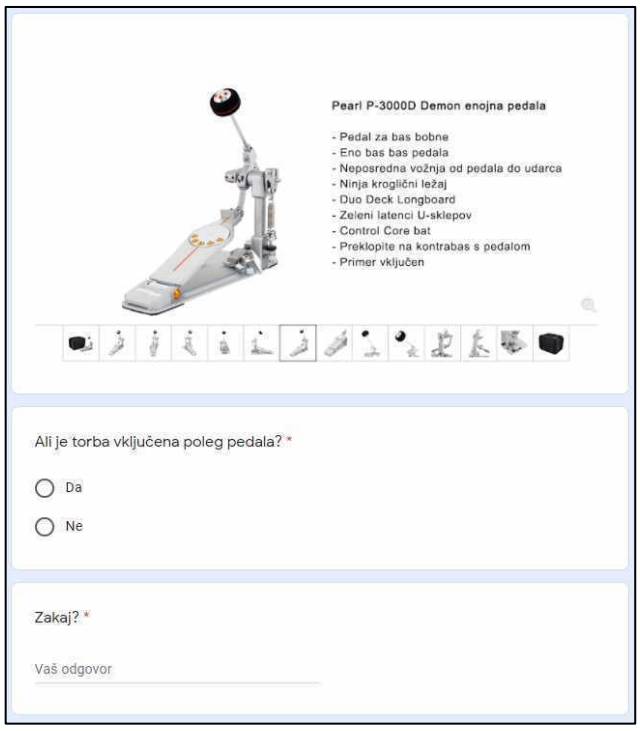
3.5. Oblikovanje vprašalnika, format odgovorov na vprašanja in udeleženci

Anketo sem ustvaril na platformi Google Forms, ki nudi podporo za prikaz slik in dober vmesnik za pregled in izvoz rezultatov. Pomembno je poudariti, da anketirancem nisem razkril, da bodo brali strojno prevedena besedila. Omenil sem, da bodo »prebrali več kratkih besedil, ki so napisana v nekoliko okorni slovenščini«.

Vrste odgovorov so bile omejene s funkcionalnostjo platforme Google Forms in niso sledile nobeni logični metodi; določil sem jih subjektivno glede na vsebino primera in vrsto napake. Gre za najbolj nezanesljivo spremenljivko v metodi, saj bi s formulacijo vprašanja lahko sugeriral pravilen odgovor, zanimalo pa me je predvsem to, če prihaja do večjega odstopanja glede na tip odgovora, denimo, če bi bili odgovori odprtega tipa, kjer anketiranci vnesejo svoj odgovor v prazno vnosno polje, bistveno slabši kot tisti, kjer izbirajo med štirimi predlaganimi odgovori. S tem bi lahko preveril konsistenco pravilnosti oz. odstopanja glede na vrsto odgovora.

Vprašalnik sem delil na družbenih omrežjih Facebook in Instagram in znance pozval, naj ga posredujejo naprej svojcem in svojim znancem, če je le mogoče starejšim. Demografskih podatkov nisem zbiral, kar je mogoče ena izmed pomanjkljivosti raziskave. Glede na razmeroma majhen vzorec sodelujočih in morebiten efekt odmevne komore bi bilo vsekakor raziskavo potrebno nadgraditi in ponoviti na bolj naključnem in predvsem večjem vzorcu, toda glede na čas zbiranja odzivov, ki je sovpadal s prvo omejitvijo gibanja vezano na epidemijo Covid-19, nisem imel druge izbire.

Na vprašalnik sem prejel 120 odgovorov.



Slika 1: Primer vprašanja. Izbor z razlago.

4. Rezultati

Rezultate predstavljam po naslednjih parametrih:

- splošno razumevanje,
- razumevanje glede na prevajalnik,
- razumevanje glede na tip napake,
- razumevanje glede na tip konteksta,
- razumevanje glede na tip odgovora.

4.1. Splošno razumevanje

Vprašalnik je obsegal 24 vprašanj, s 120 odzivi je bilo vseh možnih odgovorov 2880. Vseh pravilnih odgovorov je bilo 1697 oz. 58,96 %. Daljša razčlemba je na voljo v celotni raziskavi (Bordon, 2021).

4.2. Razumevanje glede na prevajalnik

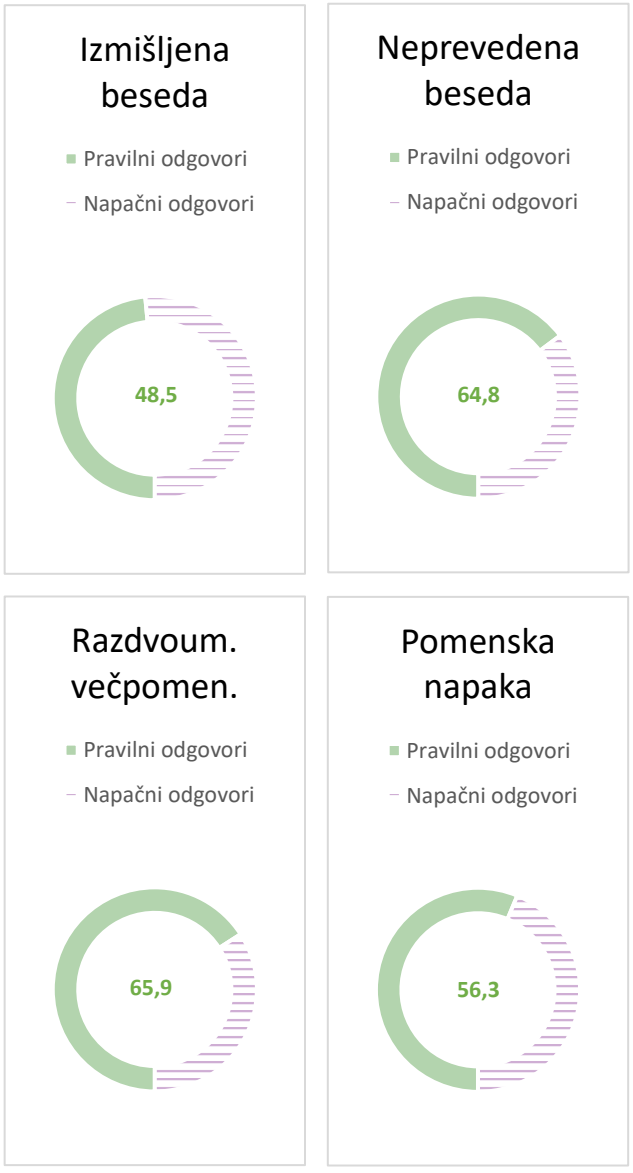
Odgovori na vprašanja, vezana na prevajalnik Google Translate so bili pravilni v 51,3 % primerov oz. 739 od 1440 odgovorov. Prevajalnik eTranslation je pokazal boljše rezultate, delež pravilnih odgovorov je znašal 66,6 %.

[3] Denimo, prevod za rob zaslona (ang. *bezel*, je prevajalnik prevedel kot »bezela«).

4.3. Razumevanje glede na tip napake

V vprašalniku so bili vključeni štirje tipi različnih napak. V alinejah nizam tip napake in odstotek pravilnih odgovorov:

- izmišljena beseda: 48,5 %,
- neprevedena beseda: 64,8 %,
- napačno razdvoumljene večpomenske besede: 65,9 %,
- hujša pomenska napaka: 56,3 %.



Slika 2: Diagrami 1–4. Rezultati glede na tip napake v %.

4.4. Razumevanje glede na kontekst

V naslednjem segmentu predstavljam delež pravilnih odgovorov vezanih na kontekst.

- Izključno besedilni: 60,4 %,
- besedilni in slikovni; slika ne vpliva na razumevanje: 44 %,
- besedilni in slikovni; slika vpliva na razumevanje: 69,8 %,
- izbor ene izmed več predlaganih slik glede na to, kaj piše v besedilu: 64,2 %.

4.5. Razumevanje glede na tip odgovora

V tem segmentu predstavljam rezultate glede na način izbora odgovora. Primarna funkcija te analize je preveriti konsistenco oz. morebitna odstopanja npr.; če so odgovori odprtega tipa, kjer anketiranci v prazno vnosno polje vnesejo poljuben odgovor, bistveno slabši kot tisti, kjer imajo na voljo denimo štiri predlagane odgovore, izberejo pa enega.

- Odgovor odprtega tipa (vnosno polje): 36,3 %,
- odgovor zaprtega tipa (A, B, C ali D): 60,8 %,
- izbor z razlago (A ali B, zakaj?): 68,3 %.

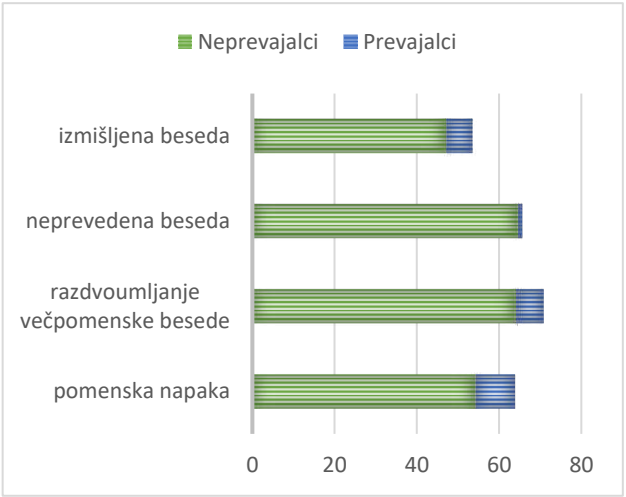
Slabši rezultat pri odgovorih zaprtega tipa je potrebno jemati z rezervo, saj so bili primeri s tako vrsto odgovora zgolj štirje. Samo določanje pravilnosti odgovora je pri takih primerih težje, osebno pa sem bil strog ocenjevalec, saj sem vse odgovore, ki niso bili popolnoma pravilni, označil za napačne.

4.6. Skupina prevajalcev

Edini demografski podatek, ki sem ga zbiral, je, ali se oseba, ki odgovarja na vprašalnik, ukvarja s prevajanjem. Pritrdilno je odgovorilo 24 udeležencev od 120. Pri teh osebah sem analiziral odgovore glede na vrsto napake in jih primerjal z neprevajalci. Nasploh so bili njihovi rezultati za 6 % boljši (63,7 %), po kategorijah pa:

- izmišljena beseda 53,5 % (+ 6,3 %),
- neprevedena beseda 65,6 % (+ 1 %),
- razdvoumljanje večpomenske besede 70,8 % (+ 6,7 %),
- pomenska napaka 63,9 % (+ 9,6 %).

Ostalih demografskih podatkov nisem zbiral, kar je ena od slabosti raziskave. V primeru da bi podatki sovpadali z mojo predpostavko, da niso relevantni, jih ne bi vključil, sedaj pa preprosto nimam podatkov, na katerih bi lahko utemeljil svojo odločitev.



Graf 1: Rezultati skupine prevajalcev proti ostalim.

5. Razprava

Pri pregledu rezultatov sem ugotovil, da povprečna stopnja razumevanja znaša 59 %. Od vseh 2880 odgovorov je bilo 1697 pravilnih.

Na tej točki je potrebno izpostaviti primer št. 6, ki je bil nasploh najslabše razumljen in je znižal povprečje rezultatov v vseh kategorijah, v katerih se je nahajal. Daljša razlaga z razčlembo je na voljo v celotni raziskavi (Bordon, 2021).

Izvirnik	Prevod
En zmagovalec bo prejel grafično kartico GeForce RTX 2080 Ti Cyberpunk 2077 Edition. Vstop v predavanje je enostaven: 1. Prijavite se na forume ali ustvarite forumski račun . 2. Komentirajte to temo (BREZ CITIRANJA TE POSTAJE) in nam povejte, kaj želite narediti najbolj v Cyberpunku 2077. 3. Za potrditev vpisa vpišite svoje uporabniško ime v naš pripomoček za oddajo. KAKO VSTOPITI: Če želite vstopiti, vnesite **mednopni vložek** in sledite navodilom za vstop v nagradne igrače.	One winner will receive the GeForce RTX 2080 Ti Cyberpunk 2077 Edition graphics card. Entering the giveaway is easy: Sign in to the forums or create a forum account. Comment on this thread (WITHOUT QUOTING THIS POST) and tell us what you want to do most in Cyberpunk 2077. Sign your username in our giveaway widget to confirm your entry. HOW TO ENTER: To enter, submit **your entry during the Sweepstakes Period** and follow the directions to enter the Sweepstakes.

Tabela 3: Primer št. 6; »Mednopni vložek.«

eTranslation je bil v povprečju za 15 % boljši od prevajalnika Google Translate, v katerem je bil omenjen primer. Nasploh pa je eTranslation kazal boljše rezultate. Najboljši rezultati glede na tip napake so bili vezani na razdvoumljanje besednega pomena (65,9 %), kar kaže, da znamo ljudje nasploh dobro razbrati pomen iz sobesedila, na drugem mestu pa so bile neprevedene besede (64,8 %), kar lahko pripišemo dobremu znanju angleščine med udeleženci v anketi.

Rezultati so bili slabši, ko je prevajalnik napravil hujšo pomensko napako, ki je oteževala razumevanje celotnega segmenta (56,3 %), daleč najslabše rezultate pa je bilo moč opaziti v kategoriji izmišljena beseda (48, 5 %), v kateri je sicer bil prej omenjeni primer št. 6.

Glede na tip konteksta so bili najboljši rezultati pri primerih, kjer je slika vplivala na razumevanje (69,8 %) in kjer so udeleženci morali izbrati sliko, na katero se je nanašalo besedilo (64,2 %). Rezultati so bili nekoliko slabši v izključno tekstovnem kontekstu (60,4 %), najslabši rezultati pa so bili v kategoriji, kjer je bila besedilu priložena slika, ki ne vpliva na razumevanje oz. potencialno zmede udeleženca (44 %) – v tej kategoriji je bil tudi primer št. 6. Izkazalo se je, da slikovni kontekst, ki lahko potencialno vpliva na razumevanje besedilnega segmenta, pri strojnih prevodih v realnih okoliščinah, torej na spletu, z vsem pomožnim gradivom, igra pomembno vlogo.

Udeleženci, ki se sicer ukvarjajo s prevajanjem, so na splošno odgovarjali boljše od povprečja. Njihov delež uspešnosti je bil največji v kategoriji hujša pomenska napaka (+ 9,6 %), kar kaže na to, da zaradi »poklicne deformacije« bolj učinkovito razumejo kontekst.

6. Zaključek

V članku sem predstavil raziskavo o razumljivost nerevidiranih strojno prevedenih spletnih besedil pri končnih uporabnikih, ki niso bili posebej obveščeni, da prebirajo strojne prevode. Razumevanje besedilnih segmentov, ki so vključevali štiri različne tipe napak, ki nastanejo pri strojnem prevajanju NMT-sistemov, sem preverjal z anketo. Ta je vsebovala strojne prevode splošnih besedil, ki sem jih prevedel s prevajalnikoma Google Translate in eTranslation. Besedila so bila nerevidirana, vsebovala so napake, ki so bile predstavljene v več različnih kontekstih, bodisi s slikovnim gradivom bodisi brez.

Rezultati so pokazali, da je splošna stopnja razumevanja 59 %, pri čemer se je izkazalo, da so prevodi eTranslationa nasploh razumljivejši od prevodov Googlovega prevajalnika. Število pravilnih odgovorov je bilo najvišje v kategoriji razdvoumljanja večpomenskih besed, kar nakazuje na to, da ljudje lažje razumemo pomen strojnih prevodov, če nam je dan kontekst. Pri tem je bilo najbolj učinkovito slikovno gradivo, s katerim so si lahko udeleženci v raziskavi pomagali razjasniti pomen določenega besedilnega segmenta. Druga najuspešnejša kategorija je bila razumevanje neprevedenih besed, kar pomeni, da je bilo znanje angleškega jezika med udeleženci na visokem nivoju.

Po analizi se je izkazalo, da je bil nekoliko problematičen način izbire odgovorov, saj sem anketirancem naključno vnaprej določil, na kakšen način bodo odgovarjali. Odgovori odprtega tipa so kazali slabše rezultate kot izbirni odgovori in odgovori zaprtega tipa, toda zaradi majhnega števila vprašanj je težko izpeljati kakšen razumen zaključek. Podobno velja za samo metodo odgovarjanja na anketo, ki je bila pogojena pandemičnemu času. Za bolj relevantne rezultate bi bilo potrebno izvajati test razumljivosti v živo, na razpravljalen način. Enako velja tudi za vzorec sodelujočih – večji in bolj raznolik vzorec bi dal jasnejše rezultate.

V bodoče bi bilo zanimivo raziskati, če se razumevanje nerevidiranih strojno prevedenih besedil izboljšuje skupaj z nadgradnjami strojnih prevajalnikov, hkrati pa bi se lahko osredotočil še na avtomatsko generirana besedila in jezik spletnih robotov.

Menim, da bo v prihodnje nekoliko manj raziskav storilnosti pri popravljanju strojnih prevodov in veliko več raziskav, ki bodo vezane na razumljivost strojno prevedenih ali avtomatsko generiranih besedil v praktičnih situacijah. Končni bralec se vedno bolj pogosto srečuje s takimi besedili, lahko pa pričakujemo, da bo zaradi še dodatnih izboljšav strojnih prevajalnikov, novih metod in razširjenosti prakse tovrstnih potencialnih stikov med stroji in bralci brez vmesnega posega človeškega popravljalca vedno več.

7. Literatura

David Bordon. 2021. Razumevati nevronščino: Kako si ljudje razlagamo jezik strojnih prevajalnikov. Magistrsko delo. Univerza v Ljubljani. Dostop 30. 5. 2022. https://repozitorij.uni- lj.si/IzpisGradiva.php?id=125328.

Sheila Castilho in Ana Guerberof Arenas. 2018. Reading Comprehension of Machine Translation Output: What Makes for a Better Read?. V: Juan Antonio Perez-Ortiz, Felipe Sanchez-Martinez, Miquel Espla-Gomis, Maja Popovič, Celia Rico, Andre Martins, Joachim Van den Bogaert in Mikel L. Forcada, ur., *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, str. 79–88, Alacant, Španija. Dostop 30. 5. 2022. http://doras.dcu.ie/23071/.

Gregor Donaj in Mirjam Sepesy Maučec. 2018. Prehod iz statističnega strojnega prevajanja na prevajanje z nevronskimi omrežji za jezikovni par slovenščina-angleščina. V: *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2018*, str. 62–68, Ljubljana. Dostop 30. 5. 2022. http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Donaj-et-al_Prehod-iz-statisticnega-strojnega-prevajanja-na-prevajanje-z-nevronskimi-omrezji-za-jezikovni-par-slovenscina-anglescina.pdf.

Evropska komisija, 2020 European Language Industry Survey 2020 Before & After Covid-19. Dostop 30. 5. 2022. https://ec.europa.eu/info/sites/default/files/2019_language_industry_survey_report.pdf.

Philipp Koehn in Rebecca Knowles. 2017. Six challenges for neural machine translation. V: *Proceedings of the First Workshop on Neural Machine Translation*, str. 28–39. Association for Computational Linguistics, Vancouver, Canada. . Dostop 30. 5. 2022. https://arxiv.org/pdf/1706.03872.pdf.

Chiraag Lala in Lucia Specia. 2018. Multimodal Lexical Translation. V: *Proceedings of the 11th international conference on language resources and evaluation (LREC)*, str. 3810–3817. Miyazaki, Japonska: European Language Resources Association (ELRA). Dostop 30. 5. 2022. https://www.aclweb.org/anthology/L18-1602/.

Zsófia Lelner. 2022. »Machine Translation vs. Machine Translation Post-editing: Which One to Use and When?«. Dostop 30. 5. 2022. https://blog.memoq.com/machine-translation-vs.-machine-translation-post-editing-which-one-to-use-and-when.

Jiatong Liu. Multimodal Machine Translation. Dostop 30. 5. 2022. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9547270.

Lieve Macken in Iris Ghysele. 2018. Measuring Comprehension and User Perception of Neural Machine Translated Texts: A Pilot Study. V: *Translating and the Computer 40 (TC40): Proceedings*, str. 120–126. Geneva: Editions Tradulex. Dostop 30. 5. 2022. https://biblio.ugent.be/publication/8580951.

Lieve Macken, Laura Van Brussel in Joke Daems. 2019. NMT's wonderland where people turn into rabbits. A study on the comprehensibility of newly invented words in NMT output. V: *Computational Linguistics in the Netherlands Journal 9 (2019)*, str. 67–80. Dostop 30. 5. 2022. https://www.clinjournal.org/clinj/article/view/93.

Marianna J. Martindale in Marine Carpuat. 2018. Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT. Dostop 30. 5. 2022. https://arxiv.org/abs/1802.06041.

Maja Popović. 2020. Relations between comprehensibility and adequacy errors in machine translation output. V: Raquel Fernández in Tal Linzen. ur., *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL 2020)*, str. 256–264. Dostop 30. 5. 2022. https://aclanthology.org/2020.conll-1.19.pdf.

Rico Sennrich, Barry Haddow in Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. Dostop 30. 5. 2022. https://arxiv.org/abs/1508.07909.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia in Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. Dostop 30. 5. 2022. https://arxiv.org/abs/1911.12798.

Ngo Thi-Vinh, Thanh-Le Ha, Phuong-Thai Nguyen in Le-Minh Nguyen. 2019. Overcoming the Rare Word Problem for Low-Resource Language Pairs in Neural Machine Translation. V: *Proceedings of the 6th Workshop on Asian Translation*, str. 207–214. Association for Computational Linguistics. Hong Kong, Kitajska. Dostop 30. 5. 2022. https://arxiv.org/abs/1910.03467.

Diana Voroniak. Post-Editing of Machine Translation: Best Practices. Dostop 30. 5. 2022. https://blog.crowdin.com/2022/03/30/mt-post-editing/.

Dan Zdarek. Machine Translation Post-editing Best Practices. Dostop 30. 5. 2022. https://www.memsource.com/blog/post-editing-machine-translation-best-practices/.

Data Collection and Definition Annotation for Semantic Relation Extraction

Jasna Cindrič, Lara Kuhelj, Sara Sever, Živa Simonišek, Miha Šemen

Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana
jasna.cindric@gmail.com
larakuhelj@gmail.com
seversara@gmail.com
ziva.sim@gmail.com
miha.semen@gmail.com

Abstract

This paper presents the process of data collection, definition extraction and annotation for the purpose of semantic relation extraction based on English and Slovene texts related to geology, glaciology, and geomorphology. Automatic semantic relation extraction is an important task in NLP; its potential applications include information retrieval, information extraction, text summarization, machine translation, and question answering. This approach was based on the TermFrame project. The texts for the corpora were collected manually, while definitions were identified through targeted queries in SketchEngine and then semantically annotated using the WebAnno tool. Our research showed some significant differences between languages resulting in some difficulties during the annotation process.

1. Introduction

This paper describes the process of definition extraction, annotation and curation based on corpora created for a research project carried out by Master's students as part of the module Corpora and Localisation at the Department of Translation Studies, Faculty of Arts (University of Ljubljana). Translation students collaborated with their peers from the Faculty of Computer and Information Science (University of Ljubljana) on a project focusing on the automatic extraction of semantic relations, which required the creation of an English and a Slovene corpus and the provision of an additional data set annotated for semantic relations. We describe the process of corpus building, the identification and extraction of definitions, followed by the annotation and curation using the WebAnno annotation tool. Finally, the paper illustrates the results and obstacles as well as discusses possible further work and research.

Corpus-based automatic semantic relation extraction has become one of the main topics in corpus linguistics. Domain-specific annotated corpora are the basis for the design of many NLP systems for relation extraction (Thanopoulos et al., 2000) and are considered knowledge sources on natural language use. It is imperative to obtain corpora large enough to provide a sufficient number of instances of relation pairs for extraction (Huang et al., 2015). This is especially true for Slovene, a language with complex morphology and free word order, which currently lacks readily available large domain-specific corpora (Pollak et al., 2012).

The layout of the project relied heavily on a similar dataset, TermFrame[1] – a trilingual knowledge base that contains Karst terminology in English, Slovene and Croatian. The knowledge base was developed on the basis of the frame-based approach in terminology (Pollak et al., 2019; Vintar et al., 2021; Vintar and Stepišnik 2020; Vintar et al., 2019; Vrtovec et al., 2019), a cognitive approach to terminology that considers context, language and culture and focuses on specialised texts (Faber and Medina-Rull, 2017). Frame-based terminology is mainly used for the creation of multimodal specialised knowledge bases, where "frames" are used as a "representation that integrates various ways of combining semantic generalisations about one category or a group of categories" (Faber, 2015). Additionally, "templates" are used as a representation of parts of one category, and "templates" cover the cultural component (Faber, 2015).

Following the process of the TermFrame project, the team began with compiling an English and a Slovene domain-specific corpus, then extracting definitions and annotating them using the WebAnno tool (Castilho et al., 2016). This paper describes these steps in detail, followed by an analysis of the annotated definitions. It also highlights the obstacles the team faced during the conversion of texts and the annotation process.

The main goal of the project was to create an English and a Slovene corpus covering the fields of geomorphology, glaciology and geology, which would serve as a basis for definition extraction, annotation and curation.

2. Building the corpora

2.1. Text collection

For the purposes of our research, the linguist team compiled two corpora, one Slovene and one English. The entire project lasted for approximately one month.

The first step was to search for texts in both languages covering predefined topics, namely geology, glaciology, and geomorphology. These areas were chosen because they were semantically related to the domain of karstology, but had not yet been used in the TermFrame database. More specifically, the texts from neighbouring domains to karstology were assumed to contain the same semantic relations, so that our to-be-created data set could be fully compatible with the existing ones.

The linguist team was particularly interested in collecting scientific texts (scientific papers, articles, books,

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

doctoral and master's theses). Many of these texts can be found through the Digital Library of Slovenia[2] or through the Co-operative Online Bibliographic System & Services – COBISS[3], and through ResearchGate, a social networking site for scientists and researchers[4]. Ultimately, our team proposed 32 Slovene texts and 26 English texts as candidates. The proposed titles were validated by a domain expert and assessed as relevant.

The next step was to ensure that the texts were in a format that could be read by Sketch Engine[5], which proved difficult in some cases. Fortunately, most of the texts on dLib.si are available in TXT and PDF format. As a result, the team was able to access the texts in the appropriate format using Notepad. Texts that were suitable to the topic but could not be accessed in the correct format were omitted. Document conversion and text cleaning proved cumbersome (see Section 2.2). The team had one week to prepare the texts according to this process.

2.2. Creating the corpora

After collecting a sufficient amount of documents and successfully converting them into the appropriate formats, the team proceeded to create the corpora. As all team members had full access to Sketch Engine, we decided this would be the most efficient and straightforward tool for corpus creation and subsequent querying. Table 1 provides an overview and detailed information about both corpora.

	English	Slovene
Tokens	1,588,085	493,107
Words	1,284,564	358,731
Sentences	52,147	18,373
Documents	26	32

Table 1: Data on the English and Slovene corpus.

As can be seen from Table 1, the Slovene corpus was significantly smaller. This was due to the fact that longer Slovene texts were harder to find, which was to be expected, considering there are not as many Slovene sources as there are English ones.

As previously mentioned, arguably the most important challenge the team faced occurred after selecting the texts for the Slovene corpus. As most of them were in the form of PDF files, the team had to ensure they were searchable before converting them into text (TXT) files. Due to some language-specific characters, particularly diacritics, such as *č*, *š*, and *ž*, most of the widely available online converters failed to produce satisfactory results.

After a few unsuccessful attempts, we managed to convert them with Notepad++, but we still had to review the files and manually correct some errors before adding the documents to the corpus. Since the final text was corrected manually, man-made errors such as the inclusion of some elements, like the table of contents, English abstracts and reference lists that were unintentionally added to the final version of the corpus caused some difficulties when searching for potential definitions. Ultimately, it was impossible to rely entirely on conversion tools – this seemingly undemanding step required additional time and attention.

3. Definition extraction

In order to obtain the sentences containing definienda, definitors and genera, we had one week to extract the definitions from the corpora using targeted queries in SketchEngine. Searching for typical definition-like sentences can be done by searching for specific words or phrases and by CQL queries.

To some extent, the structure of definitions can be predicted. Typical definition structures in Slovene include "X je Y", "Y imenujemo (tudi) X". "izraz X pomeni Y", "izraz X označuje Y", "med Y štejemo (tudi) X" etc., while typical definition structures in English include "X means …", "X is a Y", "X is a kind of …", "The term X is …" or "X is defined as". (...) In this context, X is typically a hyponym and Y is a hypernym. Sketch Engine allows searching for such definitions in multiple ways. One method is to use a simple Sketch Engine query and search for words or phrases that are often included in the definitions, such as "imenujemo" or "izraz" in Slovene and "is a" or "is a term used to describe" in English. We were able to identify multiple definitions using this method, for example *"Tip kraškega površja, kjer je prevladujoča oblika vrtače, imenujemo vrtačasti kras."*

Another method is to use a CQL query in Sketch Engine and check for definitions with advanced filtering commands such as [tag="S.*"][word="je"][tag="S.*"] in Slovene or [tag="NN"][word="is"][word="a"]?[tag="N.*"] in English. This command combines a search for a specific part of speech (S.* – noun) and a specific word (je). An example of a definition identified by using the CQL query in Slovene is *"Uvala je večja kraška globel skledaste oblike z neravnim dnom in sklenjenim višjim obodom."* Another example in English is *"A coral reef is a ridge or mound built of the skeletal remains of generations of coral animals, upon which grow living coral polyps."*

Since not all definitions fit these typical structures, we used another strategy. We checked the keywords suggested by Sketch Engine and search for them with a simple query. In this way, we were able to identify various definitions which could not be found otherwise. An example of such a definition is *Slovenska kraška terminologija navaja, da je vrtača: depresijska oblika okroglaste oblike, navadno globoka več metrov in je bolj široka kot globoka*.

In addition to these strategies, the English team also utilised a glossary from the English corpus and extracted some of the definitions from there.

By combining all of these strategies, we were able to identify definition candidates suitable for annotation. The selected definitions were then verified by a terminology specialist. Some of the definitions were judged to be unsuitable, either due to their wording or for semantic reasons. After discarding the inadequate definitions, we retained 100 definitions from the Slovene corpus and 104

[2] https://www.dlib.si.

[3] https://www.cobiss.si.

[4] https://www.researchgate.net/.

[5] https://www.sketchengine.eu/.

definitions from the English corpus. All of them were then uploaded to WebAnno[6] to be manually annotated.

4. Definition annotation

The definitions were annotated using WebAnno – a web-based annotation tool, which allowed for a faster collaborative annotation process as well as a comparative evaluation of the annotations (Castilho et al., 2016). The annotation process took approximately ten days.

Altogether, the team annotated 100 Slovene and 104 English definitions, whereby four layers of information were considered. The layers were introduced to the linguist team by the course instructor and were, in term, selected because they had already been used in the TermFrame project (Vintar and Stepišnik, 2020). We believed that relying on the same categories that had already been adapted to karstology – a domain closely related to the ones chosen for this research – would ensure a straightforward annotation process with little to no ambiguities. Furthermore, the resulting data set would be fully compatible to the existing one in the TermFrame project. The layers of information include:

1. **Semantic category:** This layer covers the main semantic categories for **A. Landform** (A.1 Surface Landform, A.2 Underground Landform, A.3 Hydrological Landform or A.4 Other)**, B. Process** (B.1 Movement, B.2 Loss, B.3 Addition or B.4 Transformation)**, C. Geome, D. Element/Entity/Property** (D.1 Abiotic, D.2 Biotic, D.3 Property and D.3.1 Geolocation) and **E. Instrument/Method** (E.1 Instrument or E.2 Method). The semantic category was defined primarily for the definiendum and genus. Semantic categories are presented in Figure 1.
2. **Definition element:** Here, the term in question was marked as DEFINIENDUM, its hypernym or superordinate term as GENUS, the defining phrase (the phrase between the DEFINIENDUM and the GENUS – e.g. the phrase *is a*) as DEFINITOR and any of its hyponyms or subordinate terms as SPECIES.
3. **Semantic relation:** A set of 15 relations was used for annotating different features of the defined term: AFFECTS, HAS_ATTRIBUTE, HAS_CAUSE, CONTAINS, COMPOSITION_MEDIUM, DEFINED_AS, HAS_FORM, HAS_FUNCTION, HAS_LOCATION, MEASURES, HAS_POSITION, HAS_RESULT, HAS_SIZE, STUDIES and OCCURS_IN_TIME.
4. **Relation definitor:** This layer is associated with semantic relations and marks words or phrases that precede particular semantic relations (e.g. *in the* ocean).

WebAnno also offers an additional layer for the **canonical form**, which is used to ensure the full form of a term when it appears in an elliptic construction. The canonical form layer has been mostly used when annotating definitions in the Slovene corpus. One of the reasons for this is that ellipses are more common in Slovene. Another reason is

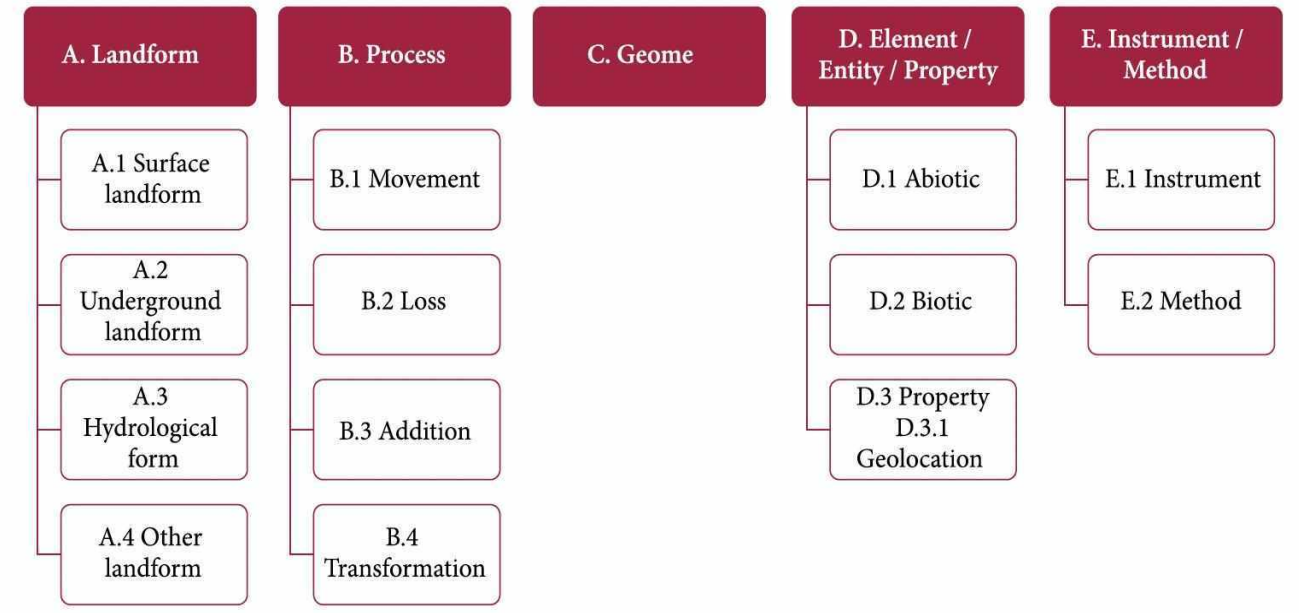


Figure 1: Semantic categories (Vintar and Stepišnik, 2021).

[6] https://www.clarin.si/webanno/login.html.

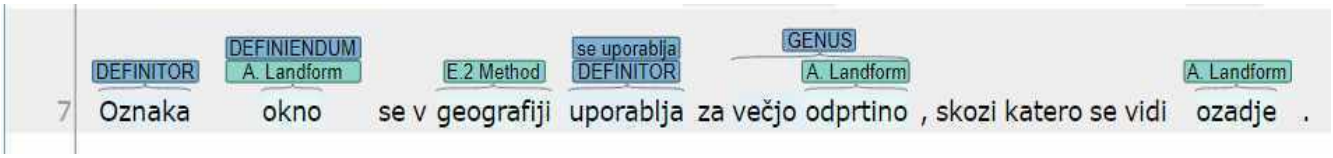


Figure 2: Use of the term canonical layer for pairing the words "uporablja" and "se" to show they form a single unit.

that the predicate and the pronoun "se" are often separated by other words.

As seen from Figure 2, which shows the example of the use of the term canonical layer in the Slovene corpus, the predicate "se uporablja" consists of two words that act as a definitor. Hence, the team used the term canonical layer to pair the two words together.

For the purpose of this project, three students annotated English definitions, while two students annotated the Slovene ones. Afterward, in the process of curation, both teams jointly annotated the definitions with the course instructor's assistance. We observed that the annotation of definition elements (definiendum, genus and definitor) was the most straightforward, although the annotators' solutions still varied in some cases (See Figure 3). On the other hand, annotation of semantic categories, semantic relations and relation definitors proved to be more dubious since the annotations often differed from one another. When variations occurred, the team managed to resolve such dilemmas through discussions.

As Figure 3 shows, all three students who annotated English definitions chose "tephra" as the definiendum. Two students annotated the phrase "is a term covering" as the definitor and one student annotated only "is a term". The word "material" was determined to be a genus by two students, whereas one student extended the genus and annotated "pyroclastic material" – "pyroclastic" was later defined as COMPOSITION_MEDIUM.

5. Analysis

After annotating all of the extracted definitions, the linguist team wanted to take a closer look at the results. Each English definition had one definiendum, giving a total of 104 definienda, while the Slovene definitions had one or more definienda, 113 in total.

The most common definitor in English was "is a", followed by "are", and in Slovene "imenujemo" and "je".

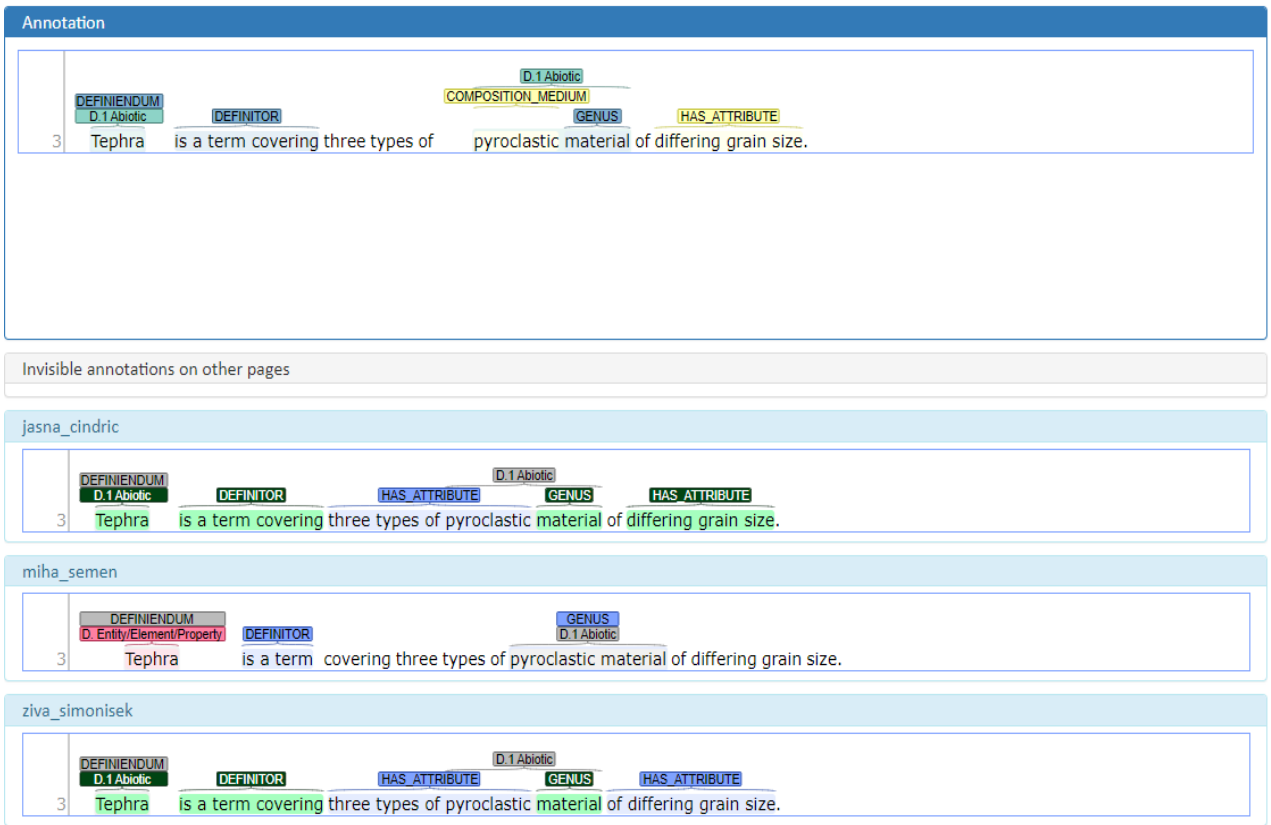


Figure 3: Curation process in WebAnno.

One or more genera were found in all English definitions, 112 in total, while not all Slovene definitions had a genus.

Figures 4 and 5 show the distribution of semantic categories for the annotated terms in Slovene and English. In total, 183 English and 334 Slovene terms were assigned categories. The most frequent category in English was D.1 Abiotic, followed by A.1 Surface landform. Similarly, A.1

Surface landform was the most frequent category in Slovene, followed by D.1 Abiotic.

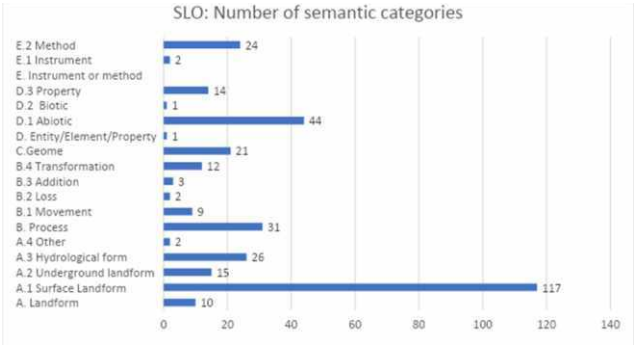


Figure 4: Semantic categories in the Slovene corpus.

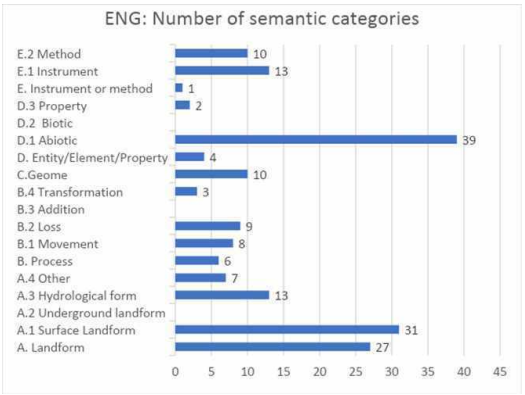


Figure 5: Semantic categories in the English corpus.

Figures 6 and 7 show the distribution of semantic relations for Slovene and English. A total of 186 relations were marked in English and 156 in Slovene. The most common relations in English were HAS_CAUSE (morphogenesis) and HAS_LOCATION (spatial distribution). On the other hand, the two most common relations in Slovene were HAS_FORM (morphography) and HAS_LOCATION (spatial distribution).

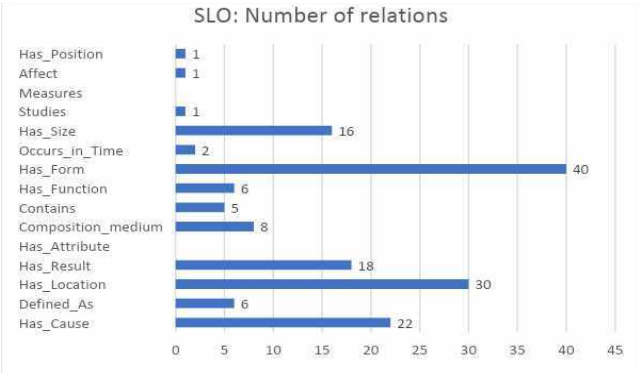


Figure 6: Number of semantic relations in the Slovene corpus.

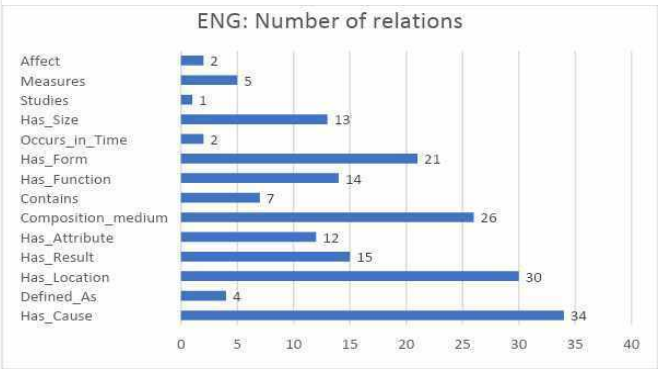


Figure 7: Number of semantic relations in the English corpus.

5.1. Annotation difficulties

During the annotation and curation process, the team encountered some complex cases, in particular when reviewing Slovene definitions, which required further discussion and careful attention. While annotating the definition element proved fairly straightforward, semantic relations posed some challenges.

The analysis showed ambiguities in 37 out of 65 sentences in the Slovene corpus. We have divided the ambiguities into the following categories.

5.1.1. Phrases that could be placed in multiple categories

The most recurring ambiguity concerned phrases that could be classified into a number of categories, while others were difficult to associate with any of the possible labels. In many cases, the team had to determine how the annotators would deal with these ambiguous words and establish agreement on a consistent annotation strategy.

For example, the phrase “kraški izviri” in Figure 8 could semantically be understood as a hydrological form, a surface form, an underground form or an abiotic.

As in the previous example, the word “obala” in Figure 9 can be understood as a hydrological form, a surface form, an abiotic or a geome.

Although the word “kras” is most likely understood as geome, depending on the context, it can also be understood as karstology, the study of karst. In line with the decision to annotate “geomorphology” as a method, “kras” could therefore be annotated as a method as well as shown in Figure 10.

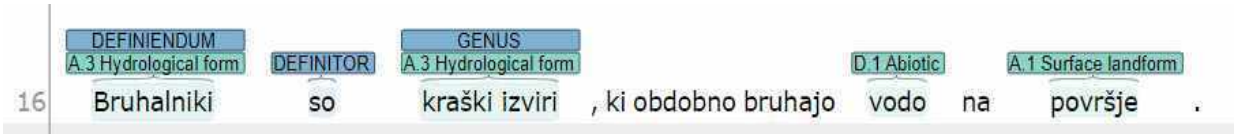


Figure 8: Example of an ambiguous annotation.

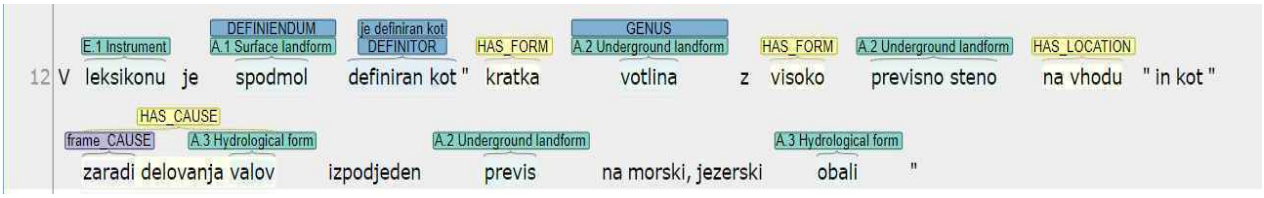


Figure 9: Example of an ambiguous annotation.

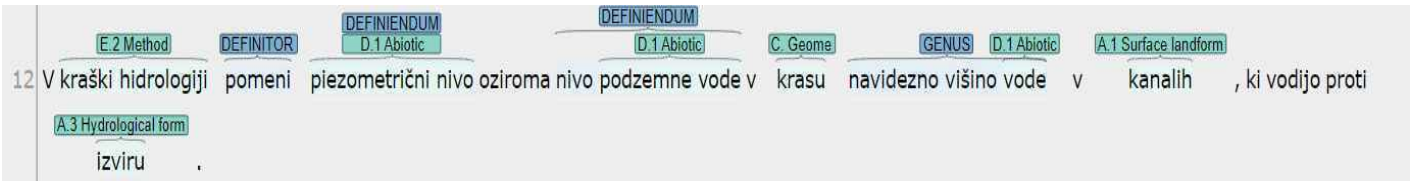


Figure 10: Example of an ambiguous annotation.

Another example was “gravitacija” (see Figure 11). It was extremely difficult to annotate a word denoting such a complex concept. In discussions with the course instructor, the team decided to annotate it as a method, as the names of the studies had to be annotated in the same way. However, it should be noted that the word could also be annotated according to other criteria.

and definiendum would share the same semantic category, since genus is a hypernym or superordinate term, but this was not the case for all definitions. For example, the definiendum “aquifer” was annotated as A.3 Hydrological form, but the genus “body of rock” was annotated as D.1 Abiotic in the same definition. This is because “body of rock” is not necessarily a hydrological form and can also be found on the surface. Another example is the definiendum “weathering”, which was annotated as B.4 Transformation

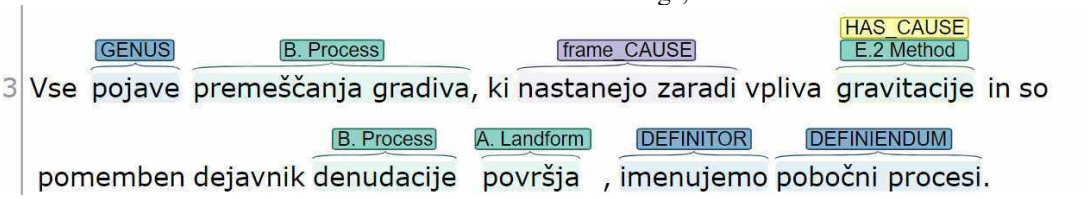


Figure 11: Example of an ambiguous annotation.

5.1.2. HAS_FORM

In a handful of cases annotating the Slovene definitions, it became clear that the semantic relation HAS_FORM manifests itself in different ways, as shown in Figures 12, 13 and 14.

Since HAS_FORM relations are more abstract and harder to grasp, annotation proved to be more difficult and required double-checking.

5.1.3. Annotation of genus

Sentences in the English corpus also posed some challenges, however their amount was significantly lower compared to their Slovene counterparts.

Before the annotation process, it was decided not to choose long phrases for the genus, but preferably just one word, e.g. “unloading of mountains” could be considered for the genus as a whole, but the team annotated only the word “unloading” as the genus. It was expected that genus and the genus “process” was annotated as B. Process. The reason for this is that “process” is a hypernym of “transformation”.

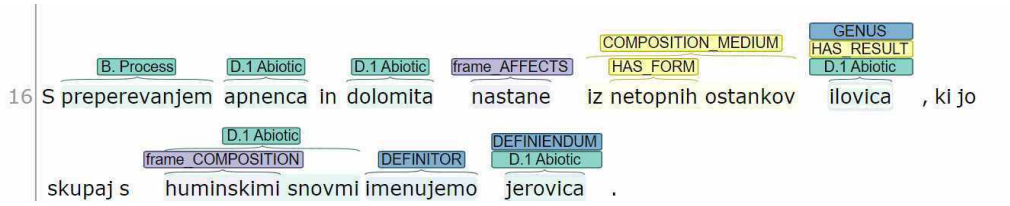


Figure 12: HAS_FORM introduced by a preposition.

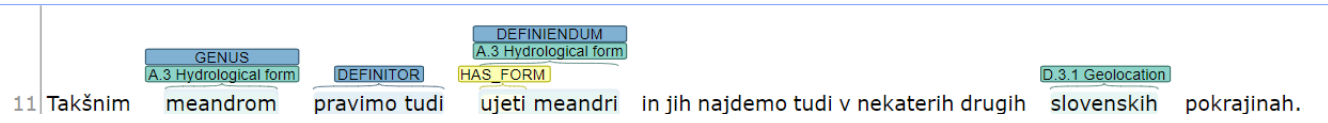


Figure 13: HAS_FORM expressed with an adjective.

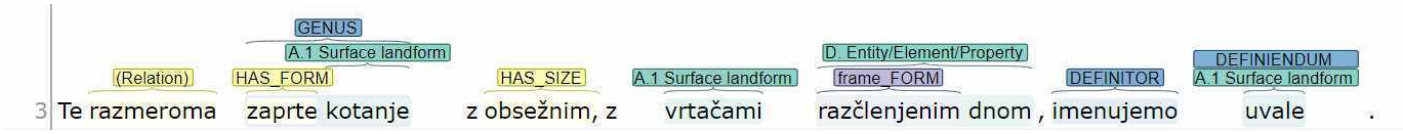


Figure 14: HAS_FORM expressed with an adjective (1) and introduced by a preposition (2).

6. Conclusion

This article describes the process of corpus creation and definition annotation for semantic relation extraction. When building corpora, linguists had to pay close attention to both the format and nature of the texts. The conversion of Slovene data proved to be quite challenging and required a great deal of attention to detail. It might be useful to develop a conversion tool specifically for language-specific characters, such as diacritics, to facilitate the study of data originating from languages, namely Slovene.

Definition extraction, on the other hand, did not pose any significant challenge.

In contrast, definition annotation followed by the curation entailed a great deal of debate and additional research. Since the team consisted only of linguists/translation students lacking domain-specific terminological knowledge, it was sometimes difficult to comment on the nature of the extracted terms. For any similar research endeavours, it could be useful to seek expert's input so as to facilitate the annotation process and prompt better results. Overall, definition elements were easier to identify and annotate than relation definitors and semantic categories and relations. The result of this work is a dataset with multi-layer semantic annotations in English and Slovene which can be used for future relation extraction experiments. It complements the TermFrame dataset and will be added to the Clarin.si repository.

The paper also draws attention to the differences between the two languages. English seems to favour shorter and more concise definitions, such as "is a" or "are", while Slovene tends to introduce longer structures, namely "imenujemo" and "se uporablja", and sometimes shorter ones, such as "je".

This research provides insight into the various language-specific barriers that arise when studying smaller languages that do not enjoy the same exposure and presence as widespread world languages such as English.

Further research could examine how definitions in both languages manifest themselves in different contexts and domains.

Large data collections serve as a basis for the development of tools for automatic semantic relation extraction. Semantic relation extraction can be used to create different computer applications that can make domain-specific knowledge more accessible, not only to experts but to the general public as well. The corpora that were built during this project can be used for future creation of specialised knowledge bases on geology, geomorphology and glaciology.

7. References

Richard Eckart de Castilho, Chries Biemann, Iryna Gurevych, and Seid Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. In: *Proceedings of the CLARIN Annual Conference (CAC) 2014*, pages 4505–4512, Soesterberg, Netherlands.

Pamela Faber. 2015. Frames as a framework for Terminology. In: H. Kockaert and F. Steurs, (eds.) *Handbook of Terminology*, Vol. 1, pages 14–33. John Benjamins, Amsterdam/Philadelphia.

Pamela Faber and Laura Medina-Rull. 2017. Written in the Wind: Cultural Variation in Terminology. In: M. Gryviel (ed.) *Cognitive Approaches to Specialist Languages*, pages 419–442. Cambridge Scholars, Newcastle upon Tyne.

Chu-Ren Huang, Jia-Fei Hong, Wei-Yun Ma, and Petr Šimon. 2015. From Corpus to Grammar: Automatic Extraction of Grammatical Relations from Annotated Corpus. In T'sou & Kwong (eds.) *Journal of Chinese Linguistics Monograph Series*, Vol. 25, pages 192–221. Chinese University of Hong Kong Press, Hong Kong.

Senja Pollak, Andraž Repar, Matej Martinc, and Vid Podpečan. 2019. Karst exploration: extracting terms and definitions from karst domain corpus. In: *Proceedings of eLex 2019*, pages 934–956. Lexical Computing CZ, s.r.o., Brno

Senja Pollak, Anže Vavpetič, Janez Kranjc, Nada Lavrač, and Špela Vintar. 2012. NLP workflow for on-line definition extraction from English and Slovene text corpora. In: J. Jancsary (ed.) *Proceedings of KONVENS 2012 (Main track: oral presentations)*, Vol. 5, pages 53–60. ÖGAI, Vienna.

Aristomenis Thanopoulos, Nikos Fakotakis, and Georg Kokkinakis. 2000. Automatic Extraction of Semantic Relations from Specialized Corpora. In: *Coling 2000, 18th International Conference on Computational Linguistics*, Vol. 1, pages 836–842. Universität des Saarlandes, Saarbrücken.

Špela Vintar, Vid Podpečan, and Vid Ribič. 2021. Frame-based terminography: a multi-modal knowledge base for karstology. In: *Proceedings of eLex 2021*, pages 164–176. Lexical Computing CZ, s.r.o., Brno.

Špela Vintar, Amanda Saksida, Uroš Stepišnik, and Katarina Vrtovec. 2019 Modelling specialised knowledge with conceptual frames: the TermFrame approach to a structured visual domain representation. In: *Proceedings of eLex 2019*, pages 305–318. Lexical Computing CZ, s.r.o., Brno.

Špela Vintar and Uroš Stepišnik. 2020. TermFrame: A Systematic Approach to Karst Terminology. In: *Dela*, Vol. 54, pages 149–167. Znanstvena založba Filozofske fakultete Univerze v Ljubljani, Ljubljana. https://doi.org/10.4312/dela.54.149-167.

Katarina Vrtovec, Špela Vintar, Amanda Saksida, and Uroš Stepišnik. 2019. TermFrame: Knowledge frames in Karstology. In: *Proceedings of ToTh 2019*, pages 109–126. Presses Universitaires Savoie Mont Blanc, Chambéry

Serbo-Croatian Wikipedia Between Serbian and Croatian Wikipedia

Ružica Farmakovski,[*] Natalija Tomić[**]

*Faculty of Philology, University of Belgrade
Studentski trg 3, 11 000 Belgrade
ruzicamarinkovic12@gmail.com
**Faculty of Philology, University of Belgrade
Studentski trg 3, 11 000 Belgrade ntomic801@gmail.com

Abstract

In this paper, we try to establish the linguistic identity of the corpus of texts CLASSLAWIKI-sh (Serbo-Croatian Wikipedia), comparing it with the corpus of texts The CLASSLAWIKI-sr (Serbian Wikipedia) and the corpus of texts CLASSLAWIKI-hr (Croatian Wikipedia), that are available at CLARIN.SI, Slovene national consortium of the European research infrastructure CLARIN Wikipedia, i. e. we are trying to determine whether it is closer to the Serbian or Croatian language standard. For this comparison, we used as variables the distinguishing features between Serbian and Croatian described in grammars and manuals of Serbo-Croatian, Serbian and Croatian languages. We came to the conclusion that according to the basic characteristics (orthographic, most phonetic, and derivational morphology features), the CLASSLAWIKI-sh is closer to the CLASSLAWIKI-hr, and according to morphosyntactic, lexical, and semantic features it is closer to the CLASSLAWIKI-sr.

1. Introduction

Wikipedia is a free online encyclopedia launched in 2001 by a community of volunteers. It is available in 326 languages and it has more than 302,906 active editors and more than 101,868,334 registered users.[1] Its specificity is its editing system. It is open to its audience for writing and contributing different content. One of the languages with considerable content is Serbo-Croatian, a language that does not officially exist since the split of former Yugoslavia.

In recent decades linguistic research has increasingly been conducted on materials and data from the Internet. They are available to everyone, free and easy to use and there are plenty of them. This makes it suitable for linguistic research as well.

Wikipedia, along with Twitter and other similar sources, offers plenty of materials and data, but to use them at all, we need to know their true identity. That is how the phenomenon of linguistic identification (and automatic linguistic identification) is becoming increasingly important.

In this sense, discriminating between related languages, considered "as a sub-task in automatic language identification" (Tiedemann and Ljubešić, 2012: 2620), also gaining more and more attention from researchers.

But this is not an easy task, especially when it comes to related languages. Since they have a common origin, they share many grammatical features and lexemes, so it is often very difficult to distinguish between them. Therefore, for many researchers, this task is a special challenge, i. e. "both necessity and a challenge (Ljubešić and Klubička, 2014: 32).

We hope that our research, which is more linguistically oriented, will provide some useful linguistic data for automatic text recognition research. Also, we hope that we will show how important it is to choose the right and reliable features as variable for this type of research (based on corpus). For example, we had to drop one of the most important and stable features, a feature that is cited everywhere in the literature (*ko:tko*), because it poses a problem for corpus lemmatization (Section 5.2).

Our paper consists of 7 sections. In Section 2, we describe the goal and present the initial hypothesis. In Section 3, we present the genetic and historical relationship between the Serbian and Croatian standards. In Section 3, we describe two types of related works that we used. On the one hand, there are works related to linguistic identification or the discrimination between related languages, and on the other hand, there are works dealing with the differences between Serbian and Croatian. Section 5 deals with the methodology, where we list and describe the variables we used, and in Section 6, we present the data we have obtained from the corpus and their analysis. In Section 7, we present the conclusion and some suggestions for further research. Finally, in Section 8, we list the literature that we used and cited in the paper.

2. Goal of the paper

In this paper, our goal is to determine the linguistic identity of the corpus of texts CLASSLAWIKI-sh (Serbo-Croatian Wikipedia, hereinafter: SCW), that is available at CLARIN.SI, Slovene national consortium of the European research infrastructure CLARIN.[2] The CLASSLAWIKI-sr (Serbian Wikipedia, hereinafter: SW) and CLASSLAWIKI-hr (Croatian Wikipedia, hereinafter: CW) corpora can also be found here. When we compare the linguistic characteristics of our target corpus with the other two corpora, we hope to determine its linguistic identity, i. e. whether SCW is closer to SW or CW or if it is somewhere in the middle. In Figure 1, we show our hypothesis schematically. Our initial hypothesis is that SCW is somewhere in the middle between SW and CW, perhaps with a tendency towards SW, due to the larger

[1] https://www.wikipedia.org/

[2] https://www.clarin.si/kontext/corpora/corplist

number of its users, less resistance to the use of Serbo-Croatian resources, etc.



Figure 1: Is SCW closer to SW or CW or it is somewhere in the middle?

We also hope to get answers to some other related questions: Does SCW represent a language that existed in the former Yugoslavia under the name of Serbo-Croatian language? Is SCW a mixture of characteristics of Serbian and Croatian varieties? Or is SCW a mixture of Serbian and Croatian texts?

3. Serbo-Croatian vs. Serbian and Croatian

Without the desire (and possibility) to determine precisely whether Serbian and Croatian are two languages, one language with two names, two dialects, two varieties, or two standards, we will present in basic terms their historical relationship.

These two entities lived under the common name Serbo-Croatian language in the former Yugoslavia for almost a century and were considered one language. It is an open question of how much they mixed, how much they influenced each other and how many linguistic features passed from one entity to another, and how much each of them preserved their identity.

They undoubtedly have the same origin. Before the Slavs immigrated to the Balkans, the Southern Slavs separated from Eastern and Western Slavs. During historical development, the western linguistic community of the Southern Slavs developed, from which the Slovene and Serbo-Croatian languages developed. The Serbo-Croatian language consisted of three dialects – Štokavian, Kajkavian, and Chakavian, according to the interrogative pronoun: *što/šta:kaj:ča* (′what′). Until the 19th century, all three dialects were in use. The foundations of the new standard language were established in the 19th century. After the Illyrian movement and the reform of the language and orthographic system by Vuk Karadžić, the Štokavian dialect (ekavian and (i)jekavian variant) was taken as the basis of the standard language.

Even before the break-up of the former Yugoslavia, this language was polycentrically standardized, and the break-up of Yugoslavia practically created four new languages: Serbian, Croatian, Bosnian, and Montenegrin.

4. Related work

Our research is based on two types of sources. On the one hand, there are works related to linguistic identification or the discrimination between related languages, and on the other hand, there are works dealing with the differences between Serbian and Croatian.

4.1. Literature on linguistic identification and the discrimination between related languages

Martins and Silva (2005) start with a well-known *n*-gram-based algorithm "that measures similarity according to the prevalence of short letter sequences (*n*-gram)" (767), but they also add that linkage information and the text from hypertext anchors could improve overall results.

Padró and Padró (2004) presented and compared three different statistical methods for language identification: Markov Models, Trigram Frequency Vectors, and Gram Based Text Categorisation (mentioned as *n*-gram above). They concluded that "for texts over 500 characters, all the systems get a precision higher than 95%, and for texts of 5,000 characters the precision is higher than 99% with all systems" (161), but for the small texts Markov Model System has the highest precision. Also, all three systems tend to fail when it comes to the problem of distinguishing similar languages (Catalan and Spanish).

So we come to the paper of Ljubešić et al. (2007) dealing with the language identification problem of the Croatian language. To identify the Croatian language, authors have to distinguish it from similar languages – Serbian, Slovenian, or Slovak. They applied the method of most frequent words and combined it with the character *n*-gram models. Finally, to improve the precision of identifying Croatian documents (where the biggest problem was distinguishing them from Serbian documents), the authors made a list of forbidden words for Croatian and Serbian. Forbidden words (or "blacklisted words"') are words that occur often in one language but never in the other language. Forbidden words (or blacklisted words) are also used (along with a document classification method) in another article dealing with the problem of discrimination between closely related languages, or more precisely between Bosnian, Croatian and Serbian (Tiedemann and Ljubešić, 2012).

Zampieri and Gebrekidan (2012) also agree that methods for discrimination similar languages or varieties are not "substantially explored". In their article, they try to define a model for the automatic classification of two varieties of Portuguese: European and Brazilian. They state that these two varieties "are considered to be the same language [although] there are substantial differences between European and Brazilian Portuguese in terms of phonetics, syntax, lexicon, and orthography" (235). Although they recognize the problem with similar entities, they use the character-based model using 4-grams. It is practically a standard character *n*-gram model, just with larger character *n*-grams.

This group of works is more mathematically oriented and does not deal with linguistic features like our work.

4.2. Literature on the differences between Serbian and Croatian

As we said at the beginning of this section, another group of papers is dealing with the differences between Serbian and Croatian. Among them, we paid special attention to two papers, whose methodology was also used for our examination – Ljubešić et al. (2018) and Ljubešić et al. (2019).[3] Namely, this group of authors states phonetic, morphological, syntactic, and lexical differences between Serbian and Croatian, which represent variables

[3] Both papers have the same authors.

through which a certain phenomenon is examined. In the first paper, it is the spatial distribution of 16 linguistic features and the question is, "do state borders correspond to linguistic boundaries". In the second paper it is the phenomenon of linguistic accommodation among the speakers of BCMS[4] languages, i. e. the question of whether BCMS speakers adapt their language when they are in contact with speakers of other BCMS languages (do they change their accent, some grammar construction, do they use specific lexemes, etc.).

This part also includes works that deal with differences in BCMS languages, but they are more descriptive, ie. differences do not represent methodological instruments for research. From Piper (2009) we learn more about the historical, social, political, and cultural circumstances of these two languages, and then follow the description of the language differences (537–552). Branko Tošović and Arno Wonisch are the editors of a series of collections of papers from 2009 to 2013 that also deal with the relationship of the BCMS languages in general (historical, social, political, and cultural perspectives), and then with many individual language problems – adjectival aspect, noun motion, nouns of nomina agentis type, distribution of future tenses, participial and reflexive passive, etc. (Tošović and Wonisch, 2009; 2010; 2012; 2013). In Ćevriz-Nišić (2009) we could find various phonological, derivational, lexical, and syntactic distinctive features between Serbian, Croatian, and Bosnian standard languages from administrative style. Article Badurina (2004) follows recent changes (late 20th century) in orthography and vocabulary; in Karavdić (2011) 16 syntactic differences are pointed out (apart from well-known *da*+present or an infinitive): possessive genitive and the adjective with noun, future 2nd or present tense, *kod*+accusative or *k*+dative, etc. In Bekavac et al. (2008) differences are organized on five levels, from phonological to semantic levels. The last one is especially interesting because it is rarely mentioned in the literature. Authors state lexeme *čas* meaning ′one moment′ in Croatian and ′one hour′ in Serbian, lexeme *persons* translated in Serbian by ′lica′ and in Croatian by ′osobe′, etc.[5]

We also consulted the most relevant grammars and manuals of the Serbian and Croatian languages, and for certain variables some special papers dealing with them. For more linguistic details of these, but also of the all listed literature units in this section, see Section 5.

All papers in this second group, except for the second of the two papers that we highlighted at the beginning of Section 4.2. (Ljubešić et al. (2019)), state the differences between Serbian and Croatian, without examining them in the corpus. Ljubešić et al. (2019) use a corpus, but it is about shorter texts (Twitter), and for a different purpose – to describe the phenomenon of linguistic accommodation. Also, our choice of variables differs from the variables used in this paper (see explanation in Section 5.2).

5. Methodology

5.1. Data and metadata

In the Introduction, we defined Wikipedia as a free online encyclopedia. But it is not entirely, nor could it be, the subject of linguistic inquiry. The subject of our research are three special corpora composed of texts from Wikipedia. These three corpora is, as we stated in Section 2: CLASSLAWIKI-sh, CLASSLAWIKI-sr, and CLASSLAWIKI-hr, available at CLARIN.SI, Slovene national consortium of the European research infrastructure CLARIN. All free corpora are part of the project CLASSLA Wikipedia which involved generating corpora for seven south-Slavic languages: Macedonian, Bulgarian, Serbian, Croatian, Serbo-Croatian, Slovene, and Bosnian. The corpora were generated using Wikipedia dumps that were downloaded on October 17th, 2020.[6]

Some important metadata for our three corpora is given in Table 1.

Corpus	Documents	Tokens	Words
CLASSLAWIKI-sh (Serbo-Croatian Wikipedia corpus CLASSLAWIKI-sh 1.0)	453,404	80,669,281	63,541,966
CLASSLAWIKI-sr (Serbian Wikipedia corpus CLASSLAWIKI-sr 1.0)	639,277	122,530,226	97,258,485
CLASSLAWIKI-hr (Croatian Wikipedia corpus CLASSLAWIKI-hr 1.0)	205,898	66,484,380	51,719,524

Table 1: Number of documents, tokens, and words in SCW, SW, and CW.

5.2. Variables of interest

To select the appropriate variables, we reviewed the linguistic differences between Serbian and Croatian that are cited in the literature. As we have already said, we used Ljubešić et al. (2018) and Ljubešić et al. (2019) the most because we followed the methodology applied in these works. Then we reviewed basic grammars and manuals for Serbian, Croatian and Serbo-Croatian: Pešikan et al. (2010), Stevanović (1989), Stanojčić and Popović (2008), Piper and Klajn (2013), Ivić et al. (2004), Mrazović and Vukadinović (2009); Barić et al (1997). Then we reviewed papers whose main topic was these differences. All these sources are described in Section 4.2. We also used papers that deal with a particular variable as a special problem. These sources are mentioned in the variable in question.

[4] Bosnian, Croatian, Montenegrin, and Serbian languages. In the literature dealing with these languages, they are referred to as BCMS languages.

[5] Lexeme *persons* can also be translated into Serbian by ′osobe′; the translation ′lica′ appears in an administrative language.

[6] Links to Wikipedia Dumps can be found on https://github.com/clarinsi/classla-wikipedia

First, we had to choose a smaller number of variables. So we tried to make the variables meet the following criteria: linguistic relevance, representing stable differences, easy recognition by the speaker, and easy automatic retrieval. Therefore, we rejected unreliable variables (such as script – Cyrilic or Latin; in addition, the texts in all corpora are in Latin script), underdeveloped variables, and variables that are impossible to process due to homonymy.

For most variables, we selected words that illustrate a certain phenomenon so we could search the corpus. We chose examples that are well known to us as native speakers and for which we found confirmation in the literature mentioned above.[7] It would be better if we could present all those examples in tables, along with their mean values and proportions. But since that would require a lot of space, we decided to just list those words and present the final analysis in Section 6.

Two variables were extracted using regular expressions – morphosyntactic variable trebati and lexical variable da li:je li.

In three cases (for the pair of words *takođe:također* (′also′) – in phonetic variables; for the semantic variable čas (′hour′, ′moment′); and for the pronoun *ko:tko*) we analyzed a smaller number of examples (80). We did this in cases where something seemed suspicious to us based on the raw numbers (*takođe:također, ko:tko)* or when we wanted to get a general impression of the use of the lexeme, and a detailed analysis would require separate research (*čas*).[8] More examples and better-randomized examples would improve this research.

The selected variables belong to the following levels of linguistic structure: orthographic, phonetic, derivational morphology, morphosyntactic, syntactic, and semantic levels.

We chose this approach, to start from known and described language features in the literature and then identify them in the corpora because we believe that this is the best way of language identification. In addition, we believe that automatic text recognition should be based on theory.

Orthographic variable

1) transliteration:original

When it comes to the orthography of foreign proper names, transliteration is more frequent in Serbian (and it is also a standard) and in Croatian foreign proper names are written in original: *Njujork:New York*. Examples of this variable are found in Memić (2009).

Phonetic variables

2) e:ije/je

It concerns the Proto-Slavic vowel *jat* and its different reflexes: *je/ije* in Croatian and *e* in Serbian, although the (i)jekavian reflexes (and dialects) also belong to the Serbian standard language.

In the literature, this variable is considered "the most obvious difference between Croatian and Bosnian on one side and Serbian on the other" (Bekavac et al., 2008:35) or as one of "the biggest differences between Croatian and Serbian" (Ljubešić and Klubička, 2014:29) or "one of the features central to defining the dialects" and as "the variable whose geographical distribution is expected to be most straightforward" (Ljubešić et al., 2018: 110).

This variable was extracted through a list of words that was created manually (as we have already mentioned). Since the consonant *j* is a frequent cause of various phonetic alternations, we chose words in which there are no phonetic alternations. Otherwise, we would have to look for more results for the (i)jekavian forms and to sum them up: *sneg:snijeg, snjeg* (′snow′), *devojka:djevojka, đevojka* (′girl′), etc.

3) rdrop

The variable rdrop refers to the fact that in some words in Croatian consonant *r* is kept at the end of the word, and in Serbian it is lost: *juče:jučer* (′yesterday′).

This variable is also illustrated by a list of words that is created manually.

The nouns *veče:večer* (′evening′) are regularly cited as an illustration of this difference, but since both nouns have the same declension, we had to exclude it from the search because we can not deduce from the form what the lemma should be. We kept the words *naveče:navečer*, *predveče:predvečer* and *uveče:uvečer* (′in the evening′), that are derived from the word *veče:večer* because they are adverbs, so they have no declension.

Since the grapheme *đ* also appears as *dj*, for words *takođe:također* (′also′) we searched for both occurrences and summed them up (*takođe:također, takodje:takodjer*).

4) h:k

The variable h:k occurs in words of Greek origin. As early as the middle age, the rule was established in Serbian that Greek χ was transferred as Slavic *h*, while in Croatian *k* appeared under the influence of Western European languages.

We also used a manually created list for this variable because there are not so many of those words.

Derivational morphology variables

5) ka:ica

The suffixes *-ka* and *-ica* are used for deriving feminine nouns of nomina agentis type. But here the situation is not so simple. First, both suffixes are very productive in both Serbian and Croatian, and we can not claim that one suffix is Serbian and the other is Croatian. So we have in Serbian: *glumica, igračica, pevačica* etc., and in Croatian: *maserka, programerka, novinarka, analitičarka* etc. This also applies to other suffixes. So we find in Babić (1999) that suffixes *-ica, -ka, -kinja, -inja* are as Croatian as Serbian, and differ only in the distribution. We find the similar claim in other authors (Dražić and Vojnović, 2010).

Second, "the choice of the suffix also depends on the ending of the masculine noun from which the feminine form is derived" (Ljubešić et al., 2018: 113). Therefore, among many other suffixes, we chose the suffixes *-ar* and *-or* in the masculine gender, for which we found confirmation in several sources that they regularly give -

[7] The dictionary Ćirilov (2010) also helped us in this.

[8] See more details in those examples.

ka in Serbian and *-ica* in Croatian (Dražić and Vojnović, 2010; Ljubešić et al., 2018; Ćorić, 2010). We also manually created a list of those pairs of words.

6) isa, ova:ira

This variable is related to the morphological composition of the international verbs: *organizovati* in Serbian and *organizirati* in Croatian (′organize′). Petar Skok noticed that difference in the 1950s. According to Skok (1955–1956) suffix *-isati* is related to Belgrade and it is of Greek origin and it entered Serbian with Turkisms. The suffix *-irati* is related to Zagreb, it is of Latin origin, and it was received through French and German. The suffix *-ovati* originates from the Proto-Slavic language. Recent research also confirms this distribution: "It is also noticeable that the distribution of suffixes in certain verbs in Serbian and Croatian is differentiated […] examples of verbs with *-ira-* are registered in Croatian texts, and with *-isa-* and *-ova-* in texts by Serbian authors. " (Ivanić and Perišić, 2018: 188).

This variable is illustrated by a list of examples mostly listed in Tošović (2010), Skok (1955–1956), and Ivanić and Perišić (2018).

Morphosyntactic variable

7) trebati

In standard Serbian, the modal verb *trebati* (′need/should′) is used as an impersonal verb and has a complement *da*+present tense: *ja treba da idem, ti treba da ideš*, etc.[9] In Croatian, this verb is used as a personal verb and has an infinitive as a complement: *ja trebam ići, ti trebaš ići*, etc. For this variable, we used the regular expression found in Ljubešić et al. (2018).

Lexical variable

8) da li:je li

As we read in Ljubešić et al. (2018) *yes/no* questions in Serbian are used with interrogative expressions *da li* and *je li*. Form *da li* is more common and form *je li* is usually shortened to *je l'*, *jel'*, or *jel*. In Croatian *je li* is the standard form.

We have analyzed only full forms using regular expressions also found in Ljubešić et al. (2018): '\bda li\b' and '\bje li\b'.

Semantic variable

9) čas (′hour′: ′moment′)

Semantic differences are less common in the literature. We have already stated lexeme *čas* meaning ′one moment′ in Croatian and ′one hour′ in Serbian in Bekavac et al. (2008). Since it is a matter of meaning, we had to make our own decisions on a case-by-case basis. So we took the first 80 occurrences of the lexeme *čas* and determined whether it means ′hour′ or ′moment′.

After describing the variables used, we will only briefly mention at the end one of the very interesting problems we encountered, and that is the use of the interrogative pronoun *who*, which in Serbian has the form *ko*, and in Croatian *tko*. The first problem is that the forms *ko*, in addition to the forms *tko*, also received the lemma *tko* in all three corpora (*da je bilo kome rekao* – the form *kome* got the lemma *tko* instead of *ko*). Another problem is that the personal interrogative pronoun *ko/tko* has the same declension as the adjective pronoun *koji/tkoji* (its shorter form). In this way, many examples that were supposed to get the lemma *koji/tkoji* got the lemma *ko/tko* (*kamen od koga se obično izrađuje nakit* – the form *koga* got the lemma *tko* instead of *koji*). That is why we rejected this feature as a variable, but we analyzed 80 examples with the lemma *ko* and 80 examples with the lemma *tko* in each of the three corpora. Then we divided those examples into lemmas that they should get: *ko, tko, (t)koji*. The results we obtained are shown in Table 2.

	CLASSLA Wiki-sr Serbian Wikipedia	CLASSLA Wiki-hr Croatian Wikipedia	CLASSLA Wiki-sh Serbo-Croatian Wikipedia
Lemma=ko (80 examples)	ko: 49	-	-
	tko: 0	-	-
	(t)koji: 29	-	-
	error: 2	error: 10	error: 32
Lemma=tko (80 examples)	ko: 4	ko: 9	ko: 1
	tko: 1	tko: 41	tko: 3
	(t)koji: 71	(t)koji: 24	(t)koji: 71
	error: 4	error: 6	error: 5

Table 2: Lemmatization of the pronoun *ko/tko*.

6. Analysis

Insight into these three corpora gave us the following data. For the variables we searched using the word lists we made, we got the number of lemmas. To obtain representative values and overcome the size inequality of these three corpora, we calculated mean values and proportions. To calculate the proportion, we used the following formula: the proportion of one value of one variable in one corpus is equal to the quotient of the mean value of that variable value in that corpus and the sum of the mean values of both values of that variable in that corpus. For example, the proportion for the value *e* of the variable e:(i)je in SW = the mean for *e* in SW / (the mean for *e* in SW + the mean for *(i)je* in SW).

To visually represent these relationships, for each variable we made the same illustration. On the left (blue) is what we have defined as a Serbian feature, and on the right (red) what we have defined as a Croatian feature. Then we marked a value for each corpus. We presented the proportions as percentages because it seems easier to read the data from the image in this way. This presentation allowed us to see data for all three corpora for each variable in the same image, making it easier to compare. The figure also shows whether SCW is closer to SW or CW.

Our first variable is orthographic and it concerns the writing of foreign proper names. As we said, transliteration is more frequent in Serbian, and in Croatian foreign proper names are written in the original. To examine this we took 5 proper names: Njujork:New York, Čikago:Chicago, Dablin:Dublin, Kembridž:Cambridge,

[9] In colloquial language this verb is very often used as a personal verb, but retains the complement *da*+present tense: *ja trebam da idem, ti trebaš da ideš*, etc.

Venecija:Venezia. As we can see from the mean values and proportions, transliteration is more prevalent in SW (0.74), original writing in CW (0.80), and SCW is closer to CW in this characteristic. The proportion is 0.68 in favour of the original writing.



Figure 2: Variable transliteration:original.

The next three variables are phonetic. For the first e:ije/je, we took 10 words, according to the criteria defined above for this variable: cvet:cvijet (′flower′), reč:riječ (′word′), sveća:svijeća (′candle′), zameniti:zamijeniti (′replace′), uvek:uvijek (′always′), pesma:pjesma (′song′), vetar:vjetar (′wind′), mera:mjera (′measure′), veštica:vještica (′witch′), sesti:sjesti (′sit′). Mean values and proportions show us the following. Although the (i)jekavian dialect also belongs to the Serbian standard, in SW ekavian reflex is completely dominant (0.99). In CW the (i)jekavian reflex of the Proto-Slavic vowel has the same value (0.99), which is not surprising, because there is only one standard in Croatian. In SCW the ekavian reflex occupies approximately one-third and the (i)jekavian 2 thirds (the proportion is 0.30:0.70).

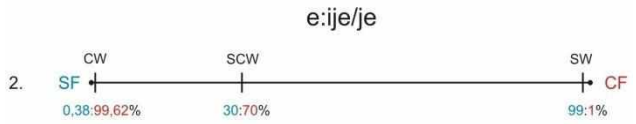


Figure 3: Variable e:ije/je.

The next phonetic variable refers to words that have a consonant *r* at the end of the word in Croatian and in Serbian it is lost. We used the following 6 words: juče:jučer (′yesterday′), prekjuče:prekjučer (′the day before yesterday′), naveče:navečer (′in the evening′), predveče:predvečer (′in the evening′), uveče:uvečer (′in the evening′), takođe:također (′also′). Analysing these words, we came to the following results. Forms without the consonant *r* at the end of the word have the expected high value in SW (0.99), as do forms with the consonant *r* at the end of the word in CW (0.99). What we did not expect is an extremely high value of the form with the consonant *r* at the end of the word in SCW (0.99). Looking at the raw numbers, we concluded that the frequency of use of the form *također* in SCW contributed to this. If we exclude this pair of words (*takođe:također*) from the analysis, the characteristic forms almost retain their values in SW and CW (0.98 and 0.98), but SCW is much more balanced (0.48:0.52 in favour of forms with the consonant *r*). We also wanted to make sure that these high values for the word *također* are not the result of a lemmatization error. We reviewed 80 examples in SCW and found 16 errors (*Brown je takođe hvalio film, On takođe uzima učešća...*). In Figure 4 we show the values that include the use of the pair of words *takođe:također*.

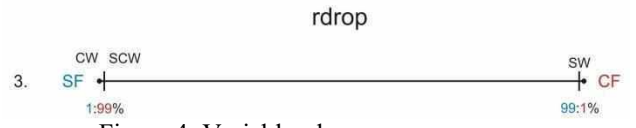


Figure 4: Variable rdrop.

The last phonetic variable h:k is found in translations of words of Greek origin – *h* in Serbian and *k* in Croatian. We used the following 7 words: haos:kaos (′chaos′), harizma:karizma (′charizma′), hemija:kemija (′chemistry′), hirurg:kirurg (′surgeon′), hronika:kronika (′chronicle′), hlor:klor (′chlorine′), hrizantema:krizantema (′chrysanthemum′). For example, we did not find the word *harizma* in CW at all, and the word *hrizantema* in CW nor SCW. This feature is very stable – words with *h* consistently appear in SW (0.99), and words with *k* consistently occur in CW (0.99). In SCW usage is balanced (0.50:0.50).

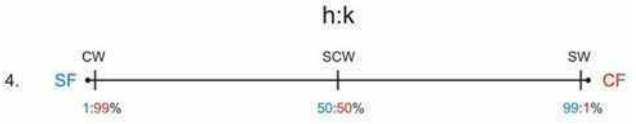


Figure 5: Variable h:k.

For our first derivational morphology variable ka:ica we used 9 words: slikarka: slikarica (′painter′, *fem*), ministarka:ministrica (′minister′, *fem*), apotekarka:apotekarica (′pharmacist′, *fem*), autorka:autorica (′author′, *fem*), doktorka:doktorica (′doctor′, *fem*), profesorka:profesorica (′professor′, *fem*), direktorka:direktorica (′director′, *fem*), lektorka:lektorica (′language editor′, *fem*), inspektorka:inspektorica (′inspektor′, *fem*). The data of the distribution of the suffixes *-ka* and *-ica* show the following. The suffix *-ka* in SW has a very high value (0.97), which confirms its consistent use in Serbian texts, just as the suffix *-ica* has a high value in CW (0.99). In SCW the suffix *-ka* reaches almost one-third (0.28), and the rest is the suffix *-ca* (0.72), which makes SCW much closer to CW according to this feature.

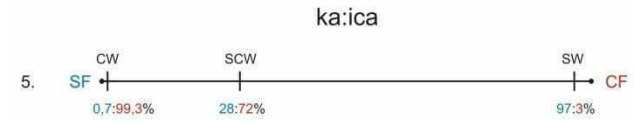


Figure 6: Variable ka:ica.

The situation is similar with verb formation. The suffixes *-isa* and *-ova*, which are related to Serbian, have a value of 0.99 in SW, the same as the suffix *-ira* in CW. In SCW, the ratio is 0.39:0.61 in favour of the suffix *-ira*, which also shows that SCW is closer to CW according to this feature. We used the 10 verbs: operisati:operirati (′operate′), fotografisati:fotografirati (′take photos′), reformisati:reformirati (′reform′), regulisati:regulirat (′regulate′), pakovati:pakirati (′pack′), kritikovati:kritizirati (′criticise′), diskutovati:diskutirati (′discuss′), identifikovati:identificirati (′identify′), promovisati:promovirati (′promote′). In SCW we did not find the form *pakirati* (′pack′), and in CW we did not find

the forms *fotografirati* (′take photos′) and *reformirati* (′reform′).



Figure 7: Variable isa, ova:ira.

Analysis of the morphosyntactic variable trebati showed that the modal verb *trebati* (′need/should′) as an impersonal verb with a complement *da*+present tense in SW has a dominant use (0.96), as does its personal variant with an infinitive as a complement in CW (0.88). In SCW this verb is used more in the impersonal form, which means that according to this feature SCW is more Serbian than Croatian (0.70:0.30)

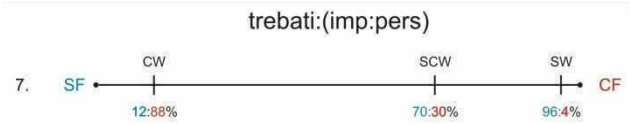


Figure 8: Variable trebati (imp:pers).

The lexical variable da li:je li represents the expressions *da li* and *je li* used for *yes/no* questions. In the description of the variable, we said that both expressions are used in Serbian, but that the form *da li* is more common in, and that the form *je li* is the standard form in Croatian. However, the results show the dominant use of *da li* in Serbian (0.98),[10] while in Croatian the use of these expressions is much more balanced – both values are close to the middle (0.46:0.54 – *je li* still has a bit more frequent use). In SCW, *da li* appears much more often (0.83:0.17), so it is closer to SW in this respect.



Figure 9: Variable da li:je li.

The semantic variable čas is stable. The lexeme *čas* is more often used in SW in the meaning of *hour* (0.90), and in CW in the meaning of *moment* (0.97). In SCW these meanings stand in relation 0.63:0.37 in favour of the meaning of *hour*, and therefore SCW is closer to SW according to this feature.

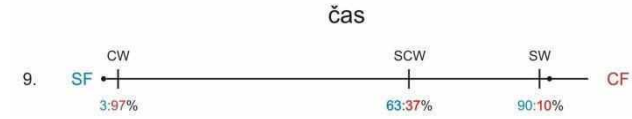


Figure 10: Variable čas.

7. Conclusion

Int the beginning, we determined that our goal was to determine the linguistic identity of the corpus of texts CLASSLAWIKI-sh and we assumed that it is midway between the corpus CLASSLAWIKI-sr and the corpus CLASSLAWIKI-hr. But we did not get a single or simple answer.

It turned out that according to orthography, most phonetic and derivational morphology features SCW is closer to CW than to SW. On the other hand, the morphosyntactic, lexical, and semantic features show that SCW is closer to SW than to CW. This may indicate that SCW contains more Croatian texts because these, so to speak, basic characteristics are more Croatian. Also, the values in SCW for most variables are closer to the extremes than they are balanced, so our initial hypothesis is confirmed in only a few cases (for example, variable h:k – 0.50:0.50). The other questions we asked at the beginning are not easy to answer in such a limited study.

To improve this research and get more accurate and precise results, some variables should be included, some unclear issues should be resolved (some problems in lemmatization), and some more advanced corpus search techniques should be used (first of all, regular expressions, randomized examples, etc.). As for the variables, there are a number of very interesting features: possessive adjective (in Serbian) / possessive genitive (in Croatian): *tetka Marin brat* / *brat tetke Mare* (′Aunt Mary's brother′); the conjunction *pošto* (′since′) – in Croatian it is used only in a temporal sense, in Serbian and in a causative sense: *Pošto je knjiga bila skupa, nisam je kupila* (′Since the book was expensive, I didn't buy it′); *kod* (in Serbian) / *k* (in Croatian): *Doći ću kod tebe.* / *Doći ću k tebi*. (′I will come to you.′); *gde* (in Serbian) / *kamo* (in Croatian) for the direction of movement: *Gde ideš?* / *Kamo ideš?* (′Where are you going?′), etc.

8. References

Božo Bekavac, Sanja Seljan, and Ivana Simeon. 2008. Corpus-based Comparison of Contemporary Croatian, Serbian and Bosnian. In: *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages*, pages 34–39, Dubrovnik, Croatia.

Božo Ćorić. 2010. Jezičke i/ili varijantske razlike na tvorbenom planu. In: Branko Tošović and Arno Wonisch, eds., *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, Book I/2*, pages 41–50. Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

Branko Tošović and Arno Wonisch, eds., 2009. *Bošnjački pogledi na odnose između bosanskog, hrvatskog i srpskog jezika.* Graz and Sarajevo: Institut für Slawistik der Karl-Franzens-Universität Graz and Institut za jezik.

Branko Tošović. 2010. Деривационные различия между сербским, хорватским и бошняцким языкам (прелиминариум). In: Branko Tošović and Arno Wonisch, eds., *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, Book I/2*, pages 65–80. Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

[10] The explanation for such a high value of *da li* in relation to *je li* in SW is that in the Serbian spoken language the full form *je li* is rarely used. Its shortened variants *je l'*, *jel'*, or *jel* are much more common.

Branko Tošović and Arno Wonisch, eds., 2010. *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, I/2.* Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

Branko Tošović and Arno Wonisch, eds.,. 2012. *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, I/4.* Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

Branko Tošović and Arno Wonisch, eds., 2013. *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, I/5.* Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

Bruno Martins and Mário J. Silva. 2005. Language Identification in Web Pages. In: *Proceedings of the 2005 ACM symposium on Applied computing, SAC '05*, pages 764–768, New York, NY, USA.

Eugenija Barić, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević, and Marija Zninka. 1997. *Hrvatska gramatika*. Zagreb: Školska knjiga.

Jasmina Dražić and Jelena Vojinović. 2010. Imenice tipa nomina agentis u srpskom i hrvatskom jeziku (tvorbeni i semantički aspekt). In: Branko Tošović and Arno Wonisch, eds., *Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika, Book I/2*, pages 41–50. Graz and Belgrade: Institut für Slawistik der Karl-Franzens-Universität Graz and Beogradska knjiga.

Jovan Ćirilov. 2010. *Hrvatsko-srpski rječnik inačica u Српско-хрватски речник варијаната.* Novi Sad:Prometej.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In: *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India.

Lada Badurina. 2004. Novije promjene u hrvatskome standardnom jeziku. *Croatian Studies Review,* 3–4:83–93

Marcos Zampieri and Binyam Gebrekidan. 2012. Automatic Identification of Language Varieties: The Case of Portuguese. In: Jeremy Jancsary, ed., *Proceedings of KONVENS 2012*, pages 233–237, ÖGAI. Main track: poster presentations.

Mihailo Stevanović. 1989. *Savremeni srpskohrvatski jezik*. Beograd: Naučna knjiga.

Mirela Ivanić and Jelena Perišić. 2018. Derivacija glagola sa osnovama stranog porekla u srpskom jeziku u svetlu (ne)jasne diferencijacije između srpskog i hrvatskog standarda. In: *Družbeni in politični procesi v sodobnih slovanskih kulturah, jezikih in literaturah*, pages 177–190.

Mitar Pešikan, Jovan Jerković, and Mato Pižurica. 2010. *Pravopis srpskoga jezika.* Novi Sad: Matica srpska.

Muntsa Padró and Lluis Padró. 2004. Comparing methods for language identification. *Procesamiento del Lenguaje Natural*, 33:155–162.

Nenad Memić. 2009. O prenošenju austrijskih i njemačkih toponima u bosanski, hrvatski i srpski jezik: o problemu egzonima u savremenom jeziku. In: Branko Tošović and Arno Wonisch, eds., 2009. *Bošnjački pogledi na odnose između bosanskog, hrvatskog i srpskog jezika.* Graz and Sarajevo: Institut für Slawistik der Karl-Franzens-Universität Graz and Institut za jezik. University Computing Centre.

Nikola Ljubešić, Maja Miličević Petrović, and Tanja Samardžić. 2018. Borders and boundaries in Bosnian, Croatian, Montenegrin and Serbian: Twitter data to the rescue. *Journal of Linguistic Geography,* 6/2:100–124, Cambridge University Press.

Nikola Ljubešić, Maja Miličević Petrović, and Tanja Samardžić. 2019. Jezična akomodacija na Twitteru: Primjer Srbije. *Slavistična revija,* 67(1):87–106.

Nikola Ljubešić, Nives Mikelić, and Damir Boras. 2007. Language identification: how to distinguish similar languages? In: Vesna Lužar-Stiffler, and Vesna Hljuz Dobrić, eds., *Proceedings of the 29th International Conference on Information Technology Interfaces*, pages 541–546, Zagreb: SRCE.

Nikola Ljubešić and Filip Klubička. 2014. {bs, hr, sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In: *Proceeding of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014*, pages 29–35, Gothenburg, Sweden.

Pavica Mrazović and Zorka Vukadinović. 2009. *Gramatika srpskog jezika za strance.* Sremski Karlovci, Novi Sad: Izdavačka knjižarnica Zorana Stojanovića.

Pavle Ivić, Ivan Klajn, Mitar Pešikan, and Branislav Brborić. 2004. *Srpski jezički priručnik.* Beograd: Beogradska knjiga.

Petar Skok. 1955–1956. O sufiksima -isati, -irati i -ovati. *Jezik,* 4(2):36–43.

Predrag Piper. 2009. O prirodi gramatičkih razlika između srpskog i hrvatskog jezika. In: Predrag Piper, ed., *Južnoslovenski jezici: gramatičke strukture i funkcije*, pages 537–552. Beograd: Beogradska knjiga.

Predrag Piper and Ivan Klajn. 2013. *Normativna gramatika srpskog jezika.* Novi Sad: Matica srpska.

Stjepan Babić. 1999. Dva tvorbena normativna problema i njihova rješenja. *Jezik,* 66(3):104–112. https://docplayer.rs/191032196-Dva-tvorbena-normativna-problema-i-njihova-rješenja-stjepan-babić.html

Vera Ćevriz-Nišić. 2009. Razlikovne crte između srpskog, hrvatskog i bošnjačkog standardnojezičkog izraza. In: *Savremena proučavanja jezika i književnosti, Zbornik radova sa I naučnog skupa mladih filologa Srbije I (1)*, pages 373–383, Kragujevac: Impres.

Zenaida Karavdić. 2011. Komparativna sintaksa bosanskog, crnogorskog, hrvatskog i srpskog jezika. In: *Njegoševi dani 3, Zbornik radova*, 357–365, Nikšić: Univerzitet Crne Gore, Filozofski fakultet.

Živojin Stanojčić and Ljubomir Popović. 2008. *Gramatika srpskog jezika za gimnazije i srednje škole*. Beograd: Zavod za udžbenike.

Ocenjevanje uporabniško dodanih sopomenk v Slovarju sopomenk sodobne slovenščine – pilotna študija

Magdalena Gapsa*

* Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
magdalena.gapsa@ff.uni-lj.si

Povzetek

V prispevku opisujem prvi korak uporabniške raziskave, znotraj katere bodo različne strokovne skupine ocenjevalcev presojale o relevantnosti določenih uporabniško dodanih sopomenk v *Slovarju sopomenk sodobne slovenščine*. V okviru raziskave želim preveriti, ali se ocene strokovnjakov, kot so lektorji, prevajalci in učitelji, razlikujejo od ocen slovaropiscev ter kako enovite so ocene znotraj posamezne skupine. Osredotočam se na potek in rezultate prvega sklopa ocenjevanja, ki ga je kot testna množica izvedla skupina študentov. Ta korak je služil tudi kot preizkus navodil in izbranih orodij za lažje načrtovanje dela preostalih predvidenih skupin ocenjevalcev. Navajam ugotovitve glede relevantnosti sopomenskega gradiva po presoji študentske skupine, kjer so zlasti zanimive mejne kategorije »pogojno« sprejemljivega gradiva, sledijo identificirane šibke točke zasnovane raziskave ter rešitve, ki bodo vključene v nadaljnji potek ocenjevanja.

Evaluation of User-Added Synonyms in the Thesaurus of Modern Slovene – a Pilot Study

The paper describes the first step of a user research in which various expert groups of evaluators will assess the relevance of certain user-added synonyms in the Thesaurus of Modern Slovene. Part of the research is to check whether the evaluations of experts such as proofreaders, translators and teachers differ from those of lexicographers, and how consistent the assessments are within each group. The main focus is on the process and results of the first set of assessments carried out by a group of students as a test set. This step also served as a test of the instructions and tools chosen to facilitate the planning of the work of the remaining intended groups of evaluators. The results are then presented in terms of the relevance of the synonymous material assessed by the group of students, with the borderline categories of "conditionally" acceptable material being of particular interest, followed by the weaknesses identified in the research designed and the solutions and improvements that will be incorporated into the further assessment process.

1. Uvod

S pojavom digitalnega medija se na področju jezikoslovja in strojne obdelave naravnega jezika spreminjajo tako potrebe kot tudi priložnosti, ki se kažejo zlasti kot možnost avtomatiziranega (hitrejšega, enostavnejšega in cenejšega) posodabljanja jezikovnih podatkov in opisov, večja povezljivost med podatki različnih vrst, neomejen prostor za njihov prikaz ter vključevanje širše skupnosti v proces priprave slovarjev[1] itd. V prispevku se osredotočam na slednje, torej možnost doprinosa širše jezikovne skupnosti, natančneje možnost, da slovarski uporabniki dodajajo sopomensko gradivo v *Slovar sopomenk sodobne slovenščine*[2] (Arhar Holdt et al., 2018, v nadaljevanju tudi *Sopomenke*), s čimer je povezano vprašanje morebitne spremembe v pogledih na sopomenskost. Na podlagi uporabniškega gradiva je možno opazovanje, kako sopomenskost dojemajo in občutijo uporabniki, zlasti v razmerju do slovaropiscev, ki podajajo končne odločitve o vključitvi sopomenskega gradiva v referenčne jezikovne vire.

Prispevek temelji na raziskovalnem vprašanju iz doktorske disertacije z naslovom *Sopomenskost v Slovarju sopomenk sodobne slovenščine in izbranih različicah wordneta*,[3] o prispevku širše jezikovne skupnosti k pogledom na sopomenskost. V disertaciji predpostavljam, da se pogled strokovne oz. širše jezikovne skupnosti razlikuje od pogleda slovaropiscev, vendar ta potencialni drugačni pogled jezikovne skupnosti lahko bistveno pripomore pri gradnji novih oz. nadgradnji obstoječih jezikovnih virov. To hipotezo bom preverila z analizo ocen sopomenskosti izbranega nabora uporabniško dodanega gradiva, ki ga ocenjujejo različne strokovne skupine (naštete v nadaljevanju). Ocene bom najprej primerjala znotraj posameznih skupin, nato pa tudi med skupinami.

Namen prispevka je predstaviti izsledke prvega, testnega oz. pilotskega ocenjevanja uporabniško dodanih sopomenk, ki ga je izvedla skupina šestih študentov jeziko(slo)vnih smeri. Evalvacijska naloga, posredovana tej skupini, je imela dva glavna namena: (I) priprava gradiva za ocenjevanje uporabniških sopomenk, preizkus modela, orodij in navodil ter morebitne dopolnitve oz. prilagoditve le-teh ter (II) zbiranje povratnih informacij za načrtovanje nadaljnjega obsega in izvedbe ocenjevanja. Raziskava se deloma povezuje s projektom *Sopomenke in Kolokacije 2.0 – SoKol, Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL*,[4] ki ga med leti 2021 in 2022 financira Ministrstvo za kulturo Republike Slovenije. Glavni cilj projekta je prenovitev *Slovarja sopomenk sodobne slovenščine* ter *Kolokacijskega slovarja sodobne slovenščine*. Projekt je omogočil dostop do študentov jezikoslovja z dobrim poznavanjem *Slovarja sopomenk*

[1] V slovenskem prostoru se tematike dotika monografija *Slovar sodobne slovenščine: problemi in rešitve* (Gorjanc et al., 2017). Podrobneje o vlogi uporabnikov v procesu priprave slovarjev in načinih sodelovanja z njimi sta pisala npr. A. Abel in C. Meyer (2013).

[2] *Slovar sopomenk sodobne slovenščine*: https://viri.cjvt.si/sopomenke/slv/ (dostop: 6. 5. 2022).

[3] Doktorska disertacija nastaja v okviru raziskovalnega programa *Jezikovni viri in tehnologije za slovenski jezik* (številka programa P6-0411) in jo v letih 2019–2023 financira Javna agencija za raziskovalno dejavnost Republike Slovenije. Mentorira jo zn. sod. dr. Špela Arhar Holdt.

[4] Spletna stran projekta SoKol: https://www.cjvt.si/sokol/ (dostop: 3. 5. 2022).

sodobne slovenščine ter izkušnjami z označevanjem pomensko povezanih podatkov.

2. Opis vira in pregled področja

Slovar sopomenk sodobne slovenščine, ki ga je leta 2018 objavil Center za jezikovne vire in tehnologije Univerze v Ljubljani, je prvi primer novega slovaropisnega koncepta, t. i. odzivnega slovarja (Arhar Holdt et al., 2018). Njegova glavna značilnost je, da se slovar stalno odziva na spremembe v jeziku ter na potrebe uporabnikov. Za namene tega prispevka je najbolj pomembna značilnost ta, da je uporabnikom omogočeno sodelovanje v procesu nastajanja slovarja, saj se podatki spreminjajo glede na aktivnost ter pripombe skupnosti, hkrati pa ta lahko prispeva k čiščenju nerelevantnih ali napačnih podatkov.[5]

Množičenje za namene slovaropisja je sicer znana praksa. Množica, ki jo želimo vključiti v ocenjevanje, ne potrebuje posebnih predznanj ali izobrazbe, saj so tudi uporabniki jezika, ki niso strokovnjaki s področja, dovolj nadarjena, ustvarjalna in učinkovita skupina, ki zmore reševanje manj zahtevnih oz. bolj rutinskih nalog, strokovnjaki pa se lahko po vključitvi množičenja osredotočijo na kompleksnejše oz. bolj analitične naloge (Kosem et al., 2013, str. 46; Čibej et al., 2015, str. 70–71). Množičenje je lahko izredno učinkovito in zanesljivo – odgovori oz. ocene nestrokovne skupnosti se skorajda ne razlikujejo od zlatega standarda oz. odgovorov, ki so jih podali slovaropisci, kar je bilo že leta 2008 dokazano z uporabo orodja *Amazon Mechanical Turk* (AMT) (Snow et al., 2008, str. 257–258), še zlasti, če zagotovimo zadostno količino ocenjevalcev (prim. Nicolas et al., 2021). Na tej predpostavki temelji vključevanje skupnosti v razvoj *Slovarja sopomenk sodobne slovenščine*.

Presojanje o (ne)sopomenskosti besed je bilo v široko razumljenem digitalnem slovaropisju mnogokrat uporabljeno še zlasti ob (nad)gradnji in čiščenju različnih wordnetov, npr. ruskega, kjer ocenjevalci presojajo o (ne)pravilnostih ter sami sestavljajo in popravljajo sinsete (Braslavski et al., 2014) ali češkega, za katerega je bil razvit poenoten sistem prijavljanja napak, ki jih uporabniki odkrijejo, ob tem pa lahko tudi predlagajo popravek (Horák in Rambousek, 2018). V slovenskem prostoru pa so ocenjevalci s pomočjo orodja za množičenje sloWCrowd (Tavčar et al., 2012) presojali, ali so avtomatsko pridobljeni predlogi sopomenski in ali spadajo v predvideni sinset ter tako pomagali odpraviti napake v slovenskem wordnetu (Fišer et al., 2014). Ocene skupnosti, tudi presojanje o (ne)sopomenskosti besed, so uporabne tudi širše, npr. pri evalvaciji natančnosti vektorskih vložitev za povezane oz. sorodne besede (prim. Schnabel et al., 2015, str. 301–303)

Ob tem se odpira vprašanje, ali se s širjenjem skupine sodelujočih širi oz. spreminja tudi sam pogled na gradivo, ki (naj) ga slovar sopomenk prinaša. Slovaropisci za identifikacijo sopomenskosti sledijo vnaprej izbranim (včasih tudi dokaj strogim) jezikoslovnim izhodiščem, slovarski uporabniki pa lahko o sopomenskosti presojajo precej bolj subjektivno, in sicer z vidika "uporabnosti" ali "relevantnosti" predloga za njihovo delo (npr. predloge tipa *brat – sorojenec*, *avto – osebno vozilo* itn., so ocenjevalci večinoma opredelili kot relevantne, čeprav gre za drugo relacijo).[6] Potrebno je opozoriti, da presojanje o podobnosti dveh besed oz. sopomenskosti tudi za slovaropisce nikakor ni lahka in nedvoumna naloga, saj univerzalne definicije sopomenskosti ni, sam koncept pa je zelo širok in tesno povezan s kontekstom in okoliščinami rabe, hkrati pa ga različni raziskovalci drugače interpretirajo in opisujejo (gl. npr. Snoj, 2019, str. 13–41; Vidovič Muha, 2013, str. 172–183; Zorman, 2000, str. 20–48).

Večina definicij sopomenke opredeljuje kot besede, ki imajo identičen pomen, vendar različno formo (Zgusta, 1971, str. 89), poudarja se tudi razlikovanje besed z istim pomenom po njihovi stilni ali zvrstni vrednosti (Toporišič, 1992, str. 294). V literaturi prevladujeta dva glavna pogleda: sopomenke so le besede, ki imajo popolnoma isti pomen (popolna sopomenskost) ali pa besede, katerih pomen je zelo podoben (delna sopomenskost). Popolna sopomenskost je zelo redka, saj krši načelo jezikovne ekonomičnosti oz. gospodarnosti, pogosta pa je delna sopomenskost (prim. Hock, 1991, str. 283; Snoj et al., 2016, str. 5; Vidovič-Muha, 2013, str. 175; Zgusta, 1971, str. 89), ki se najpogosteje kaže v primeru prenesenih pomenov, izposojenk iz tujih jezikov, arhaizmov in ekspresivnega besedišča, največ sopomenk pa imajo besede, ki so rabljene prav v prenesenem ali s kolokacijami povezanem pomenu (Apresjan, 2000, str. 37). V slovenskem prostoru je delna sopomenskost razumljena kot del stilistike in ne semantike, zato je bila dolgo, še zlasti v slovaropisju, obravnavana predvsem popolna sopomenskost, sopomenke v SSKJ pa imajo tudi normativno vlogo (usmerjajo od zaznamovanega proti nezaznamovanemu), ob izidu *Sinonimnega slovarja slovenskega jezika* (2016) je bilo več pozornosti namenjene tudi delni sopomenskosti (Vidovič Muha, 2013, str. 180; Snoj et al., 2016, str. 6). *Slovar sopomenk sodobne slovenščine* nudi nov okvir, saj osvetljuje vlogo in vrednost konteksta s prikazom kolokacij ali povezavo na korpusne zglede, obenem pa nudi tudi možnost, da uporabniki dodajo gradivo, ki po njihovem mnenju v slovarju manjka. Na podlagi skoraj 1.000 uporabniško dodanih sopomenskih predlogov želim ponovno odpreti vprašanje razumevanja sopomenskosti in preveriti, ali se je le-to spremenilo z nastankom in razvojem digitalnih jezikovnih virov, zlasti odzivnega slovarja..

3. Predviden potek in izsledki raziskave

Cilj znotraj doktorske raziskave je, da evalvacijo poleg študentov opravijo tudi predstavniki drugih skupin sodelujočih: slovaropisci (kot najbolj specializirani strokovnjaki s področja), prevajalci, lektorji, učitelji slovenščine in ljubiteljski raziskovalci jezika brez jezikoslovne izobrazbe (kot predstavniki širše jezikovne skupnosti). Interesne skupine so bile določene na podlagi tipologije ciljnih skupin uporabniških raziskav (gl. Arhar Holdt, 2015, str. 142–146), kjer so slovarski uporabniki

[5] Preostale glavne značilnosti slovarja so npr. (a) je dostopen le v digitalni obliki ob upoštevanju potreb, pogojev in prednosti le-te, hkrati pa nikoli ni zaključen, saj se podatki stalno spreminjajo in prilagajajo trenutnemu jezikovnemu stanju, (b) slovarska baza nastaja z uporabo naprednih računalniških metod, kar uporabnikom hitro ponuja veliko količino odprto in prosto dostopnih jezikovnih podatkov, ki so relevantni, a še neprečiščeni, (c) sopomenski podatki so povezani z besedilnim kontekstom s pomočjo kolokacij, korpusnih zgledov in povezav na korpus ter (č) slovar in slovarska baza sta prosto in odprto dostopni pod ustrezno licenco (prim. Arhar Holdt et al., 2018, str. 404; Čibej in Arhar Holdt, 2019, str. 339–340).

[6] Predpostavljam, da se bodo v določenih kategorijah gradiva oz. odločitev pokazale skupne točke, v drugih pa razlike.

načeloma pripadniki (vsaj) ene skupine: (I) uporabniki, ki slovarje uporabljajo v procesu izobraževanja (npr. študenti in učitelji slovenščine),[7] (II) uporabniki, ki slovarje uporabljajo v poklicne namene (npr. slovaropisci, prevajalci, lektorji in učitelji) ter (III) uporabniki, ki slovarje uporabljajo za prostočasne aktivnosti (npr. ljubiteljski raziskovalci jezika).

Tipologija je bila uporabljena tudi v raziskavi odnosa uporabnikov do novosti v *Sopomenkah*, kjer so bile v rezultatih najbolj zastopane skupine lektorji, prevajalci, učitelji slovenščine na različnih ravneh izobraževanja, pisci različnih vrst besedil (npr. beletristika, strokovno in znanstveno pisanje, kreativno pisanje, novinarstvo, blogerstvo ipd.) ter ljubiteljski raziskovalci jezika (prim. Arhar Holdt, 2020, str. 477). Iz tega lahko sklepamo, da so te skupine najbolj zainteresirane za *Sopomenke* (in sopomenske podatke na splošno), hkrati so tudi relevantne in reprezentativne. Ker je za preverbo hipoteze potrebno tudi mnenje slovaropiscev, je skupina piscev[8] nadomeščena s slovaropisci.[9] Njihovi odgovori bodo analizirani znotraj skupine, hkrati pa bodo služili kot referenčna evalvacija sopomenskih parov, s katero bodo primerjani odgovori vseh ostalih skupin.

Na podlagi izsledkov ocenjevanja po skupinah in primerjav odgovorov med njimi želim pridobiti empirično podlago o željah in pričakovanjih uporabnikov, ki bo z aplikativnega vidika podlaga za pripravo smernic za uredniške protokole, ki bodo uporabljeni pri nadgradnjah slovarja. Z znanstvenega vidika pa bodo odgovori podlaga za definiranje sopomenskosti v luči odzivnih digitalnih jezikovnih virov. Prvi sklop uporabniške raziskave, ki so ga opravili študenti, pa je poleg prej omenjenih ciljev služil še preizkusu zasnove raziskave in odkrivanju šibkih točk ter zbiranju povratnih informacij za načrtovanje nadaljnjega obsega in izvedbe.

4. Gradivo in metoda

4.1. Gradivo

Uporabniška raziskava temelji na delu podatkov, ki so podatkovni vzorec za doktorat, in sicer seznamu 546 samostalnikov, ki se pojavijo tako v podatkovni bazi *Slovarja sopomenk sodobne slovenščine* (Krek et al., 2018) in *sloWNeta* (Fišer, 2015) kot tudi v *Leksikalni bazi za slovenščino* (Gantar et al., 2013) in v *Velikem slovensko-madžarskem slovarju*, kjer so samostalniki opremljeni z oznakami semantičnih tipov (Kosem in Pori, 2021).[10]

Za to raziskavo so bili podatki za doktorsko disertacijo dodatno opremljeni s podatki o uporabniško dodanih sopomenkah v *Sopomenkah* na podlagi internega izvoza podatkov iz 18. 11. 2021.[11] Tako sem pridobila seznam 307[12] iztočnic, ki imajo vsaj eno uporabniško dodano sopomenko oz. 976 sopomenskih parov. Nekateri uporabniški vnosi (68 parov) so vsebovali dodatna pojasnila in opombe v oklepajih, največkrat opombe o zaznamovanosti predloga, npr. *arheolog – žličkar (šalj., pog.)*, *bonbon – cuker (neknj.)*, *klient – kunt (nar.)*, *preteklost – prtljaga (ekspresivno)*, *stopnica – štenga (nižje pog.)*. Ker ocenjevalcem nisem želela sugerirati odgovorov, so bile tovrstne opombe odstranjene. Zabeleženih je bilo tudi 5 primerov, kjer so uporabniki znotraj oklepajev dodajali bolj kontekstualne razlage in ne kvalifikatorje. Ti primeri so bili brez sprememb vključeni v končni nabor za ocenjevanje, saj sem s tem želela preveriti reakcijo ocenjevalcev na tovrstne oznake. Gre za predloge: *interier – ambient (v zaprtem prostoru)*, *kmet – kmet (šahovska figura)*, *koncentracija – (velika/majhna) vsebnost*, *priloga – priponka (k e-pismu)* in *torbica – (torbica) pismo*. Z odstranitvijo opomb je prišlo do podvojevanja sopomenskih parov,[13] zaznala sem 4 take primere, ki so bili prav tako izločeni iz seznama za ocenjevanje. Ta je v končni fazi obsegal 972 parov.

4.2. Navodila ocenjevalcem

Ocenjevalci so prejeli nagovor s kratkim pojasnilom, da se ocenjevanje izvaja v okviru doktorske raziskave in kakšne podatke želim zbrati. Bili so naprošeni, naj med ocenjevanjem ne uporabljajo drugih jezikovnih virov in priročnikov. Navedeno je bilo, da je naloga sestavljena iz dveh obveznih delov: preglednice s sopomenskimi pari, kjer bodo podajali svoje odgovore in morebitne pripombe, ter vprašalnika, kjer bodo podajali demografske podatke o sebi ter povratne informacije o evalvacijski nalogi sami. V primeru dvomov so udeleženci lahko zastavljali dodatna vprašanja po e-pošti.

Glavno navodilo ocenjevalcem je bilo odgovoriti na vprašanje: »Ali sta besedi v paru sopomenki?«. Vsak sopomenski par so lahko uvrstili v eno izmed štirih kategorij, oz. izbrali enega izmed štirih možnih odgovorov, in sicer DA, NE, POGOJNO DA ter NISEM PREPRIČAN/NE VEM. Odgovor DA je bil predviden za primere, ko so bili prepričani, da sta besedi sopomenki, odgovor NE za primere, ko so bili prepričani, da besedi nista sopomenki ter kadar je šlo za očitne napake oz.

[7] Študenti, zlasti jezikovnih smeri, so na prehodu med izobraževanjem in poklicno rabo, podobno učitelji, ki slovarje uporabljajo v poklicne namene, vendar je njihov poklic vezan na izobraževalni proces.

[8] V primeru piscev bi bilo najtežje pridobiti koherentno skupino, ki bi pokrivala različne prej naštete žanre, po drugi strani preostale skupine zadoščajo potrebi po predstavnikih skupine, ki slovarje uporablja v poklicne namene.

[9] Treba se je zavedati, da so slovaropisci zaradi svoje izobrazbe in specializiranosti zelo atipična uporabniška skupina za slovarske raziskave (prim. Arhar Holdt, 2015, str. 140), vendar prav to na tem mestu služi namenu raziskave.

[10] Slednja dva vira sta upoštevana, saj želim v (preostalih) analizah v okviru doktorske disertacije upoštevati tudi korpusno osnovan pomenski opis (potencialno) sopomenskega gradiva.

[11] Slovarska baza *Sopomenk*, ki je dostopna v okviru repozitorija CLARIN.SI, ne vsebuje uporabniško dodanih sopomenk. Slednje izvozimo iz slovarskega vmesnika s pomočjo prilagojene skripte, kar omogoča, da so uporabniško dodani podatki ažurni.

[12] Ne razlikujem med iztočnicami z veliko in malo začetnico. V naboru gradiva za nalogo je to samo en primer, in sicer *zemlja* in *Zemlja*, ki je tukaj obravnavan kot ena iztočnica.

[13] Slovarski vmesnik ima sicer preprosto varovalko, ki uporabnikom preprečuje ponovni vnos že dodanega predloga, vendar temelji na prepoznavi znakov in dovoljuje vnos tako alfanumeričnih kot nealfanumeričnih znakov, npr. oklepajev. Ko uporabnik obstoječemu sopomenskemu predlogu doda opombo, sistem to prepozna kot nov vnos. V mojem vzorcu se je to zgodilo štirikrat, in sicer dvakrat znotraj gesla *babica*, kjer sta bila predloga *nona* in *nona (lokalno)* ter *oma* in *oma ;)*, znotraj gesla *živina*, kjer sta bili predlagani *živad* in *živad (star.)* ter znotraj gesla *nakup*, kjer sta bili predlagani *kupilo* in *kupilo (star.)*.

zatipkane besede. Odgovor POGOJNO DA je bil predviden za pare, ko so ocenjevalci sicer menili, da gre za sopomenki, vendar so hkrati videli tudi omejitve oz. imeli pomisleke, dvome, npr. da sta besedi sopomenki samo v določenem pomenu, kontekstu, ena ali obe besedi sta zaznamovani itn. Odgovor NISEM PREPRIČAN/NE VEM je bil predviden za pare, ko niso poznali ene ali obeh besed v paru, pomena ene ali obeh besed v paru ali niso bili prepričani, ali so težko podajali svoje mnenje. Pri vsakem paru je bila možnost dodajanja opomb, ki so bile zahtevane pri odgovoru POGOJNO DA, zaželene pri NISEM PREPRIČAN/NE VEM in možne pri drugih odgovorih.

Ker je eden glavnih ciljev raziskave preverjanje, kaj ocenjevalci razumejo kot relevantno sopomensko gradivo, so bila navodila, v izogib sugeriranju odgovorov, zelo splošna. Zato »sopomenka« ni natančneje definirana, možni odgovori so vsebovali le kratek opis, ne pa tudi primerov. Prav tako ni bilo navodil, kam umestiti mejne primere.

4.3. Ocenjevanje

Sopomenski pari so bili ocenjevalcem posredovani v obliki tabele, ki je bila dostopna kot Googlova preglednica.[14] Datoteka je bila sestavljena iz dveh listov. Prvi list je obsegal skrajšano verzijo navodil, da so jih ocenjevalci imeli vedno pri roki, drug list pa seznam 972 sopomenskih parov za ocenjevanje. V prvem stolpcu tabele je zaporedna številka para, v drugem iztočnica, v tretjem pa predlagana sopomenka, npr. *vonj – vzduh*, *stigma – brazda*, *reforma – sprememba*, *pošta – sporočila*, *dopust – vakance*. Celice v teh stolpcih so bile zaklenjene v izogib namernim in nenamernim spremembam podatkov. V četrtem stolpcu so ocenjevalci iz spustnega seznama (v izogib zatipkom) izbirali enega izmed štirih odgovorov. Zadnji, peti stolpec, je bil predviden za komentarje in opombe ocenjevalcev. To je tudi edini stolpec, kjer so lahko prosto vnašali podatke. Ocenjevalci so do podatkov dostopali po principu en ocenjevalec – ena preglednica, da odgovori drugih ocenjevalcev ne bi vplivali na posameznikove odločitve.

4.4. Vprašalnik

Ocenjevalci so dobili tudi povezavo do spletnega vprašalnika, ki je bil sestavni del ocenjevanja. Vprašalnik je bil sestavljen in dostopen v spletnem orodju za anketiranje 1ka.[15] V prvem delu vprašalnika so sodelujoči odgovarjali na vprašanja o sebi: starost, zaposlitveni status, izobrazba (jezikoslovna ali ne), načini, na katere se z jezikom ukvarjajo in glavna področja, ki so zanje najbolj pomembna v zvezi z jezikom. V drugem delu so odgovarjali na vprašanja, povezana z evalvacijo samo: koliko časa so potrebovali, ali so imeli pri reševanju kakršne koli težave, ali so bila navodila jasna in ali so pri njih kaj pogrešali. Vprašalnik je bil dostopen brez omejitev, ocenjevalci so si lahko vprašanja vnaprej ogledali, njihovi odgovori so se sproti shranjevali, da so lahko npr. najprej podali podatke o sebi, informacije o nalogi pa naknadno.

5. Rezultati

Skupina pilotskih ocenjevalcev je obsegala 6 študentov. Dostop do preglednic je bil študentom dodeljen 15. 2. 2022, ocenjevanje so lahko začeli takoj in si ga prilagodili glede na druge obveznosti. Prvi študent je obvestilo o zaključenem ocenjevanju podal 16. 2. 2022, zadnji pa 8. 3. 2022. Vse želene odgovore sem pridobila v roku treh tednov. Pridobljene podatke je mogoče razdeliti v dva glavna sklopa, in sicer ocene vzorca sopomenskih parov in odgovore, pridobljene z vprašalnikom.

5.1. Izsledki ocenjevanja

Vsi odgovori, ki so jih dali ocenjevalci, so bili združeni v tabele z uporabo programa MS Excel. Prva tabela je obsegala podatke o izbranem odgovoru (brez opomb), kar je omogočilo preverjanje ujemanja oz. enotnosti ocenjevalcev. V drugi tabeli so bile zabeležene podane opombe. Te so bile ročno pregledane in dodeljene v eno izmed kategorij, ki so se oblikovale med pregledovanjem: samo v določenem pomenu ali kontekstu, zaznamovano, neznana beseda ali pomen besede, nad- oz. podpomenka, razlaga ter drugo (npr. nepravilno črkovanje, pomenske nianse, druga medbesedna razmerja, nenavadne oblike besed neujemanje besednih vrst, redkost rabe itn.). V primerih, kadar so ocenjevalci opredelili tudi vrsto zaznamovanosti (npr. ljubkovalno, pogovorno, zastarelo itn.), so bili tudi ti podatki ohranjeni.[16] Številčni podatki o opombah so predstavljeni v tabeli 1. 914 parov je imelo pripisano vsaj eno izmed šestih kategorij opomb, 435 je imelo pripisani vsaj dve kategoriji, 75 parov vsaj tri kategorije in 3 pari so imeli pripisane štiri kategorije opomb. Tretja tabela je obsegala združene podatke o ujemanju oz. enotnosti ocenjevalcev ter o že kategoriziranih opombah.

Kategorija	Število
samo v določenem pomenu/kontekstu	406
zaznamovano	375
neznana beseda ali pomen besede	266
nad- oz. podpomenka	182
razlaga	65
drugo	122
skupaj opomb	1.416

Tabela 1: Številčna razporeditev kategorij opomb.

Popolnega ujemanja, kjer je vseh šest ocenjevalcev podalo isti odgovor, je bilo zelo malo: le 34 znotraj seznama 972 parov oz. približno 3,5 % celotnega nabora. 17 parov je vseh šest ocenjevalcev prepoznalo kot nedvomno sopomenske (6 odgovorov DA), 5 parov kot pogojno sopomenskih (6 odgovorov POGOJNO DA), 5 parov kot nedvomno nesopomenskih (6 odgovorov NE) ter 7 parov kot neznane oz. neopredeljive (6 odgovorov NISEM PREPRIČAN/NE VEM). Bistveno več je bilo večinskega ujemanja med ocenjevalci, kjer izstopa samo en odgovor. Takih parov je bilo skupno 132 oz. približno

[14] Googlova preglednica od ocenjevalcev ne zahteva posebne strojne opreme, hkrati pa se vneseni odgovori sproti shranjujejo, zato naloge ni bilo potrebno reševati brez prekinitev.

[15] Spletno orodje za anketiranje 1ka: https://www.1ka.si/ (dostop: 5. 5. 2022).

[16] Podrobnejša analiza dejanskih opomb in komentarjev, ki so jih podali ocenjevalci, presega okvirje in namen tega prispevka, je pa zagotovo relevantna in zanimiva, tudi z vidika razumevanja sopomenskosti, zato bo naslovljena v prihodnosti.

13,5 % nabora gradiva. V 50 primerih je 5 ocenjevalcev izbralo odgovor DA, v 46 POGOJNO DA, v 19 NE in v 17 NISEM PREPRIČAN/NE VEM.

Skupaj je bilo parov z visokim ujemanjem ocenjevalcev 166 oz. 17 % nabora gradiva. 67 parov (40 %) je bilo umeščenih v kategorijo DA, 51 parov (31 %) v kategorijo POGOJNO DA, 24 parov (14,5 %) v kategorijo NE, preostalih 24 parov (14,5 %) pa je bilo uvrščenih v kategorijo NISEM PREPRIČAN/NE VEM. Razporeditev ocen v štiri predvidene kategorije prikazuje Tabela 2.

Odgovor	Popolno ujemanje	Večinsko ujemanje	Skupaj
DA	17 parov	50 parov	67 parov
POGOJNO DA	5 parov	46 parov	51 parov
NE	5 parov	19 parov	24 parov
NISEM PREPRIČAN/NE VEM	7 parov	17 parov	24 parov
Skupaj	34 parov	132 parov	166 parov

Tabela 2: Številčna razporeditev odgovorov.

V primerih, kjer so se ocenjevalci strinjali (vsi so izbrali isti odgovor), je bilo skupno podanih 22 opomb za 15 sopomenskih parov. V 132 primerih, kjer so se ocenjevalci večinoma strinjali (izstopal je en odgovor), je bilo skupno za 109 parov podanih 158 opomb. V primeru opombe iz kategorije drugo so ocenjevalci največkrat navajali pomenske nianse, črkovanje oz. zapis, redkost rabe, prevzete besed itn. Razporeditev opomb po kategorijah in številčni podatki so prikazani v Tabeli 3, zaradi večje preglednosti in lažje primerjave je ohranjeno zaporedje kategorij iz Tabele 1.

Kategorija	Popolno ujemanje	Večinsko ujemanje
samo v določenem pomenu/kontekstu	3	37
zaznamovano	5	55
neznana beseda ali pomen besede	11	24
nad- oz. podpomenka	1	20
razlaga	0	2
drugo	2	20
skupaj opomb	22	158
skupaj parov	15	109

Tabela 3: Številčna razporeditev kategorij opomb ob popolnem in večinskem ujemanju odgovorov.

Med pari, ki so jih sodelujoči označevali kot sprejemljive (DA), so najpogosteje opozarjali, da sta besedi sopomenski samo v enem pomenu ali kontekstu, npr. *dilema – težava*, *identiteta – osebnost*, *koncentracija – osredotočenost*, *privilegij – ugodnost*, *stigma – zaznamovanost*. Pogosto se pojavljajo tudi opombe glede zaznamovanosti besedišča, npr. *beluš – asparagus* (citatno), *cedilo – cedilka* (pogovorno), *morilec – krvnik* (zastarelo), *pes – kuža* (pogovorno), *strpnost – potrpežljivost* (pogovorno), da sta besedi nad- in podpomenka (npr. *avto – osebno vozilo*, *avtomobil – osebno vozilo*, *brat – sorojenec*, *poroka – ženitev*, *kašelj – pokašljevanje*), da ne poznajo besed ali pomenov besed (*modrček – nedrc*, *oklevanje – obiranje*, *rit – zadnja plat*), da gre za razlago (*jok – pretakanje solz*) ter drugo (npr. *dež – dežne kaplje*: mero- oz. holonimija, *elita – veljaki* in *elita – pomembneži*: neujemanje slovničnega števila, *prerok – profet*: nenavadna oblika, *sestra – sorojenka*: redka raba). Zanimivo, da so par *brat – sorojenec* ocenjevalci občutili kot razmerje nad- oz. podpomenskosti, pri paru *sestra – sorojenka* pa je en ocenjevalec opozoril na redkost rabe, drugih komentarjev ni bilo.

Med pari, ki so jih sodelujoči označevali kot pogojno sprejemljive (POGOJNO DA), se najpogosteje pojavljajo primeri zaznamovanega besedišča, npr. *avto – kripa* (slabšalno), *deček – mulec* (slabšalno, negativni odnos), *juha – župca* (pogovorno, manjšalnica), *krema – maža* (pogovorno), *zadrga – fršlus* (pogovorno, narečno). Pogosti so tudi primeri, kjer sta besedi sopomenski samo v določenem pomenu oz. kontekstu, npr. *izkušnja – dogodivščina*, *kaos – štala*, *jesen – starost*, *posluh – čut*, *preteklost – prtljaga*, ter kjer gre za nad- in podpomenke, npr. *alkohol – etanol*, *aorta – arterija*, *avto – prevozno sredstvo*, *fotoaparat – digič*, *priseljenec – tujec*. Zaznani so tudi pari, kjer so ocenjevalci navedli, da ne poznajo besed ali pomenov besed, npr. *koder – krauželj*, *pivo – pirček*, *rit – prdulja*, *telovnik – lajbič*, ter kjer so jim pripisali druge opombe, npr. *pogum – jajca*: manjka del zveze, *policija – murja* in *rit – guza*: tujka.

Med pari, ki so jih sodelujoči označevali kot nesprejemljive (NE), so najpogosteje navajali, da gre za besedi, ki sta sopomenski samo v določenem pomenu oz. kontekstu, npr. *ljubezen – življenjski tok*, *stopnica – terasa*, *živina – blago*. Pojavljale so se tudi opombe glede zaznamovanosti besedišča, npr. *čarovnica – ćudežnica* (pozitiven odnos), *nedelja – teden* (zastarelo), da ocenjevalci besede ali pomena besed ne poznajo (*čik – žvečilni gumi* in *čik – žvečilka*,[17] *laboratorij – pospeševalnik*) je predlagana sopomenka bolj razlagalne narave (*rekreacija – raztezne vaje in vaje za moč*), da gre za nad- in podpomenki (*projekcija – podatek*) ter druge opombe (*davek – dan*: nepravilno črkovanje, *nedelja – teden*: tujka).

Pri vseh parih, kjer so ocenjevalci izbrali odgovor NISEM PREPRIČAN/NE VEM, je bila zabeležena

[17] Med opombami na ta dva para so ocenjevalci opozarjali tudi, da sicer besedi sta sopomenski v določenem pomenu oz. kontekstu, vendar so ju označili kot nesopomenska, kar so utemeljili, da za študentsko generacijo *čik* pomeni izključno cigareto.

opomba, da gre za besede ali pomene, ki jih ocenjevalci ne poznajo, npr. *čarovnica – bela žena*, *civilist – legist*, *obrok – rata*, *stranka – kunt*, *zaliv – olmun*. Dodatno so se pojavljale tudi opombe, da gre za zaznamovane besede, npr. *avto – gare* (slabšalno), *kašelj – brehanje* (pogovorno), *koder – loken*, *krona – dika* (arhaično), *zdravilo – arcnije* (zastarelo, arhaično), da sta besedi sopomenski samo v enem pomenu ali kontekstu (npr. *postava – geštel*), da gre za nad- in podpomenke (*torbica – nabočnica*) ter drugo, npr. *srajca – košilja*: nepravilno črkovanje, *zdravilo – biofarmacevtik*: dvomi o dejanski rabe besede, *zdravilo – arcnije*: neujemanje slovničnega števila.

Glede na popolno in večinsko ujemanje je bilo v kategoriji DA ali POGOJNO DA umeščenih skupaj 118 izmed 166 parov ali 71 %, v kategoriji NE in NISEM PREPRIČAN/NE VEM je bilo umeščenih po 24 izmed 166 parov oz. po 14,5 %. V kategoriji DA ali POGOJNO DA torej v sprejemljivo oz. relevantno gradivo, so ocenjevalci načeloma umeščali pare, kjer so tudi opozarjali, da gre za zaznamovane besede, npr. *babica – starejša gospa* (pogovorno, zastarelo, pozitiven odnos), *debelost – zašpehanost* (slabšalno, pogovorno, negativen odnos), *kmet – seljak* (slabšalno, negativen odnos), *novinar – pisun* (slabšalno, negativen odnos), *steklenica – flaša* (pogovorno), da sta besedi sopomenski samo v enem pomenu ali kontekstu (npr. *blago – capa*, *izrazoslovje – izrazje*, *legenda – štorija*, *rit – zahrbtnež*, *žarnica – sijalka*) ter nad- oz. podpomenke (npr. *kovanec – novčič*, *nakup – fasunga*). V ti kategoriji so bili umeščeni tudi pari, kjer so ocenjevalci opozarjali na npr. pomenske nianse, redkost rabe ali slovnična neujemanja (npr. predlagana sopomenka je v množini), npr. *cedilo – sito*, *stereotip – predsodek*, *pes – štirinožni prijatelj*. V kategorijo NE, torej nesprejemljivo oz. nerelevantno gradivo, so najpogosteje spadale tujke, nepravilno zapisane besede ter besedi, ki bi lahko bili sopomenki samo v enem pomenu ali kontekstu, ampak na ta pogoj v primeru večinskega odgovora NE je opozarjal le en ocenjevalec, npr. *davek – dan*, *nedelja – teden*, *živina – blago*, *stopnica – terasa*, *projekcija – podatek*. V kategorijo NISEM PREPRIČAN/NE VEM, torej gradivo, ki zahteva dodaten in podroben pregled, so ocenjevalci načeloma umeščali pare, kjer besed ali pomenov niso poznali ali kjer so ocenjevalci menili, da se besede sploh oz. redko uporablja, npr. *avto – sinhronka*, *cigareta – španjoleta*, *fotografija – heliotipija*, *moka – mlevina*, *zdravilo – biofarmacevtik*, *torbica – nabočnica*. Pri večinskem ujemanju sta le dva para dobila opombo, da je predlagana sopomenka razlagalne narave, od tega je bil en par (*jok – pretakanje solz*) opredeljen kot sprejemljiv (večina odgovorov DA), drug (*rekreacija – raztezne vaje in vaje za moč*) pa kot nesprejemljiv (večina odgovorov NE).

5.2. Podatki o ocenjevalcih in nalogi

V prvem delu vprašalnika sem zbirala podatke o ocenjevalcih. Vsi ocenjevalci v pilotni skupini spadajo v starostno skupino 20–30 let, najmlajši je rojen leta 2001, najstarejši pa leta 1995. Ker je šlo za študentsko populacijo, so vsi ocenjevalci navedli, da študirajo, večina je navedla tudi, da je jezik osrednji predmet njihovega študija. Le en študent je opredelil, da jezik ni v ospredju njegovega študija, ker študira filozofijo. V naslednjem vprašanju sem spraševala, na katerih področjih ima jezik zanje osrednjo vlogo, možnih je bilo več odgovorov. Na voljo so imeli tri odgovore, in sicer da jih jezik zanima, ker se z njim pretežno ukvarjajo v procesu izobraževanja, da jezik uporabljajo v poklicne namene ali da se z jezikom ukvarjajo zgolj ljubiteljsko. Vsi so označili, da se z jezikom ukvarjajo v izobraževalnem procesu, polovica je dodatno označila, da jezik uporabljajo tudi v poklicne namene. Nato so opredeljevali največ tri področja oz. dejavnosti, kjer je jezik v ospredju njihovega zanimanja, ki so jih izbirali iz ponujenega seznama devetih možnosti. Najpogostejši odgovor je bil raziskovanje oz. študij jezika (5 odgovorov), lektoriranje (4), prevajanje (3), poučevanje slovenščine (2) ter predavanje jezikoslovnih predmetov na višji oz. univerzitetni ravni in tvorjenje besedil (po 1 odgovor). Nihče ni izbral področja leksikografije in ljubiteljskega raziskovanja jezika, drugih odgovorov prav tako ni bilo. V naslednjem vprašanju so morali ocenjevalci izbrati le eno izmed prej navedenih področij oz. dejavnosti, ki je za njih glavno oz. najbolj relevantno. Kot glavno so raziskovanje oz. študij jezika izbrali trije, po en je navedel lektoriranje, prevajanje in tvorjenje besedil.

Sledila so vprašanja o nalogi. Najprej so ocenjevalci opredeljevali, koliko ur jim je reševanje naloge vzelo. V povprečju so študenti za izpolnitev preglednice potrebovali približno 6 ur, najhitrejši jo je rešil v treh urah, najpočasnejši pa v enajstih. Vsi so zatrdili, da so bila navodila jasno opredeljena. Le en študent je navedel, da je imel med reševanjem naloge težave, in sicer da veliko besed ni poznal, zato se je težko opredeljeval do potencialne sopomenskosti. Ocenjevalci so imeli tudi možnost podati svoje pomisleke, opazke in komentarje, ki niso bili zajeti v navodilih. Te so podali trije ocenjevalci, ki so navedli, da bi si želeli, da bi bila kategorija POGOJNO DA bolje opredeljena, da niso bili prepričani, v katero skupino naj uvrstijo nad- in podpomenke, razlage besed in neuveljavljene tujke ter da so pogrešali možnost preverbe sopomenk v drugih virih, saj bi s tem lahko podali boljše odgovore, vendar hkrati razumejo, zakaj jih niso smeli uporabljati. Dodatnih komentarjev niso imeli.

6. Diskusija

Glede ustreznosti uporabniško dodanih sopomenk se je izkazalo, da je bilo nedvomno nesopomenskega gradiva, ki so ga prispevali uporabniki, zelo malo. Primeri, kjer so se odgovori ocenjevalcev popolnoma ujemali, so majhen del vzorca (34 parov oz. približno 3,5 %), kar je bilo sicer predvidljivo glede na obseg podatkov in število ocenjevalcev. Nekoliko več je bilo primerov, kjer je izstopal odgovor enega ocenjevalca (132 parov oz. približno 13,5 % nabora). Skupaj je torej parov z večinskim ujemanjem 166 oz. 17 % nabora. Znotraj tega je parov, kjer je izstopajoč odgovor iz nasprotnega pola (npr. vsi odgovori NE in en POGOJNO DA, vsi odgovori DA in en NE), približno ena tretjina (42 parov). Veliko večino teh parov so ocenjevalci ocenili kot sprejemljive. Parov, ki so bili večinsko umeščeni v kategorijo NE ali NISEM PREPRIČAN/NE VEM, je bilo le 24 izmed 166 oz. 14,5 %. To kaže, da so uporabniško dodani predlogi sopomenk načeloma relevantni in konstruktivni, saj je bilo skupaj v kategoriji DA in POGOJNO DA umeščenih 118 izmed 166 parov (71 %), kjer so se odgovori ocenjevalcev večinsko ujemali, četudi gre za primere, ki bistveno presegajo tradicionalno jezikoslovno dojemanje sopomenskosti. Ta ugotovitev je v skladu z raziskavo iz leta 2020, kjer se je po analizi uravnoteženega dela vzorca 1.662 sopomenk (največ 10 predlogov na uporabnika) izkazalo, da je okoli

70 % uporabniško dodanih predlogov konstruktivnih in hkrati nezaznamovanih, okoli 20 % konstruktivnih in zaznamovanih ter le dobrih 6 % odstotkov nekonstruktivnih oz. zlonamernih (prim. Arhar Holdt in Čibej, 2020, str. 6). Ocene le ene sodelujoče skupine razumljivo niso in ne smejo biti zadostna podlaga za generalizacijo, a vendar so podatki glede relevantnosti uporabniško dodanega gradiva spodbudni. Posebno pozornost bo potrebno nameniti kategoriji NISEM PREPRIČAN/NE VEM, saj so vsi v to kategorijo umeščeni potencialni sopomenski pari dobili opombo, da ocenjevalci ne poznajo besede ali pomena besede. To ne pomeni, da so tovrstni predlogi nerelevantni, bodo pa v procesu posodobitev *Sopomenk* zahtevali več pozornosti s strani urednikov, npr. natančnejše iskanje korpusnih zgledov, uporabo dodatnih virov za preverjanje dejanske rabe itn.

Na podlagi povratnih informacij iz vprašalnika in korespondence s študenti, ki so se name obračali med ocenjevanjem, je razvidno, da je pred nadaljnjim ocenjevanjem treba dopolniti navodila ter podrobneje pojasniti ozadje in cilje raziskave. Študenti so prejeli le zelo kratko pojasnilo, da se ocenjevanje izvaja v okviru doktorske naloge ter kakšen je glavni cilj, brez podrobnejših opisov in pojasnitev. V navodilih so sicer dobili tudi informacijo, naj pri ocenjevanju ne uporabljajo drugih jezikovnih virov, a brez obrazložitve, zakaj to ni zaželeno. Iz opomb, ki so jih dajali, je razvidno, da so nekateri to navodilo kršili, saj so bile pogoste opombe tipa: *v Gigafidi sem zasledil_a*, *Iz Gigafide je razvidno*, *Ne v Franu ne v Gigafidi nisem zasledil_*a, ali so v tabelo celo pripenjali povezave na druge priročnike. To je zelo verjetno posledica pomanjkanja utemeljitve, zakaj to ni zaželeno. Veliko vprašanj, zastavljenih po e-pošti, je spremljala izjava, da želijo nalogo "pravilno rešiti". Možna razlaga za to je, da so študenti vajeni na ocenjevanje odgovorov po principu prav–narobe, v opisu naloge pa ni bilo izrecno navedeno, da ni pravilnih ali napačnih odgovorov oz. da so vsi odgovori pravilni, saj sprašujemo po njihovem mnenju. Ta navedba bo mogoče manj relevantna za ostale predvidene skupine ocenjevalcev, a vendarle se kaže kot smiselna dopolnitev opisa naloge in njenega namena.

V navodilih so študenti dobili informacijo, da gre za ocenjevanje uporabniško dodanih sopomenk, vendar brez razlage, kakšni vse predlogi se lahko na seznamu pojavijo. Pogosta so bila vprašanja, kaj narediti z nad- oz. podpomenkami, neuveljavljenimi tujkami ali popačenkami ter predlogi, ki so bolj razlagalne narave. Na podlagi povratnih informacij se kot smiselna dopolnitev kaže tudi dodatni opis, katere vrste podatkov lahko ocenjevalci pričakujejo na seznamu. Na drugi strani so bili zatipki oz. očitne napake izpostavljeni kot primeri, ki se smatrajo za nerelevantne oz. nesopomenske, vendar jih niso vedno vsi ocenjevalci umestili v kategorijo NE (npr. par *Zemlja – e*, kjer gre za očitno napako, je eden izmed ocenjevalcev umestil v kategorijo NISEM PREPRIČAN/NE VEM). Kot so ocenjevalci izpostavili sami, so pogrešali natančnejšo opredelitev kategorije POGOJNO DA. V navodilih za ostale skupine uporabnikov je zato treba natančneje opredeliti, da v ta sklop spadajo pari, kjer ocenjevalci sicer lahko povedo nekaj glede sopomenskosti, a ta ni nedvomna in bi želeli ob sopomenskem paru dodatne informacije.

7. Zaključek in naslednji koraki

V prispevku sem opisala načrt raziskave, s katero želim nasloviti raziskovalno vprašanje doktorske disertacije, in sicer da se pogled strokovne oz. širše jezikovne skupnosti razlikuje od pogleda slovaropiscev, vendar je ta potencialni drugačni pogled jezikovne skupnosti lahko uporaben in pomemben za razvoj virov. Predstavila sem tudi potek evalvacijske naloge, ki so jo kot testna množica opravili študenti. S tem sem želela preizkusiti sestavljena navodila in izbrana orodja ter določiti časovno-finančni obseg in zahtevnost naloge, kar mi bo pomagalo pri načrtovanju dela in rekrutaciji preostalih predvidenih skupin ocenjevalcev.

Na podlagi odgovorov in povratnih informacij v pilotni skupini se je ocenjevanje izkazalo kot izvedljivo. Izbrani orodji, in sicer Googlova preglednica za evalvacijo sopomenskih parov ter spletno orodje za anketiranje 1ka, sta se izkazala kot ustrezna, enostavna za uporabo ter finančno in časovno vzdržna.[18] Potrebno bo izboljšati navodila za ocenjevalce, hkrati se zdi smiselno ocenjevalcem podrobneje pojasniti kontekst raziskave in namen ocenjevanja (pridobitev njihovega subjektivnega mnenja, ne "pravilnih" odgovorov). Naslovljena so bila problematična mesta v navodilih, navodila za ocenjevalce pa ustrezno preoblikovana in dopolnjena za naslednje predvidene skupine.

Predvideno je, da isti seznam v ocenjevanje dobi še 5 skupin ocenjevalcev, in sicer slovaropisci, poklicni prevajalci, lektorji, učitelji slovenščine in ljubitelji jezika brez jezikoslovne izobrazbe. V času pisanja prispevka poteka rekrutacija sodelujočih, podatki pa naj bi bili pridobljeni do poletja 2022. Sledijo analize rezultatov znotraj skupin, nato pa še primerjalno med skupinami.

Čeprav prvi rezultati še niso primerni za generalizacijo, ponujajo dober uvid v dileme pri presojanju sopomenskosti uporabniško dodanega gradiva. Spodbudno je, da je bilo (vsaj po prvem koraku raziskave) veliko uporabniško dodanega gradiva ocenjenega kot relevantnega. V primeru, da ocene drugih predvidenih sodelujočih skupin prinesejo podobne rezultate, bo to ugotovitev mogoče upoštevati pri nadaljnjem razvoju sopomenskih virov za slovenščino, v smer širitve in bogatenja z novimi podatki. Hkrati se potrjujejo predhodna spoznanja, da so uporabniški predlogi v kar največji meri konstruktivni in dobronamerni, kar je ključno za delovanje in nadaljnji razvoj odzivnih slovarjev.

8. Zahvala

Prispevek je nastal v okviru raziskovalnega programa Jezikovni viri in tehnologije za slovenski jezik (številka programa P6-0411), ki ga sofinancira Javna agencija za raziskovalno dejavnost Republike Slovenije.

9. Literatura

Andrea Abel in Christian M. Meyer. 2013. The dynamics outside the paper: user contributions to online dictionaries. V: *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*, str. 179–94. Trojina, Institute for Applied Slovene Studies in Eesti Keele Instituut.

Jurij Apresjan. 2000. *Systematic Lexicography* (prev. Kevin Windle). Oxford University Press, Oxford.

[18] Obe orodji sta namreč brezplačni za uporabo tako za ocenjevalce kot za raziskovalce ter ne zahtevata predznanja oz. dodatnega usposabljanja, posebne strojne opreme ali dodatne registracije.

Špela Arhar Holdt. 2015. Uporabniške raziskave za potrebe slovenskega slovaropisja: prvi korak. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 136–49. Znanstvena založba Filozofske fakultete.

Špela Arhar Holdt. 2020. How Users Responded to a Responsive Dictionary: the Case of the Thesaurus of Modern Slovene. *Rasprave Instituta za hrvatski jezik i jezikoslovlje,* 46(2): 465–82. doi:10.31724/rihjj.46.2.1

Špela Arhar Holdt in Jaka Čibej. 2020. Rezultati projekta "Slovar sopomenk sodobne slovenščine: Od skupnosti za skupnost". V: *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 24. – 25. september 2020, Ljubljana, Slovenija*, str. 3–9. Inštitut za novejšo zgodovino.

Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Bojan Klemenc, Iztok Kosem, Simon Krek, Cyprian Laskowski in Marko Robnik-Šikonja. 2018. Thesaurus of modern Slovene: by the community for the community. V: *Proceedings of the XVIII EURALEX International Congress, Lexicography in Global Contexts, 17-21 July 2018, Ljubljana*, str. 401–10. Znanstvena založba Filozofske fakultete.

Pavel Braslavski, Dmitry Ustalov in Mikhail Mukhin. 2014. A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. V: *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, str. 101–104. Association for Computational Linguistics. doi: 10.3115/v1/E14-2026

Jaka Čibej in Špela Arhar Holdt. 2019. Repel the syntruders! A crowdsourcing cleanup of the thesaurus of modern Slovene. V: *Electronic lexicography in the 21st century: Smart lexicography. Proceedings of the eLex 2019 conference, 1–3 October 2019, Sintra, Portugal*, str. 338–56. Lexical Computing CZ s.r.o.

Jaka Čibej, Darja Fišer in Iztok Kosem. 2015. The role of crowdsourcing in lexicography. V: *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of eLex 2015 Conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*, str. 70–83. Trojina, Institute for Applied Slovene Studies in Lexical Computing Ltd.

Darja Fišer, Aleš Tavčar in Tomaž Erjavec. 2014. sloWCrowd: A crowdsourcing tool for lexicographic tasks. V: *Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC'14*, str. 3471–75. European Language Resources Association (ELRA).

Darja Fišer. 2015. *Semantic lexicon of Slovene sloWNet 3.1*. Repozitorij raziskovalne strukture CLARIN.SI, http://hdl.handle.net/11356/1026

Polona Gantar, Simon Krek, Iztok Kosem, Mojca Šorli, Polonca Kocjančič, Katja Grabnar, Olga Yerošina, Petra Zaranšek in Nina Drstvenšek. 2013. *Leksikalna baza za slovenščino 1.0*. Repozitorij raziskovalne strukture CLARIN.SI. Pridobljeno iz http://hdl.handle.net/11356/1030

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek, ur. 2017. *Slovar sodobne slovenščine: problemi in rešitve.* Znanstvena založba Filozofske fakultete Univerze v Ljubljani, Ljubljana. doi:10.4312/9789612379759

Hans Henrich Hock. 1991. *Principles of Historical Linguistics* (druga, razširjena in dopolnjena izdaja izd.). Mouton de Gruyter, Berlin, New York.

Aleš Horák in Adam Rambousek. 2018. Wordnet Consistency Checking via Crowdsourcing. V: *Proceedings of the XVIII EURALEX International Congress, Lexicography in Global Contexts, 17–21 July 2018, Ljubljana*, str. 1023–29). Znanstvena založba Filozofske fakultete.

Iztok Kosem in Eva Pori. 2021. Slovenske ontologije semantičnih tipov: samostalniki. V: I. Kosem, ur., *Kolokacije v slovenščini*, str. 159–202. Znanstvena založba Filozofske fakultete Univerze v Ljubljani, Ljubljana. doi:10.4312/9789610605379

Iztok Kosem, Polona Gantar in Simon Krek. 2013. Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing for both lexicographers and crowd-sourcing. V: *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013*, str. 32–48. Trojina, Institute for Applied Slovene Studies in Eesti Keele Instituut.

Simon Krek, Cyprian Laskowski, Marko Robnik-Šikonja, Iztok Kosem, Špela Arhar Holdt, Polona Gantar, Jaka Čibej, Vojko Gorjanc, Bojan Klemenc in Kaja Dobrovoljc. 2018. *Thesaurus of Modern Slovene 1.0*. Repozitorij raziskovalne strukture CLARIN.SI., http://hdl.handle.net/11356/1166

Lionel Nicolas, Lavinia Aparaschivei, Verena Lyding, Christos Rodosthenous, Federico Sangati, Alexander König in Corina Forascu. 2021. An Experiment on Implicitly Crowdsourcing Expert Knowledge about Romanian Synonyms from Language Learners. V: *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, str. 1–14. LiU Electronic Press.

Tobias Schnabel, Igor Labutov, David Mimno in Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. V: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, str. 298–307. Association for Computational Linguistics. doi: 10.18653/v1/D15-1

Jerica Snoj. 2019. *Leksikalna sinonimija v Sinonimnem slovarju slovenskega jezika.* Založba ZRC, ZRC SAZU, Ljubljana.

Jerica Snoj, Martin Ahlin, Branka Lazar in Zvonka Praznik. 2016. *Sinonimni slovar slovenskega jezika.* Založba ZRC, ZRC SAZU, Ljubljana.

Rion Snow, Brendan O'Connor, Daniel Jurafsky in Andrew Y. Ng. 2008. Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. V: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 25-27 October 2008, Honolulu, Hawaii, USA*, str. 254–63. Omnipress Inc.

Tavčar, Aleš, Darja Fišer in Tomaž Erjavec. 2012. sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic. V: *Zbornik Osme konference Jezikovne tehnologije*, str. 197–202. Inštitut Jožef Stefan.

Jože Toporišič. 1992. *Enciklopedija slovenskega jezika.* Cankarjeva založba, Ljubljana.

Ada Vidovič-Muha. 2013. *Slovensko leksikalno pomenoslovje.* Znanstvena založba Filozofske fakultete, Ljubljana.

Ladislav Zgusta. 1971. *Manual of Lexicography.* Academia, Publishing House of the Czechoslovak Academy of Sciences, Praga.

Marina Zorman. 2000. *O sinonimiji.* Znanstveni inštitut Filozofske fakultete, Ljubljana.

Angleško-slovenska šahovska terminološka baza

Vili Grdič, Alja Križanec, Kaja Perme, Lea Turšič

Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
grdic.vili@gmail.com
alja.manja@gmail.com
kaja.perme@gmail.com
lea.tursic@gmail.com

Povzetek

V okviru univerzitetnega predmeta Terminologija smo izdelali angleško-slovensko šahovsko terminološko bazo, saj smo želeli ustvariti zanesljiv dvojezični vir šahovske terminologije, ki vsebuje tudi slovenščino. Baza je nastala po korpusnem pristopu. Izdelali smo angleški in slovenski korpus ter iz njiju izluščili 82 angleških in 109 slovenskih terminov. Razdelili smo jih v pet podpodročij (taktika, strategija, otvoritev, končnica in ostalo) ter jih opremili z definicijami, kolokacijami, zgledi rabe, podatki o statusu in opombami.

English-Slovenian Chess Terminology Database

In our university Terminology course, we built an English-Slovenian chess terminology database because we wanted to create a reliable bilingual source of chess terminology that includes the Slovenian language. The database is based on the corpus approach. We built an English and a Slovenian corpus and extracted 82 English and 109 Slovenian terms. We divided them into five subfields (tactics, strategy, opening, endgame and other) and added definitions, collocations, usage examples, status information and notes.

1. Uvod

Šah lahko razumemo le kot prostočasno dejavnost v obliki namizne igre, v resnici pa gre za športno disciplino ter pestro interdisciplinarno in terminološko zelo kompleksno področje. Je tudi predmet raziskav številnih področij, tako naravoslovnih kot družboslovnih.

Slovenska šahovska terminologija lahko deluje zelo zapleteno, včasih je tudi težko najti slovenske ustreznike tujejezičnih terminov, ki šahiste zaradi močnega vpliva spleta kar naprej obkrožajo (sploh angleški). Je pa njeno poznavanje ključno za pravilno izražanje in opisovanje šahovskih partij. Da bi prevajalcem in jezikoslovcem pri iskanju slovenskih ustreznikov olajšali delo, smo se na podlagi korpusnega pristopa lotili izdelave dvojezične šahovske terminološke baze. S šahom se eden od avtorjev prispevka tudi sam dejavno ukvarja, kar je prispevalo k motivaciji za raziskavo in vsebinskim izhodiščem.

Terminologija je veda, ki preučuje specializirano izrazje določenega strokovnega področja, imenovano termini. Poleg tega so njen predmet obravnave tudi pojmi in njihova razmerja ter poimenovanja v različnih jezikih, eden od glavnih ciljev pa je izdelava terminoloških priročnikov. Tako jo lahko označimo tudi kot normativno vedo, saj z izdajanjem priročnikov predpisuje rabo izrazja in pripomore pri postopku terminološke standardizacije (povzeto po Vintar, 2017: 17–18).

Pri zbiranju terminoloških podatkov smo izbrali korpusni pristop, ki v zadnjih letih prevladuje na področju terminografije. Večina gradiv, iz katerih črpamo jezikovne podatke, je danes prostodostopnih. Najhitrejši in najlažji način pri opisovanju jezika je zato s pomočjo programske opreme, ki omogoča analizo besedila (npr. *Sketch Engine*), in izdelavo lastnega korpusa, iz katerega enostavno pridobimo sezname besed, ki jih nato poljubno urejamo, samodejno luščimo termine in dalje analiziramo (povzeto po Vintar, 2017: 83).

Pri projektu sta nam zelo pomagala šahovska mojstra Iztok Jelen in Matjaž Mikac, svetovala nam je tudi mojstrica Monika Rozman.

2. Namen članka

Namen članka je opisati projekt gradnje dvojezične terminološke baze, ki smo jo ustvarili v okviru predmeta Terminologija. Kot prevajalci se zavedamo, da so dandanes jezikovni viri izjemno pomembni, zato smo tudi sami skušali ustvariti koristen vir, osnovan po sodobnem jezikoslovnem pristopu. Projekt razumemo tudi kot nadaljevanje že opravljenega dela na področju slovenske šahovske terminologije, hkrati pa slovenskemu prostoru približamo tudi angleške termine. Želimo spodbuditi k nadaljnji gradnji slovenskih in večjezičnih šahovskih priročnikov.

3. Oris področja in sorodne raziskave

Šah je okoli 1500 let stara strateška namizna igra, ki jo dandanes uvrščamo tudi med šport in je pestro interdisciplinarno področje. Obravnavajo ga številna druga področja in opravljajo aktualne raziskave, kot na primer psihologija (povezava osebnosti, vsakdanjega življenja in šaha, gl. Krivec, 2021), matematika (šah in matematika, gl. Grosar, 2017), nevrologija (šah in avtizem, gl. Gomes de Sousa, 2021), robotika (šahovski robot, gl. Goldman et al., 2021), računalništvo (šah in umetna inteligenca, gl. Guid, 2010), sociologija (ženske in spolna enakost v šahovskem svetu, gl. Vishkin, 2022), pedagogika (vpliv šaha na učenje tujih jezikov, gl. Harazińska in Harazińska, 2017) in podobno.

Po mnenju šahovskega mojstra Matjaža Mikaca (VIR 1) ima šah več dimenzij: je igra, spada pod znanost, umetnost in predvsem šport. Sicer se nekateri z zadnjim ne strinjajo, a Mikac pravi, da pogosto kritizirana premajhna fizična aktivnost v šahu ni edini kriterij, po katerem se neka disciplina uvršča med šport. Tako kot druge športne discipline ima tudi šah bogato tekmovalno tradicijo (svetovna prvenstva, šahovska olimpijada, šolska tekmovanja) ter moramo biti za partije dobro telesno in umsko pripravljeni (npr. za več ur dolge partije).

Šah ima v Sloveniji bogato tradicijo, od konca 19. stoletja do danes smo kot narod beležili izjemne dosežke,

ki sodijo v svetovni vrh, npr. uspeh Josipa Plahute, Milana Vidmarja, Luke Leniča in Laure Unuk (povzeto po Jelen, 2006: 10–12). Šahovski velemojster Marko Tratar (2003: 4) navaja, da je »[š]ah /…/ v slovenskem časopisju 20. stoletja vseskozi imel svoj prostor, tako po svoji tekmovalni plati /…/ kot tudi zaradi svojih umetniških in znanstvenih ter pedagoških razsežnosti.«

V zadnjih letih (največ v času pandemije koronavirusa med letoma 2020 in 2021) pa so številni dejavniki pripomogli k večji priljubljenosti in razširjenosti šaha po vsem svetu. Velik globalni vpliv je imela serija *Damin gambit* (Frank, 2020; gl. Jurc, 2020), ki na primeru fiktivne zgodbe realno ponazarja zaničljiv odnos do žensk v šahovskem svetu 20. stoletja, tudi vse partije in dvoboji so iz šahovskega vidika prikazani pravilno (Loeb McClain, 2020). Na najstnike in mlade odrasle, pa tudi ostale, je močno vplivala platforma Twitch za oddajanje raznih vsebin v živo (Johannson, 2021). Na njej nekateri velemojstri in drugi šahisti v živo izobražujejo in razvedrijo tudi do več deset tisoč gledalcev. Med največjimi so Hikaru Nakamura (profil *GMHikaru*), Alexandra in Andrea Botez (*BotezLive*) ter za slovenski prostor Laura Unuk, Teja Vidic in Lara Janželj (*Checkitas*). Tudi štiri spletna amaterska šahovska tekmovanja *PogChamps* so zelo pripomogla k veliki gledanosti šaha, saj so na njih sodelovali popularni oddajalci vsebin (»streamerji«) s Twitcha in Youtuba (Johannson, 2021; gl. VIR 2). Po mnenju Matjaža Mikaca (VIR 1) učinek omenjenih dejavnikov na šah v Sloveniji ni bil tako močan in očiten, saj smo kot narod šahovsko že dobro razviti.

3.1. Šahovska terminologija

Z etimologijo nekaterih slovenskih in tujejezičnih šahovskih terminov se je ukvarjal pravnik Leonid Pitamic (1950). Navaja, da večina terminov izvira iz latinščine, arabščine in perzijščine, ti pa so se pod vplivom kulturno-političnega dogajanja v Evropi od 12. stoletja dalje v evropskih jezikih razvijali različno. Nekateri termini iz več jezikov imajo dokaj podoben izvor (npr. *šah* iz srednjeveškega latinskega *scacci*), nekateri pa precej različnega in s tem tudi drugačen dobesedni pomen (npr. termini za *lovca*: ang. *bishop* 'škof', nem. *Läufer* 'tekač', fr. *fou* 'norec', rus. *слон* 'slon'). Pitamic navaja, da so besede *šah, šahovnica* in *ček* izrazoslovno vplivale na nekatere besede v več evropskih jezikih s področja prava, gospodarstva in finančništva (npr. današnja fr. beseda za šahovnico *échiquier*, ki je povezana z najvišjim sodiščem *Echiquier* v stari Normandiji; povzeto po Pitamic, 1950: 173–204).

Slovenska šahovska terminologija je nastajala pod vplivom srbščine, iz nje so prevzemali in prevajali stari slovenski šahovski mojstri, kot je Milan Vidmar (gl. Vidmar, 1946; 1951). Njihova spoznanja (in del teorije hrvaškega šahista Vladimirja Vukovića, gl. 1978; 1990) pa je v več prispevkih za učni načrt šahovskega izbirnega predmeta za osnovne šole zbral šahovski mojster Iztok Jelen (VIR 3; VIR 10; gl. Jelen, 2004a; 2004b).

Šahovska terminologija je v manjši meri večjezična, saj so se v večini jezikov uveljavili nekateri tujejezični termini iz francoščine (*en passant, j'adoube*), nemščine (*Zwischenzug, Fingerfehler, Blitz*) in italijanščine (*fianchetto, intermezzo*). V žargonu slovenskih šahistov pa lahko zasledimo tudi hrvaške (*pješak/pijun*), srbske (*dirigovanje*) in ruske termine (*пешка* 'peška'), kar je verjetno ostalo še iz časov Jugoslavije in Sovjetske zveze, ko je imel šah velik (pogosto tudi politični) pomen in so o njem večkrat poročali v medijih (VIR 1).

S šahovsko terminologijo so se ukvarjale že številne raziskave, ki so pokazale kompleksnost področja in njegovega izrazoslovja. Adylova (2017: 8) navaja, da je že samo na šahovskem terminološkem podpodročju o šahovskih otvoritvah svoja strukturna klasifikacija terminov (dvo-, tri-, štiri- in večkomponentna poimenovanja otvoritev), podobno velja tudi za ostala šahovska podpodročja (središčnica, končnica, taktike ipd.). Karayev (2016: 103) opisuje, da so nekateri splošni izrazi prešli v šahovsko terminologijo (npr. *to calculate* 'računati'), nekateri pa z determinologizacijo tudi iz področja šaha v splošni jezik, ki se običajno uporabljajo v prenesenem pomenu (v slovenščini npr. *imeti nekoga v šahu/matu/patu*). V nadaljevanju (2016: 103) navaja, da ljudje šah velikokrat asociirajo z vojno in politiko, zato se šahovska terminologija v prenesenem pomenu pogosto omenja tudi v nešahovskih kontekstih. Avtor se opre na novinarstvo in z njim povezan publicistični jezik: »Naša vlada kakor kmet ne gre nazaj« (*Moskovskij Komsomolets*, 21. 1. 2005). Dodaja, da fenomen prehajanja šahovske terminologije v splošni jezik ni nič nenavadnega, saj je ravno to značilno že za športno terminologijo (v splošnem jeziku uporabljamo npr. *napad, podajanje žoge, zadetek v črno* ipd.). Tudi Zhuravleva in Vlavatskaya (2021: 534) navajata, da šahovska terminologija ni omejena izključno na področje šaha (za šah specifični izrazi so npr. *šah, šah mat, pat, fianketo*), temveč se razteza na celotno športno sfero (npr. *zmaga, poraz, napad, obramba, sodnik*).

3.2. Jezikovni viri šahovske terminologije

Skoraj vsaka (večja) angleška spletna stran za igranje šaha in tudi druge spletne šahovske strani imajo vodnike, glosarje in ostale vire za učenje šaha, tam pa najdemo tudi sezname terminologije z definicijami, slikami ipd. (npr. na chess.com, lichess.org, chess24.com). Šahovski mojster Iztok Jelen (VIR 3) se strinja, da je virov za angleščino veliko, spletni so dobro dostopni, a se med seboj lahko zelo razlikujejo. Pravi, da je težko določiti njihovo verodostojnost, saj so definicije lahko različne, nekatere zelo splošne, druge natančnejše; avtor ni znan, ne navajajo virov informacij in ne opredelijo, kako je glosar nastal (npr. korpusni pristop). Ravno tako je vprašljiv nabor terminov, saj nekateri glosarji opisujejo kolokacije in druge besedne zveze kot termine (*control of the center* 'nadzor/varovanje središča'), drugi dodajajo tudi žargonizme (*cheapo* 'lahka past') in celo novotvorjenke (*Botez Gambit* 'nenamerna žrtev dame', ki ga je izumila šahistka Alexandra Botez; VIR 14). Nekaterih terminov, ki jih zasledimo v spletnih glosarjih, pa v našem korpusu sploh ni in jih na spletu najdemo zgolj v drugih glosarjih (torej obstajajo le v teoriji) in ne v dejanski rabi, npr. *knight fork windmill* (podvrsta taktike *windmill*). Za splošnega uporabnika so spletni viri uporabni in dovolj natančni, za jezikoslovne namene pa so zanesljivejši glosarji iz šahovskih knjig.

Sami smo se največ opirali na angleška glosarja iz knjig *Chess For Dummies* (Eade, 2016) in *Winning Chess Openings* (Seirawan, 2016) priznanega ameriškega šahista Yasserja Seirawana. Iztok Jelen priporoča *The Oxford Companion to Chess* (Hooper in Whyld, 1992).

Za slovenščino smo od spletnih virov zasledili glosarja *Šahovsko izrazoslovje* na portalu ICP (VIR 4) in *Šahovsko*

izrazoslovje na Wikipediji (VIR 5), ki imata velik nabor terminologije in sta za splošnega uporabnika dovolj natančna. Zanesljivejši so prispevki iz osnovnošolskega učnega načrta za izbirni predmet šaha Iztoka Jelena (VIR 10; 2004a; 2004b), ki vsebujejo pravila igre, obširno teorijo in slovenske termine. Avtor sam pa priporoča tudi Slovar slovenskega knjižnega jezika ter kot pomoč za nadaljnjo raziskovanje še rusko enciklopedijo *Šahmaty, Enciklopedičeski slovar* (1990) in hrvaški prevod enciklopedije *Golombek's Encyclopedia of Chess* (Golombek, 1980).

4. Metoda

Glavni cilj našega projekta je bila izdelava dvojezičnega šahovskega glosarja oz. terminološke baze, ki bi nastala na podlagi angleškega in slovenskega korpusa besedil. Pri izdelavi smo se odločili za korpusni pristop. Želeli smo raziskati dejansko rabo šahovskih terminov v obeh jezikih, v bazo vključiti najpogostejše termine v rabi in gesla opremiti z definicijami, kolokacijami, zgledi rabe, podatki o statusu in morebitnimi opombami. Na podlagi angleškega korpusa smo zgradili angleško terminološko bazo, nato pa smo z uporabo slovenskega korpusa dodali slovenske terminološke ustreznike in jih opremili z relevantnimi informacijami.

4.1. Korpusni pristop

Terminološka baza je zasnovana po korpusnem pristopu, kar pomeni, da smo jezikovne podatke zanjo pridobili iz korpusa, tega pa smo zgradili in analizirali v orodju *Sketch Engine*. Za ta pristop smo se odločili, ker je lažje ustvariti korpus besedil in ga s pomočjo računalniških konkordanc analizirati ter tako opisovati jezik določenega strokovnega področja, kot pa to početi na stari način z listkovnim gradivom (Logar in Vintar, 2008: 5). Korpus, v katerega so zajeta različna besedila z nekega področja, lahko v dovolj velikem obsegu služi kot reprezentativni vzorec jezika ter daje vpogled v dejansko rabo jezika. Takšen pristop ni le lažji, temveč tudi sodobnejši in hitrejši ter tako uporabniku prijaznejši (Logar in Vintar, 2008: 14). Že samo z osnovno analizo korpusa v orodju *Sketch Engine* dobimo seznam besed, ki ga nato lahko poljubno urejamo za nadaljnjo analizo (abecedno, po dolžini itd.), in podatek o pogostosti pojavitve besed, ki je pri prepoznavanju tipičnih terminoloških vzorcev posebej zaželen (Vintar, 2017: 84). Če pa je korpus lematiziran in oblikoskladenjsko označen, ga lahko analiziramo še podrobneje, npr. izberemo možnost, da se prikažejo vsi prislovi, pridevniki, predlogi itd., ki se pojavljajo ob nekem geslu, in tako ugotavljamo, katere kolokacije so najpogostejše (Logar in Vintar, 2008: 5). Programi za analizo korpusov so opremljeni s funkcijami, ki samodejno luščijo ključne besede in termine, eno- in večbesedne. Tako dobimo nabor terminov, uporabnik jih nato le še ročno pregleda in neprimerne odstrani.

Korpusni pristop danes ni nujno potreben le pri gradnji terminoloških priročnikov, temveč tudi pri gradnji kakršnih koli jezikovnih priročnikov, ki želijo predstavitvi aktualno stanje jezika (Gantar, 2004: 170). Poleg avtomatizacije leksikografskih postopkov so njegove prednosti še informacija o sobesedilu in rabi ter možnost izločanja irelevantnih informacij (Gantar, 2004: 177).

4.2. Angleški in slovenski korpus

Za namene projekta smo ustvarili dva korpusa, angleškega in slovenskega. Cilj pri zbiranju besedil je bil, da dosežemo čim boljšo zastopanost terminologije, zato smo izraze razdelili po terminoloških podpodročjih (taktika, strategija, otvoritev, končnica in ostalo) in za vsako vključili približno enako število besed. Pri zbiranju virov smo pazili, da smo zajeli tako splošne kot tudi specializirane šahovske vire ter po vsebini pokrili vseh pet podpodročij. Z različnostjo in enakomerno zastopanostjo besedil smo iz korpusa želeli izluščiti relevantne termine, ki bi bolje odslikavali dejansko rabo, in dobiti natančnejše podatke o pogostosti rabe. V korpus nismo zajeli takšnih virov, ki vsebujejo veliko (ali izključno) definicij, kot so na primer glosarji, in takšnih, ki poleg šaha zajemajo še ostala, za nas irelevantna področja (in s tem tudi termine). Kljub temu gre za omejen nabor virov, saj smo v slovenski korpusa vključili le prostodostopne spletne vire, v angleškega pa poleg takih tudi nekatere knjige v formatu PDF.

Slovenski korpus obsega 139.964 besed in je sestavljen iz 55 besedil. Vsi viri so prostodostopni na spletu. Za namene čim večjega nabora terminologije je največ spletnih prispevkov o šahovski teoriji, ostalo pa so splošni šahovski članki. Slovenskih knjig o šahu nismo mogli vključiti, saj te na spletu niso prostodostopne, zato je korpus v primerjavi z angleškim bistveno manjši. Korpus obsega članke različnih tem, od tega jih je 28 o pravilih igranja (npr. VIR 6), strategiji posameznih delov igre (npr. VIR 7), figurah, zgodovini šaha in splošno (npr. VIR 8; VIR 9). Vključili smo tudi 10 prispevkov s portala ICP (npr. VIR 4) in 17 iz spletne učilnice za šah kot izbirni predmet v osnovnih šolah (VIR 10).

Angleški korpus obsega 869.592 besed in je sestavljen iz 21 besedil. Tako kot slovenski tudi ta pokriva ogromno teorije o posameznih delih igre (otvoritev, središčnica in končnica, npr. VIR 11), strategiji in taktiki ter pravilnik svetovne šahovske zveze FIDE (VIR 12) – omenjena vsebina pa je tako v spletnih virih kot tudi knjižnih. V korpusu je 7 daljših spletnih člankov (npr. VIR 13), dodali pa smo tudi 14 šahovskih knjig oziroma priročnikov v formatu PDF (npr. Eade, 2016). Ker so knjige mnogo daljše od člankov, ima angleški korpus v primerjavi s slovenskim manj besedilnih vnosov, a obsega veliko več besed.

5. Terminološka baza

Pri luščenju, določevanju in razvrščanju terminov smo naleteli na nekaj težav. Pri tem nam je pomagala šahovska mojstrica Monika Rozman.

5.1. Težave z luščenjem terminov

Program je samodejno izluščil 1000 eno- in večbesednih terminov. Med njimi je bilo veliko izrazov, ki niso bili termini, zato smo morali sezname prečistiti.

S terminoloških seznamov smo odstranili naslednje (primeri so iz angleških):

napačno zaznane besede	manjkajoči deli besed *agonal*, *advan, endg* napačno branje *parry* (namesto *Garry (Kasparov)*)
glava in noga knjige	*dummies* (iz *Chess for dummies*), *Dvoretsky* (avtor), *2010* (letnica)
poteze in koordinate	poteze kmetov (tudi oznake polj) *f4, e4, g5* poteze figur *Ke6, Ra4, o-o; exd5, cxd4* *xf6* (le delni zapis poteze)
sestavljene poteze in koordinate	poimenovanje obeh polj *f4-f5, b7-b5* diagonale *a2-g8, b1-h7* (*the b1-h7 diagonal*)
kombinacije črk in drugih izrazov	*c-pawn, d6-pawn, f-file, e4-square*
imena in priimki šahistov	*Dvoretsky, Karpov, Rubinstein* (po znanih šahisti se imenujejo nekatere otvoritve, variante ipd.)
splošni izrazi	*USSR, USCF*

Tabela 1: Neterminološki izrazi z angleškega seznama izluščenih terminov.

Menimo, da je do težav z napačno zaznavo besed prišlo zato, ker so bila nekatera besedila v težje berljivem formatu. Starejše knjige vsebujejo tudi stiliziran tisk, pri katerem so zaradi poudarka ali prostorske prerazporeditve nekatere besede napisane narazen, npr. *n o r p*, te pa so bile zaznane kot večbesedni termin, čeprav nimajo pomena. Orodje *Sketch Engine* je zaradi ponavljanja informacij v glavi in nogi knjig izluščilo tudi naslove, poglavja, strani, imena igralcev idr., ki so za nas irelevantne informacije.

Najpogosteje izluščeni izrazi na seznamu so bile šahovske poteze in koordinate. To pa zato, ker so večinoma iz teh informacij sestavljene šahovske knjige, te pa so korpusu prispevale največ besed. Šahovske knjige poleg uvoda nimajo "konkretnega" besedila, kakršnega smo vajeni v člankih. Polne so diagramov s pozicijami, na podlagi katerih avtor razlaga partije, s pomočjo njih pa se učimo šahovskih otvoritev, strategije, taktik ipd.

5.2. Izbiranje terminov

Ob razvrščanju terminov v pet podpodročij (taktika, strategija, otvoritev, končnica in ostalo) smo naleteli na težave, ki so v terminologiji pogoste. Pri nekaterih enobesednih terminih smo se težko odločili, ali gre za splošni izraz ali pa je beseda šahovski termin, npr. *take* 'vzeti', *diagonal* 'poševnica', *rank* 'vrsta'. Največ težav smo imeli z ugotavljanjem, ali gre pri večbesednih terminih za svoj termin ali le kolokacijo, npr. *weak pawn* 'šibek kmet', *isolated pawn* 'osamljeni kmet', *center square/central square* 'sredično polje'.

Pozorni smo bili tudi na termine, pri katerih je prišlo do skladenjskih, pomenskih ali oblikoslovnih variant. Izrazi *to defend*, *defense, defensive* 'braniti, obramba, obrambni' se vsi uporabljajo zelo pogosto, tudi v različnih kontekstih (npr. *defense* je lahko del imena otvoritve; *Sicilian Defense* 'sicilijanska obramba'). Težko je določiti, ali gre pri različnih besednih vrstah za samostojne termine ali pa so variante enega termina.

Nekateri termini opisujejo pojav, figuro, premik ipd., ki se lahko uvrsti v več podpodročij. Kmet je na primer pomemben v otvoritvi, središčnici in tudi končnici; z njim lahko izvedemo nekaj "posebnih potez" (*pretvorba kmeta, vzetje na prehodu*), torej sodi na več podpodročij. Nekaterih terminov pa nismo mogli uvrstiti v nobeno od predvidenih podpodročij (*črni, beli*). Težav smo se delno rešili tako, da smo uvedli podpodročje *ostalo*. V veliko pomoč sta nam bila Monika Rozman in Iztok Jelen, slednji je tudi pregledal ves glosar in nam priskrbel zanesljive terminološke vire.

5.3. Izdelava terminološke baze

Potem ko smo izbrali termine in jih razvrstili na ustrezna podpodročja, smo se lotili gradnje terminološke baze. Najprej smo v programu *SDL MultiTerm* ustvarili dvojezično bazo in v angleškem jeziku določili strukturo vnosa, ki je tudi v skladu s standardom TBX. Na raven vnosa smo dodali šahovsko podpodročje (*opening, endgame, strategy, tactics, other*), na raven jezika definicijo in opombe, na raven samega termina pa rabo, status (*obsolete, colloquial, preferred, standard, variant*) in opombe. Za opredelitev podpodročja smo se odločili zato, da lahko natančneje ločimo, kateri termini spadajo v posamezno fazo šahovske igre in kaj termin sploh predstavlja (ali gre za potezo, figuro, taktiko ipd.), za status termina pa zato, da opozorimo na neustaljenost ali žargonsko rabo nekaterih terminov.

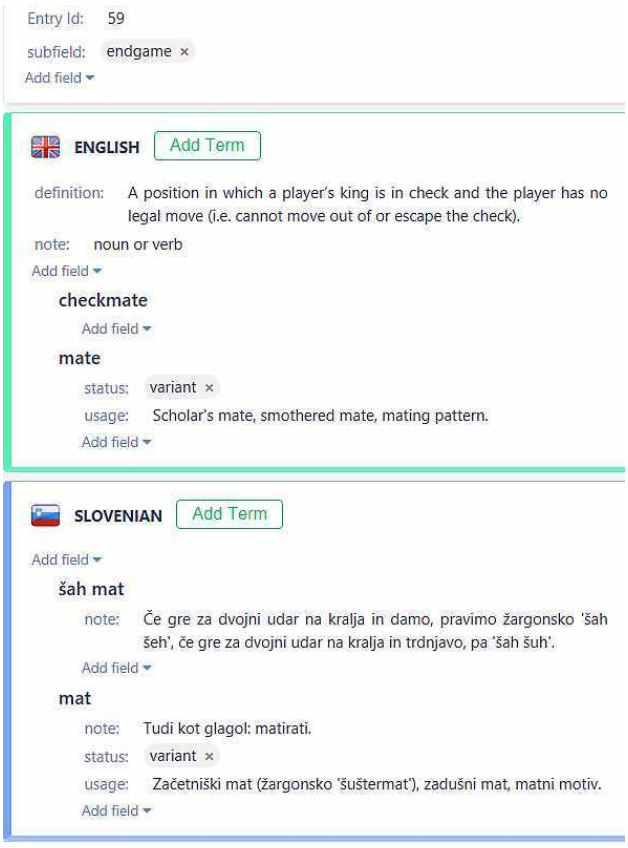
Najprej smo vnesli angleške termine in jim dodali definicije. Napisali smo jih po zgledu zanesljivih glosarjev iz šahovskih knjig, največ iz *Chess For Dummies* (Eade, 2016), in *Winning Chess Openings* (Seirawan, 2016), da so za naše potrebe bolj razumljive, pomagala nam je tudi Monika Rozman. Na podlagi korpusa smo gesla opremili še z dodatnimi informacijami (podpodročje, kolokacije idr.). Nato smo s pomočjo slovenskega korpusa in po posvetu s šahovskimi mojstri vpisali še slovenske terminološke ustreznike, kolokacije, primere rabe ipd.

Terminološka baza vsebuje 77 vnosov, od tega 82 angleških terminov s 77 definicijami in nekaterimi sopomenkami ter 109 slovenskih ustreznikov. Vsak termin vsebuje definicijo v angleščini in opredelitev podpodročja, velika večina pa tudi kolokacije, status, rabo in po potrebi opombe. Pri pregledu in širitvi baze nam je pomagal Iztok Jelen. Slovenskih definicij nismo dodali, saj nismo našli dovolj virov, ki bi vsebovali definicije za večino našega nabora slovenskih terminov, tako pa smo namesto svojega pisanja definicij to raje izpustili. Prizadevamo si, da bi v bodoče s pomočjo mojstrov in npr. Šahovske zveze Slovenije tudi to vrzel zapolnili.

Za podpodročje otvoritve smo vnesli termine, ki so značilni za ta del igre (sem spadajo tudi imena otvoritev), npr. *gambit, castling, Spanish Game, Sicilian Defense* 'gambit, rokada/rošada, španska otvoritev, sicilijanska obramba'. Pri končnicah smo vnesli termine možnih izidov igre, nekatere matne vzorce in poimenovanja določenih končnic, npr. *checkmate, stalemate, back-rank mate, Lucena position* 'šah mat, pat, mat na osnovni vrsti, Lucenova pozicija'. Podpodročje strategije obsega termine, ki jih največkrat srečamo v središčnici (*middlegame*), ko se

tvorijo pomembni strateški načrti, npr. *position, square, diagonal, file, kingside, zugzwang, tempo* 'pozicija, polje, poševnica, navpičnica, kraljevo krilo, nujnica, tempo'. Pri taktikah smo vključili osnovne taktične vzorce *fork, skewer, discovered attack, sacrifice* 'vilice, linijski udar, odkriti udar, žrtev' ipd. Dodali smo še podpodročje *ostalo*, da bi se izognili nekaterim težavam pri razvrščanju terminov na podpodročja. Tukaj zajamemo šahovske figure, posebne poteze, nazive, akronime, npr. *king, queen, promotion, grandmaster, arbiter, chessboard* 'kralj, dama, pretvorba kmeta, velemojster, sodnik/sodnica, šahovnica'.

Zapisali smo tudi pogoste kolokacije, npr. *kingside attack, strong bishop, lead in development* 'napad na kraljevem krilu, močni lovec, razvojna prednost'. Če je bila sama raba termina dvoumna, smo omenili tudi, ali se termin uporablja kot glagol, samostalnik ali pridevnik (npr. *checkmate* 'šah mat' je v angleščini lahko glagol ali samostalnik). Pri terminu *fianchetto* 'fianketo' smo dodali tudi zgled pravilne in napačne izgovorjave v angleščini: /ˌfiənˈkɛtəʊ/, */ˌfiənˈtʃɛtəʊ/ in pravilne slovenske /ˌfianˈketo/.



Slika 1: Primer terminološkega vnosa v MultiTermu.

Smo zagovorniki odprte znanosti, zato smo terminološko bazo objavili na repozitoriju CLARIN.SI (Grdič et al., 2022), kjer je prostodostopna v formatu TBX. Čeprav zajema le najpogostejše šahovske termine, jo vseeno upoštevamo kot doprinos k slovenski šahovski terminologiji. Zaradi terminoloških ustreznikov v dveh jezikih lahko služi kot pomoč prevajalcem in drugim jezikoslovcem pri pisanju besedil in raziskovanju šahovske terminologije. Prizadevamo si, da bi bazo v bodoče tudi razširili in nadgradili.

6. Pomanjkljivosti projekta

Baza je nastala na podlagi omejenega nabora virov. Slovenski korpus vsebuje le spletne vire, za dobro reprezentativnost pa bi bilo treba vključiti še nekaj knjižnih virov o različnih šahovskih podpodročjih. Na to smo sicer pazili pri angleškemu korpusu, a tudi pri njem bi bilo za boljšo reprezentativnost treba vključiti več virov.

Da bi projekt obdržali v obvladljivih razsežnostih, smo se pri končnem izboru terminov opirali na korpusno pogostost ter v bazo vključili le osnovne termine in nekatere dodatne informacije. Z večjima korpusoma ter s pomočjo več strokovnjakov in terminologov bi lahko bazo dopolnili ne le v številu terminov, temveč tudi v naboru kolokacij in primerov rabe. Naše definicije, način dodajanja in zapisa kolokacij ter ostale podatke bi moral pregledati še npr. terminolog in slovaropisec, da bi bila baza v skladu z ustaljenimi načini gradnje terminološke baze ali večjezičnega glosarja.

7. Sklep

Na podlagi korpusnega pristopa in izdelanih dveh korpusov smo ustvarili angleško-slovensko terminološko bazo, v katero smo vnesli 82 najpogosteje rabljenih angleških šahovskih terminov, jim pripisali 109 slovenskih ustreznikov ter jih opremili z definicijami, kolokacijami, primeri in informacijami o rabi.

Pri gradnji korpusa smo uporabili tako poljudne članke kot tudi specializirano gradivo, pri čemer smo stremeli k večji reprezentativnosti posameznih podpodročij. V angleški korpus smo vključili tudi knjižne vire, slovenski pa je bil omejen le na spletne.

Baza obsega nabor osnovnih terminov v obeh omenjenih jezikih. Prizadevamo si ustvariti obširnejša korpusa in izluščiti več terminov s kolokacijami in primeri rabe, dodati še slovenske definicije in sodelovati z več šahovskimi strokovnjaki, da bi bila baza v bodoče čim natančneje in pravilneje izdelana.

8. Literatura

Z. T. Adylova. 2017. System Chess Nomina of Terminological Field “Debut”. *Scientific Journal of National Pedagogical Dragomanov University. Series 9. Current Trends in Language Development*, 16:5–11. Pedagoška univerza Dragomanov, Kijev.

James Eade. 2016. *Chess For Dummies*. John Wiley & Sons, New York.

Scott Frank, režiser. *The Queen's gambit* (Damin gambit). Netflix, 2020. https://www.netflix.com/si/title/80234304.

Polona Gantar. 2004. Jezikovni viri in terminološki slovarji. V: *Terminologija v času globalizacije: zbornik prispevkov s simpozija »Terminologija v času globalizacije, Ljubljana, 5.–6. junij 2003«*, str. 169–178. ZRC SAZU, Ljubljana.

Samuel Goldman, Andrew Kwolek, Kenji Otani, Ian Ross in Jack Zender. 2021. *Chess Robot*. Univerza v Michiganu, Oddelek za strojništvo. https://deepblue.lib.umich.edu/handle/2027.42/167650.

Harry Golombek. 1980. *Šahovska enciklopedija*. Prosvjeta, Zagreb. Prevod knjige: *Golombek's Encyclopedia of Chess*. 1977. Crown publishers, New York.

Luciano Gomes de Sousa. 2021. Chess and Autism Spectrum Disorder (ASD). *Brilliant mind*, 8(4). https://revistabrilliantmind.com.br/index.php/rcmbm/article/view/52.

Vili Grdič, Alja Križanec, Kaja Perme, Lea Turšič. 2022. *English-Slovenian Chess Terminology Database 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1680.

Gari Grosar. 2017. *Šah in matematika*. Diplomsko delo, Univerza na Primorskem, Pedagoška fakulteta. https://repozitorij.upr.si/IzpisGradiva.php?id=9296&lag=eng.

Matej Guid. 2010. *Znanje in preiskovanje pri človeškem in računalniškem reševanju problemov.* Doktorsko delo, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. http://eprints.fri.uni-lj.si/1113/1/Matej__Guid.disertacija.pdf.

Joanna Harazińska in Anna Harazińska. 2017. Chess-play as the effective technique In foreign language training. *Applied Researches in Technics, Technologies and Education*, 5(3):238–242. https://www.readcube.com/articles/10.15547%2Fartte.2017.03.012.

David Hooper in Kenneth Whyld. 1992. *The Oxford Companion to Chess*. Second edition. Oxford University Press, Oxford in New York.

Iztok Jelen. 2004a. *Splošno-teoretska šahovska izhodišča izbirnega predmeta.* Skupnosti SIO. Spletna učilnica *Šah 7.–9. razred*, poglavje 6. https://skupnost.sio.si/course/view.php?id=2138.

Iztok Jelen. 2004b. *Iz teorije kombinacij.* Iz osebnega arhiva Matjaža Mikaca.

Iztok Jelen. 2006. *Šah in primerjalna analiza stanja šaha v Sloveniji.* Slovenska šahovska zveza. Iz osebnega arhiva Matjaža Mikaca.

Erik Johannson. 2021. *Chess and Twitch: Cultural Convergence Through Digital Platforms*. Magistrsko delo, Univerza v Södertörnu, School of Culture and Education, Media and Communication Studies. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1563119&dswid=6255.

Ana Jurc. 2020. *Damin gambit: kako posneti napeto nadaljevanko o šahu?* MMC RTV SLO. https://www.rtvslo.si/kultura/gledamo/damin-gambit-kako-posneti-napeto-nadaljevanko-o-sahu/543529.

Assylkhan Agbayevich Karayev. 2016. Specifics of chess terminology. *Science, technology and Education*, 6(24):102–105. LCC Olympus, Moskva.

Jana Krivec. 2021. *Improve your life by playing a game : learn how to turn your life activities into lifelong skills!* Thinkers Publishing, Landegem.

Dylan Loeb McClain. 2020. *I'm a Chess Expert. Here's What 'The Queen's Gambit' Gets Right.* The New York Times. https://www.nytimes.com/2020/11/03/arts/television/chess-queens-gambit.html.

Nataša Logar in Špela Vintar. 2008. Korpusni pristop k izdelavi terminoloških slovarjev: od besednih seznamov in konkordanc do samodejnega luščenja izrazja. *Jezik in slovstvo*, 53(5):3–17.

Leonid Pitamic. 1950. Šah v pravnem izrazoslovju. *Razprave. [Razred 2], Razred za filološke in literarne vede = Dissertationes. Classis 2, Philologia et litterae / Academia scientiarum et artium Slovenica*, 1:173–204. Slovenska akademija znanosti in umetnosti, Ljubljana.

Yasser Seirawan. 2016. *Winning Chess Openings*. Everyman Chess, London.

Šahmaty. Enciklopedičeski slovar. 1990. Sovetskaja enciklopedija, Moskva.

Marko Tratar. 2003. *Šah v slovenskem časopisu*. Diplomsko delo, Univerza v Ljubljani, Fakulteta za družbene vede. http://dk.fdv.uni-lj.si/dela/Tratar-Marko.PDF.

Milan Vidmar. 1946. *Razgovori o šahu z začetnikom*. Državna založba Slovenije, Ljubljana.

Milan Vidmar. 1951. *Pol stoletja ob šahovnici*. Državna založba Slovenije, Ljubljana.

Špela Vintar. 2017. *Terminologija: terminološka veda in računalniško podprta terminografija*. Znanstvena založba Filozofske fakultete, Ljubljana.

VIR 1 = Intervju z Matjažem Mikacem, intervjuval Vili Grdič. 4. avgust 2022, Ljubljana.

VIR 2 = *Chess.com Launches PogChamps With Top Twitch Streamers*. chess.com. https://www.chess.com/news/view/chess-com-pogchamps-twitch-rivals.

VIR 3 = Osebna korespondenca z Iztokom Jelenom, kontakt preko e-pošte. 5.–10. avgust 2022.

VIR 4 = Spletni šahovski portal ICP. Arhivirano 12. 4. 2021 na *archive.org*. https://web.archive.org/web/20210412125215/http://www.icp-si.eu/krozek/index.php?tip=glosar.

VIR 5 = *Šahovsko izrazoslovje*. Wikipedija. https://sl.wikipedia.org/wiki/%C5%A0ahovsko_izrazoslovje.

VIR 6 = *Šahovska pravila*. Wikipedija. https://sl.wikipedia.org/wiki/%C5%A0ahovska_pravila.

VIR 7 = *Šahovska strategija in taktika*. Wikipedija. https://sl.wikipedia.org/wiki/%C5%A0ahovska_strategija_in_taktika.

VIR 8 = *Slovenske šahistke v Jugoslaviji*. radiostudent.si. https://radiostudent.si/kultura/repetitio/slovenske-%C5%A1ahistke-v-jugoslaviji.

VIR 9 = Mitja Rizvič. 2016. *Avtomatsko odkrivanje zanimivih šahovskih problemov*. Diplomsko delo, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. https://core.ac.uk/download/pdf/151478793.pdf.

VIR 10 = Učni načrt za izbirni predmet šaha, spletna učilnica *Šah 7.–9. razred*. Skupnosti SIO. https://skupnost.sio.si/course/view.php?id=2138.

VIR 11 = *Chess endgame*. Wikipedia. https://en.wikipedia.org/wiki/Chess_endgame.

VIR 12 = *FIDE laws of chess*. International chess federation. https://handbook.fide.com/chapter/E012018.

VIR 13 = *Chess opening*. Wikipedia. https://en.wikipedia.org/wiki/Chess_opening.

VIR 14 = *Terms*. chess.com. https://www.chess.com/terms.

Allon Vishkin. 2022. Queen's Gambit Declined: The Gender-Equality Paradox in Chess Participation Across 160 Countries. *Psychological Science (2022)*, 33(2):276–284. https://journals.sagepub.com/doi/10.1177/09567976211034806.

Vladimir Vuković. 1978. *Škola kombiniranja.* Šahovska naklada, Zagreb.

Vladimir Vuković. 1990. *Uvod u šah na osnovi opće šahovske teorije.* Šahovska naklada, Zagreb 1990.

Irina Nikolaevna Zhuravleva in Marina Vitalevna Vlavatskaya. 2021. Structural model of chess terms in English. *Science, technology and Education*, 2(87):534–539. LCC Olympus, Moskva.

Speech-level Sentiment Analysis of Parliamentary Debates using Lexicon-based Approaches

Katja Meden[†*]

[†]Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
katja.meden@ijs.si

[*]Jožef Stefan International Postgraduate School,
Jamova cesta 39, 1000 Ljubljana

Abstract

Sentiment analysis or Opinion mining is a widely studied research area in the field of Natural Language Processing (NLP) that involves the identification of polarity (positive, negative or neutral sentiments) of the text, usually done on shorter and emotionally charged text, such as tweets and reviews. Parliamentary debates feature longer paragraphs and a very esoteric speaking style of Members of the Parliament (MPs), making them much more complex. The aim of the paper was to explore how and if lexicon-based approaches can handle the extraction of polarity from parliamentary debates, using the sentiment lexicon VADER (Valence Aware Dictionary and sEntiment Reasoner) and the Liu Hu sentiment lexicon. We performed sentiment analysis with both lexicons, together with topic modelling of positive and negative speeches to gain additional insight into the data. Lastly, we measured the performance of both lexicons, where both performed poorly. Results showed that while both VADER and Liu Hu were able to correctly identify the general sentiment of some topics (i.e., matching positive/negative keywords to positive/negative topics), most speeches themselves are very polarizing in nature, shifting perspectives multiple times. Sentiment lexicons failed to recognise the sentiment in parliamentary speeches that might not be extremely expressive or where a larger sum of intensity-boosting positive words are used to express negativity. We conclude that using lexicon-based approaches (such as VADER and Liu Hu) in their unaltered states alone do not suffice when dealing with data like parliamentary debates, at least not without any modification of lexicons.

1. Introduction

Sentiment analysis or Opinion mining is a widely studied research area in the field of Natural Language Processing (NLP) that encompasses extraction of thoughts, attitudes and subjectivity of text to identify sentiment polarity (positive, negative or neutral sentiment). Sentiment analysis is mostly used on shorter and emotionally charged text, such as tweets and reviews, though it can be used on other forms of textual data, such as parliamentary debates. Parliamentary debates are in essence transcriptions of spoken language, produced in controlled and regulated circumstance, with rich (sociodemographic) metadata (Erjavec et al., 2022).

Contrary to social media data that are usually used for sentiment analysis (tweets and other shorter social media-based text), parliamentary debates and thus parliamentary discourse vary from political environment and culture, text (or rather, speeches) itself is longer and made by the parliamentary representatives under strict(er) procedural-themed language. This alone makes parliamentary debates as an object of sentiment analysis more complex in comparison to tweets or reviews, where opinions and sentiments are usually expressed much more clearly and in the shorter span of text. The sentiment analysis for this paper was implemented on the HanDeSet parliamentary corpus that includes 1251 motion-speech units from 129 debates with manually annotated sentiment labels.

The aim of this paper is to explore lexicon-based approaches on the basis of parliamentary debates using lexical (and rule-based) approach VADER (Valence Aware Dictionary and sEntiment Reasoner) and Liu Hu sentiment lexicon to see how (and even if) lexical-based methods are able to handle sentiment analysis of longer, more complex textual data such as parliamentary debates. To complement this research question, we performed sentiment analysis with both lexicons, together with topic modelling of positive and negative sentiment clusters to gain additional insight into the data. Lastly, we measured performance of both lexicons and examined reasons for any possible misclassifications.

The paper is structured as follows: In Section 2 we present related work on sentiment analysis, VADER and Liu Hu sentiment lexicons as well as studies done on researching sentiment on parliamentary debates. In Section 3 we present the chosen methodology for our work, together with presentation of the chosen dataset *Hansard Debates with Sentiment Tags — HanDeSet*. Section 4 includes the presentation of the results of the sentiment analysis with the chosen lexicons, topic modelling results, as well as their performance. Lastly, in the Section 5 we present our conclusions and pointers for future work.

2. Related work

2.1. Sentiment analysis and lexicon-based approaches

There are several methods of applying sentiment analysis, which are divided into three approaches: supervised, lexicon-based and hybrid approaches (Catelli et al., 2022), each with its own set of advantages and disadvantages.

The lexicon-based approaches utilize sentiment lexicons to describe the polarity (positive, negative and neut-

ral) of the text. This approach involves manual construction of lexicons with positive and negative words to be used in sentiment analysis and corpus of text to which the sentiment analysis will be applied. The main advantages of this approach are the fact that they are easier to understand and have wider-term coverage, while the disadvantages lay in a finite number of words in the lexicons (i.e., we cannot cover all of the words, especially if the text is domain-specific) and the assignation of a fixed sentiment orientation and score to words - every word in the lexicon is classified as positive or negative with a numeric score, e.g., on the scale of -5 (very negative) to 5 (very positive), with 0 annotating neutrality of the text. For this paper, we will be focusing on two specific lexicon (and rule-based) approaches from the natural language toolkit (NLTK): VADER and the Liu Hu sentiment module.

2.2. VADER (Valence Aware Dictionary and sEntiment Reasoner)

VADER is established as a gold-standard sentiment lexicon that is attuned to microblog-like contexts. It is primarily designed for Twitter and other social media text (as well as editorials, movie and product reviews). VADER sentiment module was implemented in NLTK.[1] The aim of the authors was to provide computational sentiment analysis engine that works well on social media style text, yet readily generalizes to multiple domains and requires no training data, but is constructed from a generalizable, valence-based, human-curated sentiment lexicon (Hutto and Gilbert, 2014). The VADER sentiment lexicon is comprised of 7,500 lexical features with validated valence scores that indicate both the sentiment polarity (positive/negative) and the sentiment intensity on a scale from –4 to +4. For example, the word *okay* has a positive valence of 0.9, *good* is 1.9, and *great* is 3.1, whereas *horrible* is –2.5, the frowning emoticon *:(* is –2.2, and *sucks* and it's slang derivative *sux* are both –1.5 (Hutto and Gilbert, 2014).[2]

In context of parliamentary debates, VADER has been used in several different studies, such as in (Rohit and Singh, 2018), where VADER was used to extract sentiment polarity, as it uses a simple rule-based model for general sentiment analysis and generalizes more favorably across contexts than any of many benchmarks such as LIWC and SentiWordNet.

2.3. Liu Hu sentiment module

Liu Hu sentiment lexicon is a product of the research by Hu and Liu, where authors aimed to summarize all the customer reviews of a product. Contrary to the traditional summarization tasks they only mined reviews where customers have expressed their opinion on the product, trying to determine whether the opinions expressed were positive or negative (Hu and Liu, 2004). Liu Hu opinion lexicon is publicly available and consists of nearly 6,800 words (2,006 with positive semantic orientation, and 4,783 negative).[3]. The opinion lexicon has evolved over the past decade, and is, similarly to VADER, more attuned to sentiment expressions in social text and product reviews – though it still does not capture sentiment from emoticons or acronyms/initialisms (Hutto and Gilbert, 2014). The Liu Hu sentiment lexicon has been implemented in the NLTK library as a Liu Hu sentiment module (nltk.sentiment.util module),[4] where function simply counts the number of positive, negative and neutral words in the sentence and classifies it depending on which polarity is more represented. Words that do not appear in the lexicon are considered as neutral[5].

2.4. Parliamentary debates

Recently, parliamentary debates have raised an interest of researchers from various academic disciplines, especially as an object of linguistic research (Erjavec et al., 2022). Transcriptions are done by professional stenographers, familiar with the procedures, as well as with the Members of Parliament (Truan and Romary, 2021). Parliamentary discourse is shaped by the specific rules and conventions, which are in turn shaped by the socio-historical traditions that influence the organisations and operations of the Parliament. These conventions and traditions extend to language use, e.g., turn-taking or forms of address (Fišer and de Maiti, 2020). Another characteristic of the transcriptions is the fact that officially released records of parliamentary debates are not verbatim and that minute-taking varies across countries and history as well. The editing process can include elimination of obvious language or factual errors, dialectal or colloquial expressions and rude and obscene language. This, combined with the fact that editing guidelines are mostly not publicly available, can hinder research (Truan and Romary, 2021).

The main characteristics of parliamentary discourse in the UK Parliament stem from previously mentioned composition and operations of the Parliament - the UK Parliament consists of two Houses: the House of Commons and the House of Lords, where the decisions made in one House have to be approved by the other. (Parliament, 2022). The House of Commons parliamentary debates consist of three substantial elements (Abercrombie and Batista-Navarro, 2018b):

Debates are initiated with a motion —– a proposal made by an MP. When invited by the Speaker (the presiding officer of the chamber), other MPs may respond to the motion, one or more times. Lastly, the Speaker may call a division, where MPs vote by physically moving to either the '*Aye*' or '*No*' lobby of the chamber. These divisions may be called at any time, but typically occur at the end of the

[1] https://www.nltk.org/api/nltk.sentiment.vader.html

[2] The entire VADER lexicon is available at https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vader_lexicon.txt

[3] The entire Liu Hu lexicon was available on https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[4] https://www.nltk.org/api/nltk.sentiment.util.html

[5] List of positive and negative words in the lexicon can be found at https://github.com/woodrad/Twitter-Sentiment-Mining/tree/master/Hu%20and%20Liu%20Sentiment%20Lexicon

debate. Example from the corpus shows the structure of the units:

Motion: *That there shall be an early parliamentary general election.*

Speech: *Does my right hon. Friend agree that the Prime Minister, in calling this election, has essentially said that she does not have confidence in her own Government to deliver a Brexit deal for Britain? One way in which she could secure my vote and the votes of my hon. Friends is to table a motion of no confidence in her Government, which I would happily vote for.*

Vote: *'Aye' (positive).*

3. Methodology

3.1. Dataset

HanDeSeT: Hansard Debates with Sentiment Tags is a corpus that contains English parliamentary debates from 1997 to 2017 with 1251 motion-speech units taken from 129 separate debates and manually annotated with sentiment scores. The corpus itself was compiled from the *UK Hansard parliamentary corpora*. Transcripts are largely-verbatim records of the speeches made in both chambers of the UK Parliament in which repetitions and disfluencies are omitted, while supplementary information such as speaker names (speaker metadata) are added (Abercrombie and Batista-Navarro, 2018b).

The HanDeSet corpus features 1251 motion-speech units, where each unit comprises a parliamentary speech of up to five utterances and an associated debate motion. As detailed in (Abercrombie and Batista-Navarro, 2018b), parliamentary debates incorporate "much set, formulaic discourse related to the operational procedures of the chamber", i.e. speech segments used to thank the Speaker or describing the activities in the chamber.

Each speech-motion unit has several sentiment polarity labels:

- *manual speech* : manually assigned sentiment label of the speech (0 = negative, 1 = positive)
- *manual motion*: manually assigned sentiment label of the motion (0 = negative, 1 = positive)
- *gov/opp motion*: label on the relationship of the MP (who proposes the motion) to the Government (i.e. whether the MP is in Government or not: 0 = is not in Government, 1 = is in Government)
- *speech vote*: a speaker-vote label extracted from the division associated with the corresponding debate (i.e. how the MP voted to proposed motion: 0 = negative, 1 = positive)

Since our research scope covers only the parliamentary speech and the sentiment of it, we will be focusing on the *manual speech* labels.

3.2. Data cleaning and pre-processing

As extraction of polarity (or sentiment) score can heavily depend on certain text characteristics, pre-processing text data can impact the performance of the lexicon-based modules severely. As detailed in (Hutto and Gilbert, 2014), there are five generaliseable sentiment intensity characteristics: punctuation (specifically, the exclamation mark "!"), capitalization (e.g., using all caps in a text), amplifying the intensity of the text with mood booster words (e.g., using words like *extremely* or *very*) or using a combination of all of these characteristics (e.g., *"The food here is EXTREMELY GOOD!!!"*). In regard to this, we pre-processed the text using only tokenization (and keeping the punctuation) and lemmatization (using UDPipe Lemmatizer).

3.3. Experiment settings

Most work was done in the Orange Data Mining Tool[6]. Both VADER and Liu Hu sentiment modules are both already incorporated in the *Sentiment analysis* widget in Orange.

3.3.1. Sentiment analysis and performance comparison

Semantic analysis was performed on the speeches (with both VADER and Liu Hu sentiment modules). VADER outputs several scores for the semantic analysis: *pos*, *neg*, *neu* and *compound*. The *compound* feature is the combined score of all of the other features and our main indicator of sentiment in text. For Liu Hu, the score shows difference between the sum of positive and sum of negative words, normalized by the length of the document and multiplied by a 100. The final score reflects the percentage of sentiment difference in the document (Demšar et al., 2013). It is important to note that the lexicons were not modified in any way.

Next we mapped the sentiment scores, output by both sentiment modules to their respective labels: positive and negative. This was done to match the scores in the gold standard, where each speech is labelled with either 0 for negative or 1 for positive (and where neutral sentiment labels do not exist). Therefore, the main problem of mapping these labels stemmed from speeches and motions, that had a score of "0" (and are thus regarded as neutral) that needed to be mapped either as positive or negative.

After inspecting the dataset and the distributions of the positive and negative class in the dataset (presented in the Table 1), where it can be seen that the distributions for manually applied sentiment labels for speeches are slightly skewed towards the positive class, with the positive class counting 705 speeches (56.4%) and the negative of 545 (43.6%) speeches. Therefore, we decided to map these speeches as *positive*, in favor of the majority class. After obtaining the labels (positive/negative), the last step was to compare the results of the sentiment analysis to the gold standard (and our test dataset) with classification accuracy and F1 score evaluation metrics. To compare our results, a majority class baseline was added.

3.3.2. Descriptive analysis and topic modelling

As previously stated, our research aimed not only to evaluate the performance of both sentiment lexicons but to research the sentiment in the UK parliamentary debates. In regard to this, we also applied topic modelling to extract additional information on the topics of the analyzed

[6] https://orangedatamining.com/

parliamentary speeches. Descriptive analysis of the results provided by the VADER and Liu Hu sentiment modules on parliamentary debates enables insight into the positive speeches, resemblances and reasons for possible differences between the results of the lexicons.

The results of the sentiment analysis are presented with histogram of sentiment scores of both sentiment lexicons (compound score for VADER and sentiment score by Liu Hu) to visualize the distributions of positive and negative scored speeches. Deriving from this we also performed topic modelling on subsets of positive and negative speeches to identify topics and see if they correspond to the general sentiment of the topic that the keywords belong to.

To facilitate topic modelling, speeches first needed to be pre-processed: transformed to lowercase, tokenized, lemmatized with UDPipe Lemmatizer. Lastly, stopwords were filtered out list of stopwords, provided from NLTK and with a manually compiled additional list of stopwords[7] for the procedural words, that are very common in (procedural) parliamentary speech.

For topic modelling we used *Latent Dirichlet Allocation* method to extract keywords of speeches and its topics. As LDA does not give the optimal number of topics for the text itself, the exact number of topics needs to be determined by the model user (Gan and Qi, 2021). We, therefore, experimented with different numbers of topics in the range from 5 to 11, with the *Topic Coherence* metric serving as our pointer. This specific range of topics was chosen to facilitate high enough granularity of the keywords in the topics (i.e., no less than 5 topics) but at the same time keep the coherence of the keywords in the topics. Topic coherence score represents the "degree of semantic similarity between high-scoring words in the topic to help distinguish between topics that are semantically interpretable and topics that are artifacts of statistical inference" (Stevens et al., 2012). Table 1 shows the Topic Coherence score fluctuation in different settings for all chosen subsets (positive and negative clusters produced by VADER and Liu Hu), with numbers in bold representing the optimal number of topics for the subset.

Number of Topics	VADER positive	VADER negative	Liu Hu positive	Liu Hu negative
5	0.281	0.244	0.267	0.252
6	0.272	0.256	0.275	0.244
7	0.263	0.282	0.264	0.250
8	0.268	0.276	0.275	0.260
9	0.251	0.260	0.265	0.256
10	0.265	**0.303**	**0.276**	**0.279**
11	**0.284**	0.270	0.265	0.259

Table 1: Topic Coherence scores of the positive and negative subsets and their optimal number of topics.

The topics, identified with the LDA method are visualized with MDS (Multidimensional scaling)), where the size of the topic indicates *Marginal Topic Probability* (i.e. how representative a topic is to a corpus or a cluster). To get the naming of the topics as accurate as we could, we used several Orange widgets: *t-SNE widget* for the 2-D projection of the speeches with similar topics, *Extract keywords* widget to extract 5 most common keywords in those speeches and *Score documents* widget to identify the names of the documents the keywords occur in most often, inferring the topic name from the title and content of the documents.

[7]Additional list of stopwords is available at: https://drive.google.com/file/d/16kH_dV8HlUhctwmmsLn4F9zOkmJyqgg5/view?usp=sharing

4. Results

4.1. Sentiment analysis results

In this section we present the results of the sentiment analysis, done with VADER and Liu Hu. Figure 1 compares the distributions of positive and negative speeches, identified by VADER (Figure 1a) and Liu Hu (Figure 1b) sentiment lexicons.

Even at first glance, we can see that VADER results are leaned heavily towards the positive class. The compound score ranges from 0.9987 (score of the most negative speech) to 0.9992 (score of the most positive speech). Most speeches in the dataset (617 speeches, 49.32%) were classified by VADER as extremely positive in the range from 0.8 to 1 of the compound score. On the other hand, only 124 speeches (9.91%) were deemed extremely negative in range from -0.8 to -1.

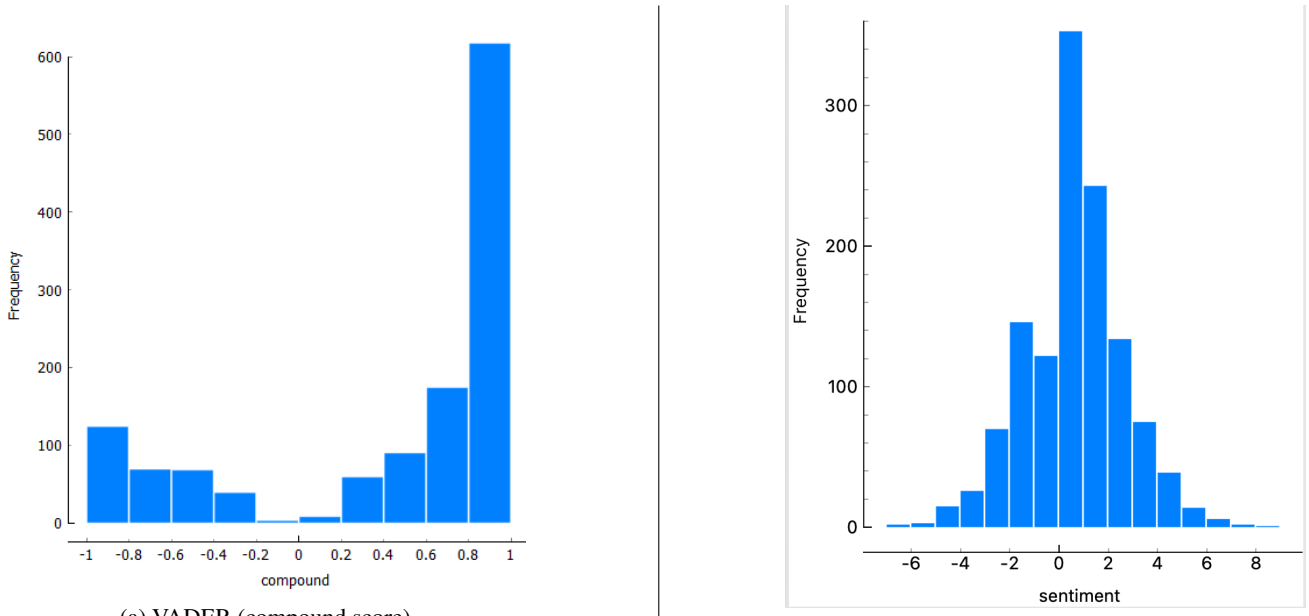
Figure 1b represents results obtained by using Liu Hu sentiment lexicon. While VADER uses a scale from -1 to 1, Liu Hu computes the sentiment score by preserving 0 as the neutral value and deems everything below 0 as negative and above as positive sentiment. As it can be seen from the figure, the distribution of sentiment in the speeches differs greatly from the VADER results. The most negative speech has a sentiment score of -6.976, the most positive a score of 8.1967, with most speeches (353 speeches, 28.22%) positioned on a sentiment score spectrum from 0 to 1. Out of those, 216 speeches were scored with 0 (neutral speeches).

In its entirety, more than 75% of the speeches were deemed positive by VADER (984 speeches, 75.78%). Similarly, Liu Hu deemed positive almost 70% of the speeches (867 speeches, 69.30%) For the topic modelling, each set was split into a positive and negative subset:

- VADER subset of positive speeches: 948 speeches (75.78%)
- VADER subset of negative speeches: 303 speeches (24.22%)
- Liu Hu subset of positive speeches: 867 speeches (69.30%)
- Liu Hu subset of negative speeches: 384 speeches (30.70%)

4.2. Topic modelling results

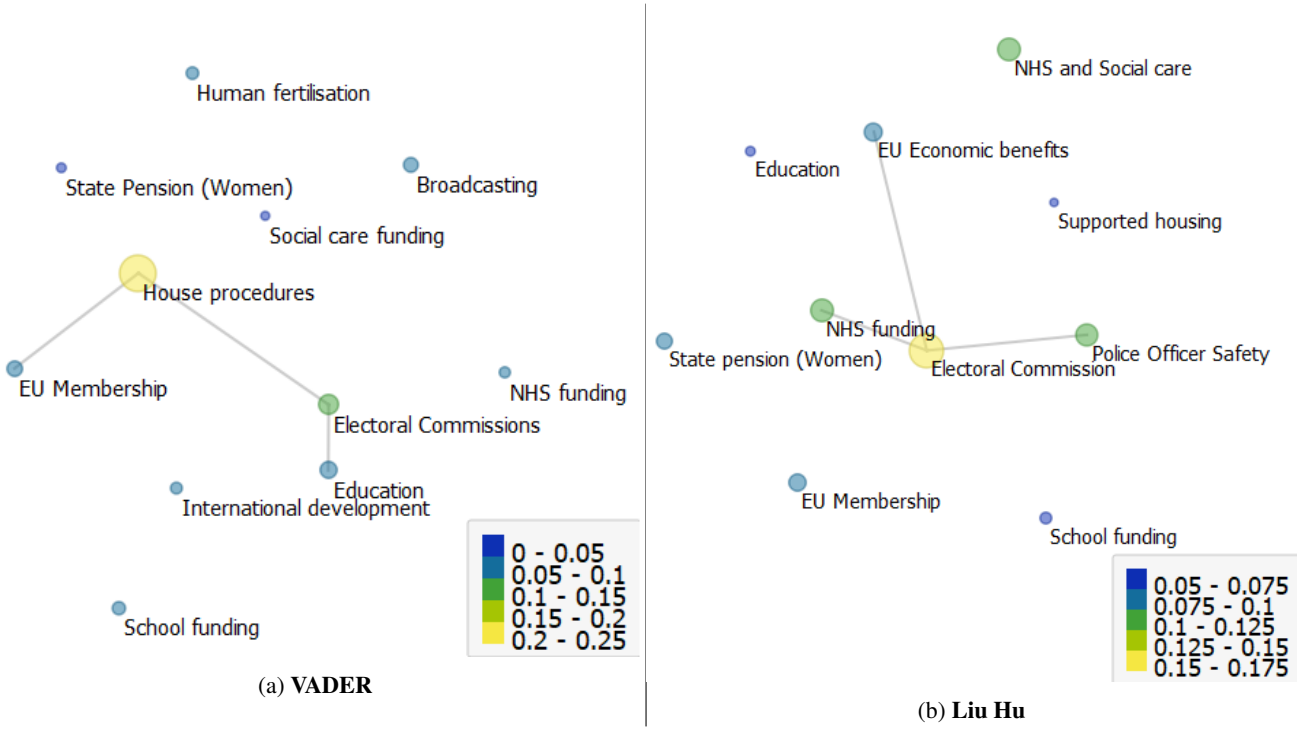
The results are presented in two parts, using MDS to aid in visualization of the topics and their labels. The first part focuses on comparison of the topics in both positive



(a) VADER (compound score)

(b) Liu Hu (sentiment score)

Figure 1: Results of the sentiment analysis and distribution of positive and negative speeches.



(a) **VADER**

(b) **Liu Hu**

Figure 2: Comparison of topics, identified in the positive speeches between VADER and Liu Hu.

clusters, while the second one presents identified topics and trends in the negative clusters.

As it can be seen from Figure 2a and 2b, the largest clusters of keywords detected among the positive speeches, produced by VADER, belong to the topic *House procedures*[8], where the topic consists of very common words throughout the corpora, e.g., *member*, *house*, *bill*, *parliament*, etc. In Liu Hu produced results, the largest topic is relatively similar to the House procedures, that being *Electoral Commission*, where most keywords, emphasised above are still present, with two explicit keywords that define the nature of the topic - *election* and *change*. Both topics are also linked together (MDS enables linking of semantically similar topics together), which makes the closeness of the keywords in both topics even more clear. Topic *Electoral Commission* appears in both positive clusters. In addition to the aforementioned *Electoral Commission*, topics like *EU*

[8]Full name of the documents, that contain most of the keywords in the topic corresponds best to *The Business of the House*, thought the name of the topic was shortened for easier visualization.

membership, *School funding* and *NHS funding* also appear in both positive speeches.

The keywords and topics, identified in the negative speeches are shown in Figure 3a and 3b.

With the Marginal Topic Probability score of 0.175, the most common keywords in the VADER negative subset are found in topic *State pension age*, followed closely by *Armed forces* (score of 0.172), *Prisons and probation* (0.150) and *Police Officer Safety*. MDS also showed that several topics are also very closely related to one another, e.g., Topic *Armed forces* is closely related to both *House procedures* and *Terrorism bill* topics. Similarly, although not surprising, a strong connection is also found between keywords in *State pension (Women)* and *State pension age (Women)*. Lastly, strong similarity is shown between keywords in *Police Officer Safety* and *Prisons and Probation*. In the Liu Hu negative speeches, the most represented topic is *State pension (Women)* with the Marginal Topic Score of 0.163, followed closely by *EU Membership* with the score of 0.159 and *Homelessness* with 0.114. All three topics (or, rather, their keywords) are also connected amongst themselves. For both VADER and Liu Hu negatively scored speeches, the keywords most present in them are found in topic on state pension and state pension age (very connected topics that share many common keywords). In addition to that, several other topics can be found in both subsets, e.g. *Armed forces*, *Police Grant* and *House procedures*.

In general, the keywords of the topics identified mostly corresponded to the general sentiment of the topics in their respective subsets. Even though, in several cases, keywords (and topics) appeared both in the positive as well as in the negative speeches. This is most likely due to the fact that parliamentary debates usually feature heavy position-taking in regard to a certain motion.

The topics in the negative speeches were harder to identify in comparison to the positive speeches - this is mostly due to the larger subset, as well as the fact that the keywords were very fragmented. This can be seen in the positive clusters, where the Marginal Topic Score of most topics (aside from the two or three very well represented ones) are not high and are in lowest score range. While in general the topics were harder to identify, most topics that were strongly present in the speeches had very obvious keywords. On the other hand, topics in the positive speeches were easier to identify, although, there were some exceptions, as some of the keywords (even though many stopwords were removed) were too general to pinpoint with human perception alone.

4.3. VADER and Liu Hu performance evaluation

To evaluate the performance of the sentiment modules we used the following evaluation metrics: classification accuracy and F1 score. Similarly, a related research (Abercrombie and Batista-Navarro, 2018b) used the dataset to develop a 2-step model for sentiment analysis task - they trained SVM and MLP to produce a one-step *Speech* model and a two-step *Motion-Speech* model, using different features (text only, text and metadata). The results for the one-step *Speech* model with text-only features (evaluated with a 10-fold validation) were added to the Table 2 for comparison.

	Acc(%)	F1 score
VADER	52.0	0.49
Liu Hu	50.0	0.47
Baseline	56.5	0.56
SVM (text only)	66.7	0.718
MLP (text only)	67.3	0.713

Table 2: Performance results with VADER and Liu Hu, accompanied with the baseline and results for SVM and MLP from the related study.

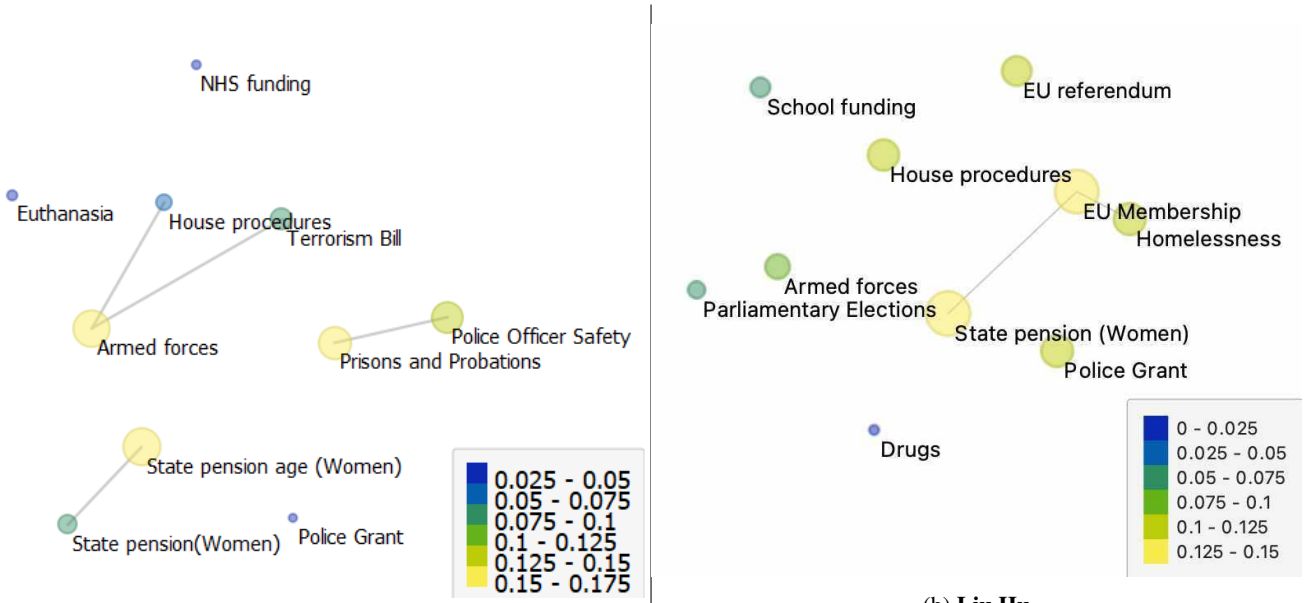
The performance of the VADER and Liu Hu sentiment lexicons is poor, not even surpassing the baseline score. However, if we want to put the results in a perspective, we need to consider the nature of parliamentary debates and parliamentary language. The language of parliamentary debates is, as we stated previously, complex - the speeches especially are longer and full of visible political procedure characteristics (such as courtesy naming, e.g., hon. Friend, hon. Lady ...).

Very poor performance scores show that sentiment lexicons (in their current, unmodified state) are not the best methodology when it comes to extracting sentiment polarity in parliamentary debates. In comparison, study, detailed in (Abercrombie and Batista-Navarro, 2018a) achieved much greater results even by using just the text features (as shown in Table 2).

To research the reason for such poor performance, we analysed several speeches in detail. Below is an example and one of the possible explanations for misclassifications:

"Our national health service is, and always has been, ***valued*** *and* ***cherished*** *by my constituents who rightly expect an* ***excellent*** *standard of care to be provided* ***free*** *at the point of use when they need treatment. We are all deeply* ***committed*** *to the future of the NHS, but to* ***ensure*** *that it can continue to provide the quality of care that our constituents expect, it cannot stand still. [...] What is certain is that the current model through which health services in Calderdale and Huddersfield are delivered is not sustainable in the long term, and that changes are needed to ensure that we have a local health service that continues to provide* ***excellent care****."*

The speech itself contains words that could influence the scoring in a positive way - VADER scored this speech with 0.9992 (making it one of the most positive speeches identified by VADER), while Liu Hu scored it with 1.578 . Words in bold are all included in the VADER lexicon with high positive scores; e.g., *committed* has a score 1.1, *valued* of 1.9, *cherished* of 2.3 and *excellent* of 2.7. Therefore, the speech could have been perceived as positive, even though the entire speech is in reality negative, as it emphasises that the current model of health services is not long-term sustainable. Similarly, Liu Hu includes words *cherished*, *quality*, *free* and *excellent* in the list of positive words, but it does not include words like *valued* or *committed* (and thus making them neutral). The sentiment of this text is, accord-



(a) **VADER** (b) **Liu Hu**

Figure 3: Comparison of topics in negative speeches between VADER and Liu Hu.

ing to Liu Hu, still positive - less than with VADER, but the process and reason for misclassification is mostly the same.

5. Conclusions

In this paper we used sentiment based approaches (VADER and Liu Hu) on the base of parliamentary data with the aim to explore how these two modules handle sentiment detection on longer, less expressive and more formal language to that of the (usually) used social media language (for which both sentiment modules are optimized for). While the both VADER and Liu Hu were able to correctly identify the general sentiment of some topics, present in negative and positive clusters (e.g., matching keywords in the *Euthenasia* topic to the negative cluster), the speeches themselves are very polarizing in nature. This can most clearly be seen in the fact, that some topics were identified in both positive and negative clusters, e.g., topics like *School funding* and *NHS funding* were identified in both positive and negative speeches, as both can be viewed from different (positive or negative) standpoints.

The most probable reason for misclassifications is the length of the speeches, as well as the matter of speeches not being extremely expressive or having a bigger sum of positive boosting words used to express negativity. The language of parliamentary discourse can be extremely complex, mostly due to the esoteric speaking style and opaque procedural language of Parliament (Abercrombie and Batista-Navarro, 2018b). Distinguishing between a positive and negative polarity of parliamentary debates can be a difficult task even for human annotators, which was proven by the poor inter-annotator agreement score in the first round of annotation of the HanDeSet dataset, detailed in (Abercrombie and Batista-Navarro, 2018a). Similar can be said for lexicon-based approaches to sentiment analysis, though despite the poor performance scores, the lexicons still gave us some insight into the general sentiment around topics and parliamentary speech characteristics. As it can be seen from the poor performance evaluation results, sentiment-based approaches like Liu Hu and VADER alone do not suffice when dealing with such a specific text data, at least not in their unmodified state. Better results could have possibly been acquired by modifying the lexicons to incorporate some of the characteristics of parliamentary debates (e.g., adding new words and changing the scoring of existing ones).

6. Acknowledgments

The paper was written in the framework of the research programme P2-0103 (B): Tehnologije znanja (Knowledge Technologies), co-financed by the Slovenian Research Agency (ARRS) from the state budget and the Slovenian research infrastructure CLARIN.SI (Common Language Resources and Technology Infrastructure, Slovenia).

7. References

Gavin Abercrombie and Riza Batista-Navarro. 2018a. 'Aye' or 'No'? Speech-level Sentiment Analysis of Hansard UK Parliamentary Debate Transcripts. In: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, eds., *Proceedings of the Eleventh International Con- ference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gavin Abercrombie and Riza Theresa Batista-Navarro. 2018b. A Sentiment-labelled corpus of Hansard Parliamentary Debate Speeches. In: D. Fišer, M. Eskevich, and F. de Jong, eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018 - ParlaMint II Workshop)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rosario Catelli, Serena Pelosi, and Massimo Esposito. 2022. Lexicon-based vs. BERT-based sentiment analysis: A comparative study in Italian. *Electronics*, 11(3):374.

Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, Miha Štajdohar, Lan Umek, Lan Žagar, Jure Žbontar, Marinka Žitnik, and Blaž Zupan. 2013. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14:2349–2353.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The Parlamint corpora of parliamentary proceedings. *Language resources and evaluation*, pages 1–34.

Darja Fišer and Kristina Pahor de Maiti. 2020. Voices of the Parliament. *Modern Languages Open*.

Jingxian Gan and Yong Qi. 2021. Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an example. *Entropy*, 23(10):1301.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Clayton Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the international AAAI conference on web and social media*, pages 216–225.

UK Parliament. 2022. *The two-House system*.

Sakala Venkata Krishna Rohit and Navjyoti Singh. 2018. Analysis of speeches in Indian parliamentary debates. *arXiv:1808.06834*.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 952–961.

Naomi Truan and Laurent Romary. 2021. Building, Encoding, and Annotating a Corpus of Parliamentary Debates in XML-TEI: A cross-linguistic account. *Journal of the Text Encoding Initiative*.

Evalvacijska kategorizacija strojno izluščenih protipomenskih parov

Tina Mozetič,[*] Miha Sever,[*] Martin Justin,[*] Jasmina Pegan[‡]

[*] Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
tina.mozetic11@gmail.com, mihasever98@gmail.com, martin1123581321@gmail.com
[‡] Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
jp2634@student.uni-lj.si

Povzetek

Namen prispevka je oceniti relevantnost strojno pridobljenih protipomenskih parov za vključitev v razširjeni Slovar sopomenk sodobne slovenščine. Nekdanje strukturalistično pojmovanje protipomenskosti vedno bolj prehaja k sodobnejšemu, ki temelji na naprednih računalniških metodah, odprtosti, množičenju, relevantnosti in uporabnosti podatkov. V raziskavi smo pregledali 2852 strojno izluščenih parov protipomenk. Primeri, ki jih označevalci niso enoznačno uvrstili med protipomenske oziroma neprotipomenske, so razvrščeni v 21 kategorij. Za protipomenke vsake kategorije je opredeljeno, ali jih je smiselno vključiti v odzivni slovar. Strojni postopek se je izkazal za uspešnega, saj je v slovar mogoče vključiti 88 % izluščenih parov. Kategorije bodo v prihodnosti uporabne tudi za oblikovanje smernic ter razvoj nadaljnje metodologije strojnega luščenja protipomenk.

Evaluative Categorisation of Automatically Extracted Pairs of Antonyms

This paper aims to assess the relevance of extracted antonym pairs that are to be included in the expanded Thesaurus of Modern Slovene. The former structuralistic conception of antonymy is shifting to a more modern one that is based on advanced computational methods, openness, crowdsourcing, relevance, and data usability. In this study, we reviewed 2852 extracted pairs of antonyms. Examples that were not uniquely classified as antonyms or non-antonyms by the evaluators are grouped into 21 categories. For each category, it is determined whether they should be included in the responsive dictionary. The process proved to be successful, as 88% of the extracted pairs could be included in the dictionary. The categories will also be useful in the future for the creation of guidelines and the development of further methodologies for automatic extraction of antonyms.

1. Uvod

Slovar sopomenk sodobne slovenščine je s 105.473 iztočnicami in 368.117 sopomenkami »najobsežnejša prosto dostopna avtomatsko generirana zbirka sopomenk za slovenščino« (Sopomenke 1.0, 2022). Slovar deluje po principu odzivnega slovarja, ki je v prvem koraku pripravljen povsem strojno. Strojno pripravljeni podatki so objavljeni takoj, ko jezikoslovna evalvacija potrdi njihovo načelno ustreznost oz. relevantnost za skupnost, nato pa se slovar razvija naprej po korakih in v sodelovanju jezikoslovcev in širše zainteresirane javnosti (Arhar Holdt et al., 2018). Pri projektu Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL bomo sopomenkam dodali protipomenke, za katere je treba opraviti tovrstno jezikoslovno evalvacijo relevantnosti.

Cilj našega prispevka je tako oceniti relevantnost strojno pridobljenih protipomenskih parov za vključitev v razširjeni Slovar sopomenk sodobne slovenščine. Pri tem nas zanima predvsem, kateri del podatkov je (1) primeren za neposredno vključitev v slovar, (2) kateri za vključitev ni primeren in (3) kateri zahteva dodaten premislek. V prispevku se natančneje ukvarjamo s tretjo točko, pri čemer dokazujemo, da je »problematične« primere mogoče kategorizirati glede na vrsto problema in tako določiti, ali jih je (a) mogoče izboljšati strojno, ali (b) morda zahtevajo uredniško odločitev, (c) jih je mogoče izboljšati s pomenskim členjenjem gesla ali kvalifikatorji, (d) jih je mogoče izboljšati s pomočjo skupnosti oziroma (e) kljub določenemu problemu pustili v naboru slovarskega gradiva in računati na to, da bodo uporabniki sami presodili o njihovi uporabnosti.

Problemske kategorije, ki jih bomo tako oblikovali, bodo služile kot izhodišče za nadaljnje delo na projektu, ki obsega nadgradnjo metodologije luščenja, pripravo smernic za uredniško obravnavo protipomenk in vključitev protipomenk v Slovar sopomenk sodobne slovenščine. Ročno pregledani protipomenski pari bodo uporabljeni kot učna množica za nadaljnje luščenje protipomenk iz korpusa Gigafida 2.0 (Krek et al., 2020). Tudi pri oblikovanju smernic pa bo naša analiza prišla zelo prav, saj smo identificirali probleme, za katere bo treba v nadaljevanju podati tudi načelne uredniške rešitve.

V drugem razdelku prispevka tako najprej predstavimo jezikoslovne raziskave protipomenskosti in koncept odzivnega slovarja. V tretjem na kratko opišemo metodi pridobivanja in označevanja podatkov. V četrtem razdelku pa predstavimo rezultate označevanja in jih analiziramo. Najprej predstavimo odločitve označevalcev glede ustreznosti protipomenskih parov, nato pa natančneje predstavimo vsako od problemskih kategorij, v katere so bili v fazi označevanja uvrščeni »problematični« primeri. Pri vsaki kategoriji predstavimo tudi njeno pogostost in ocenimo, na kakšen način bi bilo identificirani problem mogoče reševati. V zaključnem delu povzamemo bistvene ugotovitve prispevka.

2. Pregled področja

Jezikoslovje smatra protipomenskost – poleg sopomenskosti – za temeljno medleksemsko pomensko razmerje (Stramljič Breznik, 2010; Humar, 2016; Vidovič Muha, 2005, 2021). V nasprotju s sopomenkami protipomenke nujno nastopajo binarno, tj. v parih, in so vedno del skupnega pojmovnega ali celo pomenskega polja (Vidovič Muha, 2021). V slovenskem izrazoslovju sta se

enakovredno ustalila izraza protipomenka in antonim oz. protipomenskost in antonimija, čeprav Slovenski pravopis 2001 prednost daje izrazu protipomenka (Humar, 2005).

Definiranje protipomenke je razmeroma enostavno. Protipomenka je po SSKJ (2014) »beseda z nasprotnim pomenom v odnosu do druge besede«, enako jo opredeljuje tudi Toporišič (2001). Marjeta Humar (2016) definicijo razširi na »poimenovanja pojmov z eno- ali večpomensko besedo ali besedno zvezo, [pri katerem] sta v protipomenskem razmerju pomenski sestavini pojmov (navadno po ena pri vsakem od dveh), izraženih z enopomenskima besedama, z enopomenskima besednima zvezama ali pa s posameznima pomenoma dveh večpomenskih besed ali zvez« (22).

V nasprotju z definiranjem pomenska tipološka razvrstitev protipomenk predstavlja veliko oviro; tovrstnih razvrstitev je namreč toliko, kolikor je znanstvenikov, ki so se z njimi ukvarjali. Problematike se zavedajo tudi jezikoslovci sami (gl. Humar, 2016), njihova glavna naloga pa bi bila določiti meje protipomenskosti (Gao in Zheng, 2014), ki se od enega do drugega znanstvenika močno razlikujejo.

Marjeta Humar (2016) med pionirske in najpomembnejše jezikoslovne raziskovalce protipomenskosti uvršča Lyonsa, Apresjana in Novikova. Lyons je določil tri vrste protipomenk, ki izhajajo iz ene od naštetih značilnosti: komplementarnost, protipomenskost in konverzija. Pri tem loči protipomenkost v ožjem in širšem smislu; v ožjega vključuje le polarno protipomenskost, ki je zanj najčistejša oblika antonimije. Apresjan je protipomenke razčlenil veliko temeljiteje, opozoril pa je tudi na kvaziprotipomenke, ki nimajo enako nasprotnih pomenov. Novikov je na drugi strani protipomenskost razdelil na kontrarno nasprotnost – kot najpogostejšo obliko, komplementarno in vektorsko nasprotnost. Med kvaziprotipomenke je uvrstil pomensko neenake, nesorazmerne, nesimetrične, stilistično raznorodne, časovno različne protipomenke, ki izražajo druga nasprotja.

V slovenskem prostoru se je najbolj uveljavila členitev po A. Vidovič Muha (2005, 2021), ki protipomenskost opredeljuje kot pomensko nasprotnost ali dopolnjevalno protislovnost; za izhodišče tipološke členitve jemlje vpliv protipomenk na aktantske vloge znotraj stavčne povedi. V okviru tega protipomenke deli na:

- zamenjavne oz. konverzivne,
- dopolnjevalne oz. komplementarne,
- skrajnostne oz. polarne, s podskupino stopnjevalnih oz. gradualnih in
- usmerjene oz. vektorske.

V grobem kategorizacija temelji torej bodisi na enakovrednih skupinah protipomenk bodisi na osi bolj protipomensko–manj protipomensko (ožji : širši smisel, prave protipomenke : kvaziprotipomenke, popolne : nepopolne, neostra : ostra nasprotnost, binarna : nebinarna nasprotnost, izražanje nasprotja : stilistično sredstvo) (Humar, 2016).

Strukturna delitev protipomenk je jasnejša. V slovenskem prostoru se je z njo z besedotvornega vidika največ ukvarjala Irena Stramljič Breznik (2010), ki protipomenke deli na istokorenske (tudi gramatične ali tvorbene) in raznokorenske (tudi leksikalne).

Slovenski, pa tudi sicer nekdanji jugoslovanski prostor protipomenskosti dolgo ni posvečal večje pozornosti (Humar, 2016), to izražajo tudi glavni slovenski jezikovni priročniki. SSKJ je s kvalifikatorjem ant. (antonim) opremil 87 leksemov, ki se uvrščajo med kakovostne (polarne, skrajnostne) protipomenke, medtem ko usmerjenih in dopolnjevalnih ne izkazuje (Humar, 2016). Toporišič (1976) v svoji slovnici antonimijo omenja bežno pri antonimnem pridevniku, protipomenskost krajše predstavi pozneje v četrti, prenovljeni izdaji leta 2001. Kljub temu da so protipomenke leksikografsko prepoznane kot pomemben dejavnik pri določanju pravilnih pomenov besed (Toporišič, 2001), protipomenskega slovarja v slovenskem prostoru še nimamo. Imamo pa dva slovarja sopomenk, in sicer Sinonimni slovar slovenskega jezika (SSSJ), ki ga je izdal ZRC SAZU, in spletni Slovar sopomenk sodobne slovenščine (SSSS), ki je nastal pod okriljem Centra za jezikovne vire in tehnologije. Pretekli leksikografski opis slovenskega jezika se je naslanjal na strukturalistično tradicijo SSKJ-ja, ki so ji sledili tudi dosedanji najvidnejši slovenski raziskovalci protipomenskosti (Jože Toporišič, Ada Vidovič Muha, Irena Stramljič Breznik, Marjeta Humar).

Družbene spremembe kot posledica digitalizacije in razvoja informacijsko-komunikacijske tehnologije so oblikovale potrebo po popolnoma drugačnem leksikografskem opisu slovenščine, na podlagi katerega bi lahko gradili nove jezikovne vire in tehnologije. Leksikografija se namreč v sedanjem času zaradi vstopa interneta spopada z vse hitrejšimi jezikovnimi spremembami. Na eni strani je soočena z vprašanjem, kako v spremenjenih razmerah predstaviti slovarske vsebine jezikovnim uporabnikom, na drugi strani pa z novimi jezikovnimi praksami, ki jih vse težje sproti zajema in popisuje (Gantar et al., 2016). Sodobni jezikovni uporabniki vse bolj zahtevajo takojšnji dostop do slovarskih vsebin sodobnega jezika, zato moramo leksikografske analize izvajati vse hitreje, a enako kvalitetno (Gantar et al., 2016). Iz tradicionalnega leksikografskega modela prehajamo v sodobnejši, pri katerem slovarske vsebine temeljijo na naprednih računalniških metodah, odprtosti, množičenju, relevantnosti in uporabnosti podatkov.

Tako je na eni strani povsem ročni pristop luščenja podatkov zamenjal polavtomatični, ki ni le časovno in finančno manj potraten, ampak hkrati zagotavlja dodatne potencialno koristne podatke za presojo o vključevanju leksemov v slovar. Pri tem se vloga leksikografa ne spreminja, saj še vedno ostaja odločevalec na vseh ravneh odločanja o slovarskem vključevanju leksemov, spreminja pa se način pridobivanja in predstavitve leksemskega podatka (Gantar et al., 2016). Podoben princip luščenja je bil uporabljen pri pripravi SSSS-ja. Leksemska razmerja navadno luščimo iz baze več virov, SSSS tako temelji na luščenju podatkov iz korpusa Gigafida in *Velikega angleško-slovenskega slovarja OXFORD - DZS* (Arhar Holdt et al., 2018). V tujini so pri pripravi korpusnih protipomenskih slovarjev prešli že na avtomatično luščenje (Wang et al., 2010; Lobanova et al., 2010; Aldhubayi in Alyahya, 2014).

Na drugi strani pa SSSS deluje tudi po konceptu odzivnega slovarja; gre za odprto dostopno zbirko relevantnih, a še ne povsem neprečiščenih podatkov. Pri izdelavi prečiščene baze sodeluje jezikovna skupnost, s čimer izdelava slovarja ni nikdar zaključena, saj se soustvarja skladno s spremenljivo jezikovno realnostjo. Poleg soustvarjanja jezikovni uporabniki potencialne iztočnice tudi vrednotijo s svojimi odzivi (Arhar Holdt et

al., 2018). Uporabniki prednost koncepta prepoznavajo v preglednosti, dostopnosti, hitremu prilagajanju sodobni sliki jezika, soustvarjalnosti, preprosti uporabi in načinu razvrščanja iztočnic (Kojc et al., 2018; Kamenšek Krajnc et al., 2018). Temu bi moral slediti tudi sodobni slovar protipomenk..

3. Metodologija

3.1. Pridobivanje podatkov

Podatkovno množico s protipomenkami smo sestavili iz več virov. Postopek je podrobneje opisan v diplomskem delu (Pegan, 2019), z izjemo zadnjega koraka z brisanjem ponavljajočih zapisov, ki je bil dodan naknadno. Glavnino podatkov o protipomenkah smo pridobili iz baze sloWNet (Fišer, 2015), manjši delček (87) pa na osnovi klicev iz slovarja SSKJ, dostopnega na slovarskem portalu Fran. Baza sloWNet ima obliko XML, poglejmo si en primer zapisa množice sopomenk (*synset*):

```
<SYNSET>
    <ID>eng-30-00001740-a</ID>
    ...
    <SYNONYM xml:lang="en">
        <LITERAL sense="1"
                pwnid="able%3:00:00::">able
        </LITERAL>
    </SYNONYM>
    <SYNONYM xml:lang="sl">
        <LITERAL lnote="auto">sposoben</LITERAL>
        <LITERAL lnote="auto">zmožen</LITERAL>
    </SYNONYM>
    ...
    <ILR type="near_antonym">
        eng-30-00002098-a</ILR>
    ...
</SYNSET>
```

Za vsak *synset* smo poiskali protipomenski *synset* prek elementa 'near_antonym'. Uporabili smo vse kombinacije, kjer je ena beseda v izvornem *synsetu* in druga v protipomenskem *synsetu*. Na tak način smo pridobili 4.514 parov protipomenk.

Iz SSKJ smo poiskali vsa gesla, ki imajo navedene tudi protipomenke. Poenostavljen primer zapisa vidimo spodaj:

```
<div>
    <span title="Iztočnica">abstrákten</span>
    ...
    <span title="Protipomenka">ant. </span>
    <span title="Protipomenka">
        <a>konkreten</a>:</span>
    ...
</div>
```

Skupno smo iz SSKJ izluščili 87 parov protipomenk.

Zaradi maloštevilčnosti smo podatke o protipomenkah razširili tako, da smo dodajali pare besed s pripono ne-, proti-, brez-. Primeri tako pridobljenih parov so *dostopen – nedostopen*, *ustaven – protiustaven* ter *alkoholen – brezalkoholen.* Tako pridobljene podatke smo deloma ročno prečistili nesmiselnih kombinacij, kot je *no – brezno* ter odstranili besede, za katere nismo imeli vektorskih vložitev v okviru diplomske naloge. Tako smo dobili 1340 parov protipomenk. Dodatno smo upoštevali tudi pare protipomenk, kjer eno izmed obeh besed zamenjamo z njeno sopomenko, s čimer se je množica povečala na 4113 parov protipomenk. Po brisanju ponavljajočih se zapisov, kjer sta besedi le zamenjani, smo pridobili množico 2852 parov protipomenk.

3.2. Označevanje podatkov

V raziskavo je bilo vključenih 2852 parov protipomenk. Vsak izmed šestih pregledovalcev je pregledal vse primere v individualni Google Preglednici, pri čemer je vsakemu paru pripisal eno izmed možnosti d, g in n. Oznaka d označuje, da gre za protipomenki, oznaka n pove, da dani besedi nista protipomenki, oznaka g pa pomeni, da je par problematičen in ga je treba podrobneje proučiti. Označevalci pred začetkom nismo prejeli natančnejših navodil, kaj se smatra kot protipomensko in kaj ne. Namen prvega koraka je bil namreč na osnovi gradiva ugotoviti problematična področja, ki bi jih lahko podrobneje analizirali v nadaljevanju.

Med pregledovanjem smo beležili primere in sproti oblikovali 19 problemskih kategorij. V nadaljevanju smo vsakemu izmed problematičnih parov pripisali po en glavni in morebitni dodatni problem. Podatke smo si razdelili na tri dele, pri čemer je vsak pregledovalec pregledal dva dela podatkov. Med pregledovanjem smo dodali še dve novi kategoriji, in sicer *(Ne)dovršne glagolske tvorjenke* in *Dejanje in stanje*, saj sta se kot problematični izkazali šele po natančnejši analizi vseh primerov.

4. Rezultati in analiza

Po prvem krogu pregledovanj smo 1124 (39,4 %) parov enotno potrdili kot protipomenske in le 22 (0,8 %) primerov kot neprotipomenske. Pri preostalih (1706; 59,8 %) se je vsaj eden izmed pregledovalcev odločil drugače kot ostali, zato smo takšne primere označili za nadaljnjo analizo. V drugem krogu pregledovanja pa se je izkazalo, da so bili nekateri primeri problematični zgolj v zelo specifičnem pogledu oz. da je bil primer lažno označen kot problematičen. Odločitev je bilo treba spremeniti tudi pri nekaterih že potrjenih parih, saj so se po podrobnejšem pregledu izkazali kot problematični. Kategorija potrjenih protipomenk se je tako povečala na 1207 (42,3 %) primerov, medtem ko je bilo potrjenih neprotipomenskih parov 48 (1,7 %). V nadaljnjo analizo smo poslali 1597 (56 %) primerov, kot prikazuje Tabela 1.

Oznaka	Delež
Sta protipomenki	42,3 %
Nista protipomenki	1,7 %
Nadaljnji pregled	56,0 %

Tabela 1: Rezultati po drugem krogu označevanja.

Nadaljnja raziskava se bo osredotočila zgolj na primere (1597; 56 %), ki so se po drugem krogu pregledovanj izkazali za problematične. Razdelili smo jih v 21 kategorij, prikazanih v Tabeli 2, kjer smo za lažjo predstavo vsako izmed kategorij ponazorili s primerom para besed, o katerih smo presojali. Vidimo lahko tudi, kolikokrat se je vsaka izmed kategorij pojavila kot glavni in kot dodatni problem. Glavne probleme smo določili 1597 primerom, medtem ko

smo dodatni problem identificirali pri 668 (41,83 %) primerih, ki predstavljajo 23,46 % celotnega gradiva.

Iz Tabele 2 je razvidno, da se kot glavni problem najpogosteje pojavlja *Redkost in kontekstualna vezanost* pomenov (31,87 %). Pogosto se pojavljajo tudi kategorije *Zanikanost s predpono -ne in -brez* (10,58 %), *Nedoslednost na ravni prevzeto – podomačeno* (10,33 %) in *Zaznamovanost in/ali redkost besede* (9,83 %). Najredkeje so se kot problematične pojavljale kategorije *Zatipki* (0,31 %), *Drugo* (0,38 %) in *Pomensko šibki glagoli* (0,44 %).

Kategorija	Primer	Št. pojavitev (glavni problem)	Odstotek	Št. pojavitev (dodatni problem)	Odstotek
Zatipki	*čistost – nečistot*	5	0,31	/	/
Napačne leme	*alkoholne – brezalkoholne*	40	2,50	3	0,45
Različna besedna vrsta	*dopoldne – popoldanski*	16	1,00	/	/
(Ne)dovršnost	*narasti – zniževati*	87	5,45	2	0,30
(Ne)določnost	*bližnji – daljen*	11	0,69	/	/
Neobstoječe besedotvorne različice	*pritrjevanje – zanikanost*	54	3,38	7	1,05
Zanikanost s predpono ne, brez-	*občutljivost – nedražljivost*	169	10,58	201	30,09
Nedoslednost na ravni prevzeto - podomačeno	*aktiv – trpnik*	165	10,33	36	5,39
(Ne)dovršne izglagolske tvorjenke	*zmanjšanje – povečanje*	32	2,00	4	0,60
Dejanje in stanje	*brezposelnost – zaposlitev*	18	1,13	2	0,30
Povratnost	*ubogati – upirati (se)*	53	3,32	17	2,54
Pomensko šibki glagoli	*manjkati – biti (prisoten)*	7	0,44	2	0,30
Pomensko polne besede	*pridobiti – odreči (soglasje)*	15	0,94	2	0,30
Spol kot “protipomenka”	*kralj – kraljica; dolžnica – upnik*	60	3,76	3	0,45
Zaznamovanost in/ali redkost besede	*ata – mati; nenavadno – često*	157	9,83	79	11,83
Enakopisnice in večpomenke	*bistrost – motnost*	76	4,76	20	2,99
Redkost in kontekstualna vezanost pomenov	*bogat – neploden*	509	31,87	246	36,83
Lastnosti, ki si niso protipomenske, a se pogosto tako uporabljajo	*krivulja – premica*	38	2,38	11	1,65
Posredne sopomenke	*glasen – nem*	40	2,50	5	0,75
Stopenjski primeri	*prihodnji – sedanji*	39	2,44	17	2,54
Drugo	*ofenziven – nespotakljiv*	6	0,38	11	1,65

Tabela 2: 21 kategorij in njihove pojavitve kot glavni in dodatni problem.

4.1. Zatipki

V kategorijo *Zatipki* spadajo pari, pri katerih je vsaj ena izmed besed nedvoumno zatipkana, torej ne more biti isti ali drug leksem v katerikoli obliki. Iz Tabele 2 je razvidno, da se je ta kategorija pojavila zgolj petkrat (0,31 %) kot glavni problem in nikoli kot dodatni. Je ena izmed najbolj problematičnih kategorij, saj besed, ki so narobe črkovane, ne moremo vključiti v slovar.

Primeri: *čistost – nečistot, izginti – pojaviti, izvažati – uvžati*..

4.2. Napačne leme

Pod *Napačne leme* sodijo primeri, ki so sicer lahko oblikoslovno ujemajoči, vendar v neslovarski obliki. Iz Tabele 2 je razvidno, da se je ta kategorija pojavila v 40 primerih (2,5 %) kot glavni in trikrat (0,45 %) kot dodatni problem. Takšne primere moramo odstraniti s seznama parov za vključitev v slovar oz. jih spremeniti v pravo slovarsko obliko.

Primeri: *alkoholne – brezalkoholne, dolžna – nedolžna, finančne – nefinančne.*

4.3. Različna besedna vrsta

Pri kategoriji *Različna besedna vrsta* gre za besedne pare, kjer sestavini pripadata različnima besednima vrstama (npr. samostalnik in pridevnik, pridevnik in prislov). Kot glavni problem se je ta kategorija pojavila pri 16 parih (1,00 %), kot sekundarni pa sploh ne. Pri večini primerov besedi nista protipomenki, dilema se pojavi le pri parih tipa samostalnik – pridevnik, saj gre tukaj največkrat za posamostaljene pridevnike (tipa *delavnik – fraj*). V takšnih primerih sta besedi lahko rabljeni protipomensko, seveda v ustreznem kontekstu. Pare iz te kategorije se odstrani s seznama za vključitev v slovar. Izjemo predstavljajo pari tipa samostalnik – pridevnik, ki se jih ročno pregleda in vključi s potrebnimi oznakami.

Primeri: *dopoldne – popoldanski, znotraj – ven, delavnik – fraj.*

4.4. (Ne)dovršnost

Pri *(Ne)dovršnosti* govorimo o glagolskih parih z različnim glagolskih vidom. Tako je eden izmed glagolov v nedovršni, drugi pa v dovršni obliki. Takšni pari so bili v 87 (5,45 %) primerih prepoznani kot primarni in dvakrat (0,30 %) kot sekundarni problem. Jasno je, da je za protipomenko nekemu glagolu najboljša izbira glagol, ki ima enak glagolski vid, a dilema ostaja pri glagolih, ki so pomensko ustrezni in imajo drugačen glagolski vid. Takšne pare bi bilo (vsaj na prvi pogled) smiselno odstraniti.

Primeri: *napasti – braniti, narasti – zniževati, natovoriti – iztovarjati.*

4.5. (Ne)določnost

V to kategorijo sodijo pridevniški pari, pri katerih je eden izmed pridevnikov v določni, drugi pa se pojavlja v nedoločni obliki. Ta kategorija se je v 11 (0,69 %) primerih pojavila kot glavni problem, medtem ko se kot dodatni problem ni pojavila. Ker je problem v veliki meri povezan z značilnostmi lematizacije za slovenščino, ki pridevnike lematizira v nedoločno obliko, razen kadar to ni mogoče (pari so pomensko načeloma protipomenski), bi bilo tovrstno gradivo smiselno ohraniti v slovarju.

Primeri: *bližnji – daljen, mesten – podeželski, oddaljen – bližnji.*

4.6. Neobstoječe besedotvorne različice

Gre za primere, ki so pomensko sicer ustrezni, a se težava pojavi, ker je ena (ali obe) beseda(i) neobstoječa. Kot primarni problem se je ta kategorija pojavila v 54 (3,28 %) primerih, kot sekundarni pa v sedmih (1,05 %). Ta kategorija po naši presoji ne sodi v slovar, saj gre za besede, ki niso realno v rabi. Že pri luščenju protipomenskih kandidatov bi lahko dodali korak preverbe posamezne besede v referenčnem korpusu in dodali opozorilo pri tistih primerih, ki se ne pojavljajo.

Primeri: *pritrjevanje – zanikanost, eleganca – neelegantnost, nelaskav – podrepniški.*

4.7. Zanikanost s predpono ne-, brez-

V kategoriji *Zanikanost s predpono -ne, -brez* govorimo o primerih, pri katerih je vsaj ena izmed protipomenk tvorjena kot negacija nekega izraza. Gre za pare, kjer sta obe besedi negaciji dveh protipomenk ali za primere, kjer se kot protipomenski par pojavita beseda in negacija njene sopomenke. Kot je razvidno iz Tabele 2, je bila ta kategorija v 169 (10,58 %) primerih prepoznana kot glavni in v 201 (30,09 %) kot dodatni problem. Raba takšnih parov v besedilu bi bila morda slogovno problematična, zagotovo pa so protipomenski v določenih kontekstih. Pare bi zato vključili v slovar in odločitev prepustili uporabniku, ki najbolje pozna kontekst, v katerem se beseda nahaja.

Primeri: *nespremenljiv – nestalen, neugoden – škodljiv, koristen – neugoden.*

4.8. Nedoslednost na ravni prevzeto – podomačeno

Tukaj obravnavamo primere, ki so sicer protipomenski, a je ena izmed besed prevzeta in s tem pogosto (drugače) zaznamovana. Zanimivo je tudi iskati mejo med »prevzetim« izrazom (*ujemanje – inkongruenca*) in takim, ki je v jeziku že uveljavljen (*inteligenten – neumen*). Razlike se lahko pojavljajo tudi na ravni zapisa prevzete besede in ne le v njenem pomenu (npr. *software* in *softver*). Iz Tabele 2 je razvidno, da so označevalci vsaj eno besedo prepoznali kot prevzeto v 165 (10,33 %) primerih, kjer je bil to glavni problem in v 26 (5,39 %) primerih, kjer je bil to dodatni problem. Ker gre tu le za prevzete besede, ki se v jeziku (še) niso uveljavile, bi jih bilo dobro vključiti v odzivni slovar, saj jih bo uporabnik lahko s pridom uporabljal v primernih kontekstih.

Primeri: *aktiv – trpnik, politeizem – enoboštvo, skupen – individualen.*

4.9. (Ne)dovršne glagolske tvorjenke

V kategorijo *(Ne)dovršne glagolske tvorjenke* sodijo tvorjenke, pri katerih besedotvorna podstava izkazuje razlike v dovršnosti. Ena beseda je torej tvorjena iz dovršnega, druga pa iz nedovršnega glagola. Analiza je pokazala, da je primerov, kjer je bila ta kategorija prepoznana kot primarni problem, 32 (2 %), in da so takšni, kjer je bila prepoznana kot sekundarni, štirje (0,60 %). Ti primeri so podobni 4.4, zato bi jih bilo smiselno obravnavati na enak način, torej jih ne bi vključili v slovar.

Primeri: *zmanjševanje – povečanje, izkrcanje – vkrcavanje, manjšanje – povečevanje.*

4.10. Dejanje in stanje

V to kategorijo sodijo samostalniški pari, ki so sicer protipomenski, a ena beseda predstavlja neko dejanje, dogodek, drugi pa neko stanje, lastnost. Problematika je podobna kot pri *(Ne)dovršnih glagolskih tvorjenkah*, le da gre tu za samostalnike, ki niso glagolsko tvorjeni. Kot je razvidno iz Tabele 2, se je kategorija *Dejanje in stanje* kot glavni problem pojavila v 18 (1,13 %) primerih in kot dodatni v 2 (0,30 %) primerih. Ker gre pri takšnih parih za manjšo nianso v pomenu, ki so v določenih kontekstih lahko protipomenski, jih je najbolje uvrstiti v slovar in uporabniku omogočiti, da sam presoja o njihovi uporabnosti.

Primeri: *zaposlitev – brezposelnost, degeneracija – razvoj, nedolžnost – zagrešitev.*

4.11. Povratnost

V kategorijo *Povratnost* smo uvrstili glagolske pare, ki so sicer protipomenski, a vsaj enemu izmed njiju (ali obema) manjka povratni zaimek. Brez povratnega zaimka takšni glagoli nimajo smisla ali imajo drugačen pomen (ki ni protipomenski s predlagano protipomenko). Iz Tabele 2 je razvidno, da je povratni zaimek kot glavni problem manjkal v 53 (3,32 %) parih, in pri 17 (2,54 %) kot dodatni problem. Ker je pri takšnih glagolih povratni zaimek ključen za smiselnost protipomenskega para, ga je nujno dodati. Takšne primere bi zato odstranili s seznama za vključitev v slovar.

Primeri: *strinjati (se) – prepirati (se), ubogati – upirati (se), udeležiti (se) – zamuditi.*

4.12. Pomensko šibki glagoli

Pri pomensko šibkih glagolih govorimo o glagolskih parih, v katerih (vsaj) en člen ob sebi zahteva dopolnilo, če ga želimo smatrati kot protipomenko drugemu. Kategorija se je kot glavni problem pojavila sedemkrat (0,44 %) in dvakrat (0,30 %) kot dodatni. Če naj bodo tovrstni primeri uvrščeni v slovar, mora biti ob pomensko šibkem glagolu dodana ustrezna beseda ali zveza.

Primer: *manjkati – biti (prisoten), biti (statičen/pri miru) – premikati (se), biti (statičen/pri miru) – gibati (se).*

4.13. Pomensko polne besede brez konteksta

Pod *Pomensko polne besede* spadajo pari, kjer je en člen lahko uporabljen kot protipomenka drugemu le takrat, ko je uporabljen v določenem kontekstu skupaj z neko drugo besedo. V ostalih kontekstih besedi nista v protipomenskem razmerju. Kot glavni problem se je omenjena kategorija pojavila pri 15 (0,94 %) parih in kot dodatni pri 2 (0,30 %) parih. Zdi se, da bi tovrstne probleme v slovar lahko vključili, manko konteksta pa rešili na ravni kolokacij, ki jih Slovar sopomenk sodobne slovenščine trenutno vključuje za pomensko primerjavo dveh sopomenk.

Primeri: *pridobiti – odreči (soglasje), odpovedati – obdržati (naročnino), napolniti – sprožiti (pištolo).*

4.14. Spol kot »protipomenka«

V kategoriji *Spol kot »protipomenka«* sta se pojavljali dve problematiki. Najprej smo obravnavali pare, kjer sta kot protipomenki navedena izraza, ki ju uporabljamo za označevanje spolov. Vprašanje je, ali je ob upoštevanju želene družbene občutljivosti slovarja spol sploh ustrezno definirati kot »protipomenski« ter ga s tem obravnavati kot nekaj nasprotnega in binarnega (npr. *moški – ženska*). Druga problematika je razvidna tudi iz primerov, pri katerih sta bila samostalnika (tipično) protipomenska, a v različnih slovničnih spolih (*dolžnica ¬ upnik*). Kot glavni problem se je spol pojavil pri 60 (3,76 %) parih in kot dodatni pri 3 (0,45 %) parih. Če bi kategorijo kljub problematičnosti uvrstili v slovar, bi bilo smiselno natančneje opazovati odzive uporabnikov in ugotoviti, kako ocenjujejo uporabnost in primernost tovrstnega gradiva. Pari tipa *dolžnica – upnik* niso ustrezni za v slovar oz. bi treba gradivo umestiti pod ustrezno iztočnico (*dolžnica – upnica*; *dolžnik – upnik*).

Primeri: *moški – ženska, kralj – kraljica, dolžnica – upnik.*

4.15. Zaznamovanost in/ali redkost besede

V kategoriji *Zaznamovanost in/ali redkost besede* najdemo pare, kjer načeloma gre za protipomenska izraza, a je en izmed njiju zaznamovan. V nekaterih primerih gre za čustveno zaznamovanost (*fant – punči*), v drugih za zastarelo rabo (*izjemoma – često*), pogovorne izraze (*delavnik – fraj*) ali zgolj za izraze, ki se v rabi le redko pojavljajo (*debelost – mršavost*). Kot je razvidno iz Tabele 2, se je ta kategorija kot glavni problem pojavljala precej pogosto, in sicer pri 157 (9,83 %) parih, prav tako pa tudi kot dodatni problem (pri 79 parih, tj. 11,83 %). Ker gre za primere, ki semantično ustrezajo pojmu protipomenskosti, bi jih bilo najbolje vključiti v slovar, da uporabnik sam preceni, če oz. kdaj so v njegovem kontekstu uporabni. Zagotovo bi jim pa bilo dobro dodati slovarsko oznako, ki bi označevala zaznamovanost, ki jo takšni izrazi imajo.

Primeri: *brat – sestrica, izredno – vobče, dolgovezen – koncizen.*

4.16. Enakopisnice in večpomenke

V kategorijo *Enakopisnice in večpomenke* so vključeni pari, kjer je eden izmed izrazov večpomenski. Pri teh parih gre velikokrat tudi za prenesen pomen enega izmed členov (*hladen – navdušen*). Problematične so tudi prave enakopisnice, torej tiste, ki bi v slovarju imele ločene iztočnice in ne le več pomenov (*pust – masten*). Takšni pari so se kot glavni problem pojavili 76-krat (4,76 %) in 20-krat (2,99 %) kot dodatni problem. Takšni primeri seveda sodijo v slovar, treba pa bi bilo opredeliti, s katerim pomenom besede je določena beseda v protipomenskem razmerju.

Primeri: *bistrost – motnost, zajedalec – gostitelj, moder – naiven.*

4.17. Redkost in kontekstualna vezanost primerov

V to kategorijo sodijo primeri, ki so protipomenski le v določenih kontekstih. Običajno je tu eden izmed izrazov bolj uveljavljen in uporabljen v več kontekstih, zato je protipomenka drugemu le v določenih primerih. Prav tako so se tukaj znašli primeri, pri katerih bi bili sestavini para v nad/podpomenskem razmerju, če bi eno od njiju negirali (kot pri *zdrav – umobolen*, kjer bi bili pravi protipomenki *zdrav – bolan*, medtem ko je *umobolen* le ena oblika nezdravja). V kategorijo *Redkost in kontekstualna vezanost primerov* smo vključili tudi primere, kjer je bil eden izmed izrazov zelo specifičen, običajno terminološki (primer: *izdelava – delaboracija*). Odločili smo se, da terminoloških izrazov ne bomo uvrščali v posebno kategorijo, saj je težko

določiti mejo med strokovnimi, specifičnimi in »čistimi« terminološkimi izrazi. Ker gre za najširšo kategorijo, se je kot primarni problem pojavila v kar 509 (31,87 %) primerih in kot sekundarni v 246 (36,83 %) primerih. Te primere se vključi v odzivni slovar, saj lahko uporabnik v široki paleti možnosti izbere zase najustreznejšo.

Primeri: *bogat – neploden, cena – prednost, domač – nepoznan.*

4.18. Lastnosti, ki niso protipomenske, a se pogosto tako uporabljajo

V tej kategoriji so zbrani primeri, ki sicer opisujejo izključujoče lastnosti, a v strogem pomenu ne gre za protipomenki, čeprav se pogosto tako uporabljata. To so predvsem pari, ki jih v pogovornem kontekstu uporabljamo kot protipomenki, ali takšni, za katere zmotno mislimo, da to sta. Kot je razvidno iz Tabele 2, je bila ta problematika prepoznana v 38 (2,38 %) primerih kot glavni in v 11 (1,65 %) primerih kot dodatni problem. Čeprav takšni pari niso strogo gledano protipomenski, bi jih bilo najverjetneje smiselno vključiti v slovar in izbiro prepustiti uporabniku.

Primeri: *anabolizem – katabolizem, krivulja – premica, nepomemben – znamenit.*

4.19. Posredne sopomenke

Pod *Posredne sopomenke* sodijo pari tipa *glasen – nem*, ki so na prvi pogled protipomenski le v redkih primerih, če pa bi eno sestavino zamenjali z njeno sopomenko, bi dobili precej bolj očiten protipomenski par (npr. *glasen – tih*). Takšni pari so se kot primarni problem pojavili 40-krat (2,50 %), kot sekundarni pa 5-krat (0,75 %). Čeprav niso prototipsko protipomenski, bi bilo tudi takšne pare morda dobro vključiti v slovar, saj uporabniku lahko koristijo v določenih situacijah, obenem pa spremljati, ali bodo uporabniki v odzivnem slovarju tovrstne primere ocenjevali s pozitivnimi ali negativnimi glasovi.

Primeri: *profit – minus, glasen – nem, kvaren – koristen.*

4.20. Stopenjski primeri

V to kategorijo smo zbrali pare, ki jih sicer lahko razumemo kot protipomenske v določenem kontekstu, a se pojavlja zelo očitna stopnjevanost. Besedi torej sta lahko protipomenki (*prihodnji – sedanji*), a običajno obstaja še neko bolj izrazito nasprotje (*prihodnji – pretekli*).Sem smo vključili tudi stopnjevane pridevnike, ki pa niso vedno nujno na popolnoma nasprotni stopnji. Tako imamo lahko v paru npr. primernik in presežnik in ne le dva primernika (primer: *manjši – največji* in ne le *manjši – večji*). Stopenjski primeri so se kot glavni problem pojavili v 39 (2,44 %) primerih in v 17 (2,54 %) primerih kot dodatni problem. Ker so kontekstualno pogojeni, jih je dobro vključiti v odzivni slovar in tako uporabniku omogočiti širšo izbiro potencialnih protipomenk.

Primeri: *negativen – nevtralen, dvojen – enojen, maksimalen – majhen*.

4.21. Drugo

Pod *Drugo* smo vključili primere, ki niso sodili v nobeno izmed ostalih kategorij. Kot je razvidno iz Tabele 2, smo 6 (0,38 %) parov vključili pod Drugo kot glavni problem in 11 (1,65 %) parov kot dodatni problem. Takšne pare, ki so se pojavili zelo poredko (0,38 %), bi bilo smiselno vključiti v slovar in presojo uporabnosti prepustiti uporabniški skupnosti.

Primeri: *državljan – tujec, ofenziven – nespotakljiv, zamuditi – zadeti.*

5. Zaključek

Iz analize je razvidno, da imajo problemske kategorije različno težo, nekatere težave bi bilo treba nasloviti, preden se gradivo lahko vključi v slovar, medtem ko lahko pri drugih odločitev o relevantnosti prepustimo uporabniški skupnosti. V analizi smo ugotovili, da so kategorije *Zatipki*, *Napačne leme*, *Različna besedna vrsta*, *(Ne)dovršnost*, *Neobstoječe besedotvorne različice*, *(Ne)dovršne glagolske tvorjenke* in *Povratnost* najbolj problematične, vendar jih je obenem predvidoma mogoče vsaj delno reševati tudi avtomatsko, kar bomo upoštevali pri razvoju nadaljnje metodologije strojnega pridobivanja protipomenk. Ostale kategorije pa so bolj vezane na kontekst, zato jih lahko vključimo v slovar in odločitev prepustimo skupnosti.

Čeprav je bilo nedvoumno potrjenih protipomenk na prvi pogled malo (manj kot polovica), pa nadaljnja analiza kaže, da lahko v odzivni slovar vključimo veliko večino (88 %) podatkov. Prav tu se kaže prednost odzivnega slovarja, ki uporabniku ponuja možnost, da izbira med širokim naborom potencialnih protipomenk in jih ocenjuje kot bolj ali manj ustrezne. V slovar je torej najbolje vključiti čim več potencialnega gradiva in jezikovni skupnosti prepustiti odločitev, kaj je zanjo uporabno in kaj ne.

Z digitalizacijo družbe so se spremenile (in povečale) potrebe jezikovnih uporabnikov, ki želijo vedno večji nabor podatkov, med katerimi lahko izbirajo. Odzivni slovar jim ne omogoči zgolj tega, ampak tudi dodajanje novega gradiva in odzivanje na že obstoječe. Skupaj z družbo se tako spreminjajo slovarji, z njimi pa tudi mi in naša vloga pri njihovem ustvarjanju.

6. Zahvala

Projekt *Nadgradnja temeljnih slovarskih virov in podatkovnih baz CJVT UL* v letih 2021–22 financira Ministrstvo za kulturo Republike Slovenije.

Avtorji in avtorice bi se radi zahvalili tudi Špeli Arhar Holdt za vključitev v projekt in pomoč pri načrtovanju raziskave in prispevka.

7. Literatura

Luluh Aldhubayi in Maha Alyahya. 2014. Automated Arabic Antonym Extraction Using a Corpus Analysis Tool. *Journal of Theoretical and Applied Information Technology,* 70(3):422–433.

Darja Fišer. 2015. Semantic lexicon of Slovene sloWNet 3.1. *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1026.

Polona Gantar, Iztok Kosem in Simon Krek. 2016. Discovering automated lexicography: the case of the slovene lexical database. *International Journal of Lexicography*, 29(2):200–225.

Špela Arhar Holdt, Jaka Čibelj, Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Bojan Klemenc, Iztok Kosem, Simon Krek, Cyprian Laskowski in Marko Robnik Šikonja. 2018. Thesaurus of Modern Slovene: By the Community for the Community. V: *Thesaurus of Modern Slovene: By the Community for the Community.*

Proceedings of the XVIII EURALEX International Congress, str. 401–410.

Marjeta Humar. 2005. *Protipomenskost v slovenski jezikoslovni literaturi*. V: M. Jesenček, ur., *Knjižno in narečno besedoslovje slovenskega jezika,* str. 234–238, Slavistično društvo Maribor, Maribor.

Marjeta Humar. 2016. *Protipomenskost v slovenskem knjižnem jeziku: na primeru terminoloških slovarjev*. Inštitut za slovenski jezik Frana Ramovša ZRC SAZU, Ljubljana.

Elin Kamenšek Kranjc, Špela Medved in Kaja Podgoršek. 2018. Primerjava spletnega slovarja Slovar sopomenk sodobne slovenščine in knjižnega Sinonimnega slovarja slovenskega jezika. *Liter jezika,* 9(12):66–70.

Agnes Kojc, Tamara Rigler, Kaja Sluga, Anika Plešivčnik in Špela Kovačič. 2018. Slovar sopomenk sodobne slovenščine in Sinonimni slovar slovenskega jezika. *Liter jezika,* 9(12):62–65.

Simon Krek, Cyprian Laskowski, Marko Robnik Šikonja, Iztok Kosem, Špela Arhar Holdt, Polona Gantar, Jaka Čibej, Vojko Gorjanc, Bojan Klemec in Kaja Dobrovoljc. 2018. Thesaurus of Modern Slovene 1.0. *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1166.

Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraž Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem in Kaja Dobrovoljc. 2020. Gigafida 2.0: the reference corpus of written standard Slovene. V: N. Calzolari, ur., *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*, str. 3340–3345. ELRA - European Language Resources Association, Paris. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Nikola Ljubešić in Tomaž Erjavec. 2018. Word embeddings CLARIN.SI-embed.sl 1.0. *Slovenian language resource repository CLARIN.SI*. http://hdl.handle.net/11356/1204.

Anna Lobanova, Tom van der Kleij in Jennifer Spenader. 2010. Defining Antonymy: A Corpus-based Study of Opposites by Lexico-syntactic Patterns. *International Journal of Lexicography,* 23(1):19–53.

Ada Vidovič Muha. 2005. *Medleksemski pomenski razmerji – sopomenskost in protipomenskost*. V: M. Jesenšek, ur., *Knjižno in narečno besedoslovje slovenskega jezika*, str. 206–221. Slavistično društvo Maribor, Maribor.

Ada Vidovič Muha. 2021. *Slovensko leksikalno pomenoslovje. Prva e-izdaja.* Znanstvena založba FFUL, Ljubljana.

Slovar slovenskega knjižnega jezika. Druga, dopolnjena in deloma prenovljena izdaja. 2014. Cankarjeva založba, Ljubljana.

Sopomenke 1.0. O slovarju. Center za jezikovne vire in tehnologije. https://viri.cjvt.si/sopomenke/slv/about.

Irena Breznik Stramljič. 2010. *Tvorjenke slovenskega jezika med slovarjem in besedilom.* Mednarodna založba Oddelka za slovanske jezike in književnosti FFUM, Maribor.

Jasmina Pegan. 2019. *Detekcija antonimov z vektorskimi vložitvami besed.* Diplomsko delo. Fakulteta za računalništvo in informatiko Univerze v Ljubljani.

Jože Toporišič. 1976. *Slovenska slovnica*. Založba »Obzorja«, Maribor.

Jože Toporišič. 2000. *Slovenska slovnica. Četrta, prenovljena izdaja.* Založba »Obzorja«, Maribor.

Wenbo Wang, Christopher Thomas in Amit Sheth. 2010. Pattern-Based Synonym and Antonym Extraction. *ACM SE '10: Proceedings of the 48th Annual Southeast Regional Conference*: 1–4. https://dl.acm.org/doi/abs/10.1145/1900008.1900094.

Ilukana – aplikacija za učenje japonskih zlogovnih pisav hiragana in katakana s pomočjo asociacij

Nina Sangawa Hmeljak,* Anna Sangawa Hmeljak,† Jan Hrastnik‡

* Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
nina.sangawa@gmail.com
† Akademija za likovno umetnost in oblikovanje, Univerza v Ljubljani
Dolenjska cesta 83, 1000 Ljubljana
anna.sangawa@gmail.com
‡ Fakulteta za matematiko in fiziko, Univerza v Ljubljani
Jamova cesta 21, 1000 Ljubljana

Povzetek

Prispevek predstavlja zasnovo in oblikovanje digitalne aplikacije za slovensko govoreče učence oz. študente japonščine kot pomoč pri pomnjenju japonskih zlogovnih pisav hiragana in katakana s pomočjo asociacij in interaktivnega učenja. Vsak znak dveh japonskih pisav je opremljen z ilustracijo, ki vsebuje obliko tega znaka in obenem ponazarja slovensko besedo, ki se začne s tem zlogom. Aplikacija nudi tako seznam ilustracij kot tudi interaktivne vaje, s katerimi uporabnik lahko preverja svoje znanje. Aplikacija je napisana s paketom za razvoj programske opreme Flutter, v jeziku Dart, tako da deluje v poljubnem operacijskem sistemu. Je še v fazi prototipa, v bodoče načrtujemo raziskavo o učinkovitosti pri pomnjenju, testiranje med uporabniki in dodelavo uporabniškega vmesnika.

Ilukana – an app for learning the Japanese hiragana and katakana syllabaries using associations

We present the concept and implementation of a digital application for Slovene-speaking learners of Japanese, as an aid to remembering the Japanese syllabaries hiragana and katakana using associations and interactive learning. Each letter of the Japanese syllabaries is matched with an illustration containing the letter itself and representing a Slovene word beginning with the syllable represented by the letter. The application includes a list of illustrations and interactive exercises. It is written using Flutter, in Dart, and can therefore be used in any operating system. The app is a prototype, research on its effectiveness, user testing and interface upgrades are planned.

1. Uvod

V prispevku predstavljamo izgradnjo in oblikovanje digitalne aplikacije za učenje japonskih zlogovnih pisav hiragana in katakana s pomočjo asociacij in interaktivnega učenja. Aplikacija je namenjena slovensko govorečim učencem ali študentom japonščine kot pomoč pri učenju osnovnih znakov japonskih zlogovnih pisav hiragana in katakana. Osnovana je na principu asociacije med znanimi in novimi informacijami: za lažje pomnjenje oblike in izgovora znakov japonskih zlogovnic ponuja za vsak znak ilustracijo, ki nakazuje obliko tega znaka in obenem ponazarja slovensko besedo, ki se začne s tem zlogom. V aplikaciji so seznam ilustracij in interaktivne igrice za preverjanje naučenih znakov in tudi mini-igrica za učenje pravilnega vrstnega reda potez pri pisanju kane. Aplikacija je še v fazi prototipa, v prispevku predstavljamo ozadje projekta, namen aplikacije, teoretična izhodišča, podobne ilustracije in aplikacije za govorce drugih jezikov, oblikovalski koncept, tehnično implementacijo, ugotovljene pomanjkljivosti in načrte za bodoče delo.

2. Ozadje projekta

Japonščina je priljubljen jezik med ljubitelji japonskih mang in animejev, v Sloveniji se poučuje na Filozofski fakulteti Univerze v Ljubljani, v več privatnih jezikovnih šolah, mnogi mlajši se japonščine učijo tudi sami s pomočjo spleta. Pri učenju japonščine je posebej zahtevno učenje pisave, saj japonščina ne uporablja latinice ampak tri druge pisave, hiragano in katakano, ki sta japonski zlogovni pisavi, ter kanji oz. pismenke, ki izvirajo iz Kitajske (Hmeljak et al., 2020). Od treh pisav imata hiragana in katakana še najmanj znakov, vsaka ima 46 različnih znakov, medtem ko je kitajskih pismenk na tisoče. Tako kot vsak japonski otrok se tudi tuji učenci najprej naučijo teh dveh zlogovnic. Ker je učenje nove pisave, ki ima popolnoma drugačne oblike kot latinica, težavno, a tudi nujno potrebno, da lahko učenci sploh začnejo brati v japonščini, smo se odločili ustvariti aplikacijo, s katero je lahko učenje lažje in bolj zabavno.

3. Učenje z asociacijami – mnemotehnika

Mnemotehnika oz. mnemonika je tehnika učenja oz. pomnjenja, pri kateri skušamo vsebino, ki se jo želimo naučiti (tj. to, kar imamo samo v kratkoročnem spominu), urediti in povezati z že znanim (tj. s tem, kar že imamo v dolgoročnem spominu) na tak način, da si jo lažje zapomnimo. Pomembne so zlasti v začetni fazi učenja jezika, ko si mora učenec zapomniti osnovno besedišče ali pisavo, medtem ko na višjem nivoju učenja jezika imajo učenci običajno bolj razvito in povezano znanje in lahko učinkovito uporabljajo druge metode (Oxford, 2016). Primeri mnemotehnike so razne rime in besedne zveze, s katerimi si lažje zapomnimo določena pravila, kot npr. stavek “Suhi škafec hoče pasti”, s katerim si zapomnimo soglasnike, pred katerimi uporabimo predlog s in ne z, ali povezovanje oblike predmeta z obliko črke v besedi, ki si jo hočemo zapomniti v povezavi s predmetom, npr. “ob prvem krajcu se luna Debeli, ob zadnjem pa Crkuje”, kjer oblika črk D in C spominja na obliko lune ob prvem in ob zadnjem krajcu.

Med mnemotehnike spada tudi metoda ključne besede, po kateri asociiramo novo besedo, ki se jo želimo naučiti, z besedo, ki podobno zveni, s pomočjo neke vsebinske povezave (Cohen, 1987; Manalo, 2002). Več raziskav kaže, da je mnemotehnika lahko uporabna za učenje širokega

spektra snovi, kot je učenje tujih jezikov, znanstvenih zakonitosti, itd. Omenjene raziskave so pokazale, da so se tisti, ki so se učili z uporabo mnemotehnike, veliko bolje odrezali kot tisti, ki se niso. Poleg tega je bila mnemotehnika učinkovita tudi pri učenju oseb s specifičnimi učnimi težavami ali po možganskih poškodbah. Nekatere raziskave so pokazale celo, da večina ljudi spontano uporablja mnemotehniko pri učenju na pamet (Manalo et al., 2004).

Mnemotehnika je torej uporabna tudi pri učenju novih jezikov in pisav. Več raziskav je pokazalo tudi učinkovitost mnemotehnik, ki so angleškim govorcem pomagale pri učenju japonske pisave (Quackenbush et al., 1989; Manalo et al., 2004; Matsunaga, 2003) in korejske pisave (Brown, 2012).

Obstaja tudi več učbenikov za učenje kitajskih pismenk, ki se poslužujejo mnemotehničnih metod za pomnjenje in povezovanje oblike in pomena s pomočjo asociacij. Med prvimi je serija učbenikov Jamesa Heisiga (1977; 1987; Heisig in Sienko, 1994), ki pokriva vseh 2000 standardnih pismenk in je bila prevedena v francoščino (1998), španščino (2001) in nemščino (2005), za angleško govoreče pa obstaja še več podobnih učbenikov (Banno et al. 2009; Bodnaryk, 2000; McNair, 2000; McNair, 2005 in McCabe, 2012). Za slovensko govoreče učence japonščine pa še ni takega gradiva, zato smo se odločili, da ga ustvarimo.

3.1. Mnemonične slike

Za učenje hiragane z asociacijami obstaja že več primerov za učenje s pomočjo mnemoničnih slik, ki so povezane z angleščino. Oblika enega znaka zlogovnice hiragane se prekrije z ilustracijo angleške besede, ki se začne z enakim zlogom kot izbrani znak hiragane (Ogawa, 1990; Rowley, 1995; Koichi, 2014). Obstajajo tudi že aplikacije za ta namen, kot je npr. Hiragana Memory Hint in Katakana Memory Hint Japonske fundacije (Japan Foundation 2015), ki ponuja učenje v povezavi z angleščino, indonezijščino in tajščino.



Slika1. Primeri idej mnemoničnih slik za znake に ni, さ sa in な na (zgoraj Ogawa 1990, spodaj Koichi 2014).

Manalo et al. (2004) so ugotovili, da je tak način učenja hiragane učinkovit, udeleženci so bili na splošno zadovoljni, bili so mnenja, da jim je pomagalo pri pomnjenju in šolskem uspehu. Po drugi strani Matsunaga (2003) ugotavlja, da je učenje hiragane z mnemoničnimi slikami, povezanimi z angleškimi besedami, bilo učinkovito pri učencih japonščine, ki niso bili naravni govorci angleščine, le na kratek rok in to le za tiste, ki se nikoli prej niso učili jezika, ki ne uporablja latinice.

O rabi mnemoničnih slik za učenje kane med govorci slovenščine še ni bilo raziskav, a lahko domnevamo, da tudi zanje učenje preko ilustracij, ki se nanašajo na angleške besede, ni posebej učinkovito, kot ugotavlja Matsunaga (2003) za govorce drugih jezikov.

4. Aplikacija Ilukana

Učbeniki in aplikacije za učenje hiragane z mnemoničnimi slikami torej že obstajajo, vendar ne za slovensko govoreče oz. ni takih, ki bi povezale oblike znakov kane s slovenskimi besedami. Glede na to, da praktično vsi slovensko govoreči učenci japonščine že obvladajo tudi angleščino, bi lahko na prvi pogled uporabljali gradivo za angleško govoreče, kot ga ponuja npr. Ogawa (1990) ali Koichi (2014) in je prikazano v sliki 1. Toda zlasti pri angleščini, ki ima izrazito globok pravopis, lahko pri povezovanju oblike kane z izgovarjavo angleške besede pride do zmede zaradi interference zapisa angleške besede in tudi zaradi variabilnosti izgovora same angleščine (britanska ali ameriška angleščina ipd.). Za slovenske govorce, ki se angleščine večinoma učijo istočasno v govorni in pisni obliki, bi lahko bilo težko odmisliti pisno obliko (npr. “nun” za znak な *na*) in povezati samo izgovarjavo besede v angleščini (/nan/) z zlogom /na/ v japonščini, saj se sorodna beseda v slovenščini izgovori /nuna/ (glej sliko 1). Podobno bi lahko tudi rekli za znak に *ni*, za katerega je izbrana beseda knee, ki se izgovarja /nii/, vendar za tiste, ki si hkrati predstavljajo tudi pisno obliko, je težko odmisliti »k«, ki je na začetku besede. Tudi fonetično je marsikateri zlog v slovenščini bliže japonskemu kot angleški, tako je npr. zlog /sa/ praktično enak v slovenščini in japonščini, medtem ko je izgovorjava angleške besede “saw” drugačna, še dodatno pa lahko zmede razlika med ameriškim in britanskim izgovorom.

Za mlajše učence, ki morda še ne obvladajo angleščine, pa je verjetno lažje pomniti asociacije z besedami iz lastnega maternega jezika kot iz angleščine ali drugega tujega jezika, ki ga še ne obvladajo dobro. Zato smo se odločili poiskati slovenske besede, ki lahko pomagajo pri pomnjenju znakov japonske kane, zanje ustvarili ilustracije in s temi zgradili aplikacijo Ilukana.

Ilukana je aplikacija za učenje japonskih znakov hiragana in katakana. Ciljna publika so slovensko govoreči učenci ali študenti. Aplikacija ponuja ilustracije, s pomočjo katerih si uporabnik lažje zapomni povezavo med obliko znaka in njegovo izgovarjavo. Uporabnik dostopa do posameznih znakov preko seznama obeh pisav, ki se izmenjujeta z interakcijo uporabnika. Aplikacija vključuje tudi element igre v obliki kviza, prav tako za pomnjenje izgovorjave, kot tudi pravilnega vrstnega reda potez pri zapisovanju. Pri japonščini namreč pravopis določa vrstni red potez, ki vpliva na obliko znakov, zlasti bolj kompleksnih. Pri znaku あ se na primer najprej zapiše vodoravna črta, nato navpična in na koncu še krivulja.

4.1. Ustvarjanje asociacijskih slik

V aplikaciji so znaki kombinirani z ilustracijami, ki uporabniku pomagajo, da si zapomni obliko znaka z asociacijo na vsebino ilustracije. Tako na primer ilustracija za znak あ (glej sliko 2), ki se izgovori /a/, prikazuje adrenalinski park, kar uporabniku pomaga, da si preko besede "adrenalin", ki se začne z zlogom /a/, zapomni povezavo med obliko znaka あ in njegovo izgovarjavo, tj. glasom /a/.

Za ustvarjanje asociacij je bilo torej potrebno za vsak znak hiragane in katakane najti slovensko besedo, ki se začne z enakim zvokom kot izbrana hiragana in ki predstavlja nekaj, kar je podobne oblike. Pri tem je še nekaj omejitev: beseda mora pomeniti nekaj, kar je mogoče izrisati (abstraktne pojme bi težje spremenili v ilustracije), obenem pa mora biti ilustracija kolikor mogoče enoumno povezana z eno samo besedo, poleg tega ne sme imeti preveč detajlov, da se lahko jasno izriše tudi na manjšem ekranu. Za vsak znak smo preizkusili več idej in se odločili za najbolj jasno.

Primeri nekaj različnih idej za isti znak so prikazani v sliki 3.



Slika2. Primeri idej mnemoničnih slik: a kot adrenalin.



Slika3. Primeri idej mnemoničnih slik: け/ke/ kot *keramika, keglanje, kebab, kečap, Kekec*.



Slika4. Ni kot nilski konj.



Slika5. Sa kot sardele.

Na sliki 2 je prikazan primer za zlog /a/ (あ v hiragani in ア v katakani). Tu smo uspeli najti primer ilustracije (adrenalin) za isto besedo, ki se prekriva z znakoma za isti zlog v obeh zlogovnicah. Za znak hiragane あ smo kot najprimernejšo izbrali to besedo, ker komplicirani ovinki spominjajo na vlakec smrti. Pri katakani ア pa oblika lahko spominja na kajak na divjih vodah. Obe aktivnosti sta zelo dinamični in ju lahko povežemo z besedo *adrenalin*.

Na sliki 3 je prikazanih več različnih idej, ki smo jih imeli za znak hiragane け, ki se izgovori /ke/. Po vrsti od leve zgoraj so *keramika*, *kegljanje*, *kebab*, *kečap* in *Kekec*. Da bi si lažje zapomnili, da je znak sestavljen iz dveh ločenih delov, sta bila kegelj ob kegljaču in Kekec s pohodno palico najboljša kandidata. Toda kegljanje bi lahko zamenjali z bowlingom, ki je bolj razširjen, in bi tako lahko zamešali z zlogom /bo/, ki se v hiragani zapiše ぼ in je oblikovno podoben znaku け oz. /ke/, zato smo izbrali ilustracijo Kekca, ki ni dvoumen.

Sliki 4 in 5 prikazujeta še primer za /ni/ に in /sa/ さ. Za /ni/ smo izbrali izraz nilski konj, za /sa/ pa sardelo.

4.2. Oblikovanje aplikacije

Pri oblikovanju aplikacije smo se odločili za minimalistični izgled. Za celotno podobo so za izhodišče

uporabljene barve japonske zastave, tj. rdeča in bela, ter črna za besedilo. Zato da ekran ni presvetel in ne draži oči, je za ozadje uporabljena siva barva. Vse ilustracije, ki so funkcionalni del aplikacije (ikona za premikanje naprej in nazaj, vračanje na vstopno stran ipd.), so v slogu tradicionalnih japonskih grafik *ukiyo-e*.

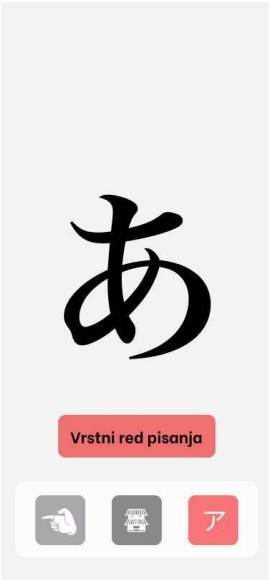
Ko prižgemo aplikacijo, se znajdemo pred vhodom japonske hiše. Ko se dotaknemo vrat, ki se nam odprejo, vstopimo na prvo stran, ki je glavni meni. Na glavnem meniju imamo štiri gumbe. V ozadju je ilustracija, ki je povzeta po znani grafiki japonskega slikarja Sharakuja, ki upodablja igralca gledališča *kabuki*. V navigacijski vrstici so trije gumbi. Na sredini je gumb za vračanje na vstopno stran (*home button*) v obliki japonske hiše, na levi je gumb za premikanje na prejšnji ekran (*back button*) v obliki roke v slogu japonskih grafik *ukiyo-e*, na desni pa gumb, ki nam omogoča preklapljanje med znaki hiragane in katakane. Prva dva gumba na vstopni strani sta namenjena učenju pismenk s pomočjo asociacije z ilustracijo. Če se dotaknemo gumba hiragana ali katakana, nas to pripelje na seznam vseh znakov (pismenk) te pisave. Ko se dotaknemo enega znaka hiragane ali katakane, nas to pripelje na stran s to pismenko čez celo širino ekrana. Pod pismenko je gumb, ki nas pripelje do gibljive slike, ki pokaže pravilni vrstni red, po katerem se zapiše. Ko se dotaknemo pismenke same, pa se nam prikaže pismenka v kombinaciji z ilustracijo. Pod pismenko je v latinici napisana izgovarjava (zlog, ki ga pismenka zapisuje) ter slovenska beseda za pojem, s katerim ga asociiramo. Tretji gumb na glavnem meniju z imenom “seznam” je le pregledni seznam, ki ga lahko z gumbom za preklapljanje uporabimo za pregledovanje in pomnjenje oblik pismenk. Četrti gumb z imenom “vaje” pripelje uporabnika do dveh vaj, kjer se lahko nauči vrstni red pisanja in izgovorjavo hiragane ali katakane.



Slika6. Začetna stran aplikacije.



Slika7. Glavni meni.



Slika8. Stran za črko あ.



Slika9. Mnemonična slika za あ.

4.3. Primerjava aplikacije Hiragana Memory Hint z aplikacijo Ilukana

Aplikacija Hiragana Memory Hint je aplikacija, ki jo Japonska fundacija ponuja na App Store in Google Play (Japan Foundation 2015). Namenjena je učenju zlogovne pisave hiragana oz. olajšanje tega za angleško govorečo populacijo. Obstajajo tudi različice za govorce drugih azijskih jezikov. Naša aplikacija Ilukana ji je podobna v tem, da obe uporabljata mnemonične slike za povezovanje besede v maternem jeziku govorca z obliko kane. Glavna razlika pa je seveda to, da je naša namenjena govorcem slovenščine. Obe aplikaciji imata dve glavni funkciji: pregledovanje in učenje japonskih znakov s pomočjo mnemonskih slik ter kviz, kjer lahko uporabnik vadi. Ilukana nudi uporabnbiku izbiro, ali si z mnemoničnimi slikami želi zapomniti hiragano ali katakano, medtem ko Hiragana Memory Hint ima na voljo le hiragano, saj za katakano obstaja ločena aplikacija. Naša aplikacija ima tudi daljši seznam vseh zlogov, saj ne vsebuje le osnovnih 46 znakov hiragane kot aplikacija v angleščini, ampak vse možne zloge, ki jih lahko napišemo z dodajanjem diakritičnih znakov: črtici ゛ za zvenečnost (npr. か /ka/ oz. が /ga/), krogec ゜ za glas /p/ (npr. は/ha/ oz. ぱ /pa/) in diakritični znaki za mehčanje soglasnikov (npr. さ /sa/ oz. しゃ /ša/). Za te posebne zloge še nimamo ilustracij v našem prototipu.

Aplikacija Hiragana Memory Hint ima več interaktivnosti in elementov igre. Ilukana ima le dve igrici. Ena je za vajo izgovorjave, ki je v obliki kviza z več možnimi odgovori, pri drugi pa uporabnik pritisne na črte znaka po pravilnem vrstnem redu pisanja (*kakijun*). Hiragana Memory Hint pa ima 4 različne tipe kvizov. Poleg branja hiragane ima še kviz z več odgovori, kjer uporabnik izbere med več znaki hiragane za dano izgovorjavo, ki je napisana v latinici, izbira znak hiragane glede na izgovorjavo posneto v zvočni obliki in kviz za izbiro hiragane glede na napisano izgovorjavo, kjer so na izbiro znaki, ki so si podobni.

Aplikacija Hiragana Memory Hint uporablja črno bele linearne ilustracije z barvnim ozadjem, medtem ko Ilukana uporablja barvne ilustracije s svetlo sivim ozadjem in črnim besedilom. Pri aplikaciji Hiragana Memory Hint so izbrali minimalističen, jasen učbeniški design s sans serifno latinico za angleščino in *gothic* črkovno vrsto v japonščini , medtem ko smo se pri Ilukana odločili za *mincho* črkovno vrsto pri japonskih pismenkah, zato da si lahko uporabnik natančno zapomni pisano obliko japonskih znakov, vključno z zaključevanjem potez po principih *tome, hane* itd., črke v latinici pa so tudi v Ilukani sans serifne zaradi boljše čitljivosti. Pri aplikaciji Hiragana Memory Hint so za oblikovanje gumbov in znakov izbrali minimalistični pristop, ki nas spominja na učbenik ali delovni zvezek s poudarkom na okroglo obliko in igrive barve, pri Ilukana pa smo želeli uporabiti elemente japonske kulture v slogu tradicionalnih japonskih grafik na gumbih in ozadjih. Razlika je tudi v uporabi barve: Hiragana Memory Hint uporablja več odtenkov barve v gumbih, kot so zelena, modra, rdeča, oranžna itd., ki daje igriv občutek, medtem ko je pri Ilukana uporabljena rožnata rdeča kot glavni odtenek s kombinacijo sivih odtenkov ter črne, ki daje bolj izčiščen, eleganten občutek.

4.4. Tehnična implementacija aplikacije

Aplikacija je napisana s paketom za razvoj programske opreme Flutter, v jeziku Dart. Ta jezik smo izbrali zato, da bo aplikacija uporabna v vsakem okolju, saj nam Flutter omogoča, da aplikacija deluje tako v operacijskih sistemih iOS in Android kot v poljubnem spletnem brskalniku.

Pri pisanju programa je bil največji izziv oblikovati vse objekte, da se oblika ohrani pri vseh možnih velikostih ekranov. To smo reševali s testiranjem na različnih telefonih in popravljanjem relativne razdalje med objekti. Ker je bil dizajn originalen, smo morali posebej implementirati vse objekte, kot je navigacija.

Spletna verzija aplikacije je dostopna na naslovu https://sninah.github.io/ilukana/.

5. Zaključek

Aplikacija je še v fazi razvoja, testirana je bila na nekaj različnih modelih telefonov, a za optimalno delovanje na manjših zaslonih potrebuje še nekaj dodelave.

V bodoče nameravamo testirati in optimizirati delovanje aplikacije, med drugim tudi z optimizacijo tempiranja prikazovanja znakov, ki se zdaj pri kvizu prikazujejo naključno, tako da se večkrat pojavljajo tisti, pri katerih je uporabnik naredil več napak.

Nameravamo tudi preveriti uporabnost in učinkovitost ilustracij pri učenju hiragane in katakane med slovenskimi govorci. Temeljite uporabniške študije aplikacije še nismo izvedli, same ilustracije pa smo pokazali nekaj študentom japonščine, ki so povedali, da so jim bile ilutracije zabavne in da so pomagale pri pomnjenju. Da bi preverili dejansko učinkovitost pri pomnjenju, bi bilo potrebno izvesti eksperiment s kontrolno skupino in preverjanjem znanja pred učenjem, takoj po učenju in čez daljši čas, kar načrtujemo izvesti v prihodnje.

6. Literatura

Eri Banno, Yôko Ikeda, Chikako Shinagawa, Kaori Tajima in Kyôko Tokashiki. 2009. *Kanji look and learn: 512 kanji with illustrations and mnemonic hints*. Tokyo: Japan Times.

Robert P. Bodnaryk. 2000. *Kanji Mnemonics: An Instruction Manual for Learning Japanese Characters*. Winnipeg, Manitoba: Kanji Mnemonics.

Lucien Brown. 2012. The use of visual/verbal and physical mnemonics in the teaching of Korean Hangul in an authentic L2 classroom context. *Writing Systems Research,* 4(1):72–90. http://dx.doi.org/10.1080/17586801.2011.635949

Andrew Cohen. 1987. The use of verbal and imagery mnemonics in second-language vocabulary learning. *Studies in Second Language Acquisition,* 9(1):43–61.

Kazumi Hatasa. 1991. Teaching Japanese syllabary with visual and verbal mnemonics. *CALICO Journal,* 8(3): 69–80. http://www.jstor.org/stable/24156286.

James Heisig. 1986. *Remembering the kanji: A complete course on how not to forget the meaning and writing of Japanese characters*. Tokyo: Japan Publications Trading Co.

James Heisig. 1987. *Remembering the kanji: A systematic guide to reading Japanese characters*. Tokyo: Japan Publications Trading Co.

James Heisig in Tanya Sienko. 1994. *Writing and reading Japanese characters for upper-level proficiency*. Tokyo: Japan Publications Trading Co.

James Heisig, Marc Bernabé in Verònica Calafell. 2001. *Kanji para recordar: curso mnemotécnico para el aprendizaje de la escritura y el significado de los caracteres japoneses*. Barcelona: Herder Editorial.

James Heisig in Yves Maniette. 1998. *Les kanji dans la tête: apprendre à ne pas oublier le sens et l'écriture des caractères japonais*. Yves Maniette.

James Heisig in Robert Rauther. 2005. *Bedeutung und Schreibweise der japanischen Schriftzeichen*. Frankfurt am Main: V. Klostermann.

Kenneth Higbee. 1977. *Your Memory: How It Works and How to Improve It*. Englewood Cliffs, NJ: Prentice-Hall.

Kristina Hmeljak Sangawa, Hyeonsook Ryu in Mateja Petrovčič. 2020. Zakaj latinica ni dovolj: o izgubi informacij pri latinizaciji vzhodnoazijskih imen v knjižničnih katalogih. *Knjižnica,* 64(1–2):47–78.

Japan Foundation. 2015. *Hiragana Memory Hint. Katakana Memory Hint. English Version.* https://minato-jf.jp/Home/JapaneseApplication.

Koichi. 2014. *Learn hiragana: The ultimate guide*. https://www.tofugu.com/japanese/learn-hiragana/

Emmanuel Manalo. 2002. Uses of mnemonics in educational settings: A brief review of selected research. *Psychologia,* 45(2):69–79. https://doi.org/10.2117/psysoc.2002.69

Emmanuel Manalo, Satomi Mizutani in Julie Trafford. 2004. Using mnemonics to facilitate learning of Japanese script characters. *Japan Association for Language Teaching Journal,* 26(1):55–77. http://jalt-publications.org/recentpdf/jj/2004a_JJ.pdf#page=57.

Sachiko Matsunaga 松永幸子. 2003. Effects of Mnemonics on Immediate and Delayed Recalls of Hiragana by Learners of Japanese as a Foreign Language. *Japanese-Language Education around the Globe,* 13: 19–40. https://doi.org/10.20649/00000331.

Glen McCabe. 2012. *Learning Japanese Hiragana & Katakana Flash Cards Kit*. Tokyo: Charles E. Tuttle.

Bruce McNair. 2005. *Kanji Learned Through Phonic-Mnemonics: Learning to Read Japanese Kanji Using the McNair Phonic-Mnemonic System*. Kanji Learning Institute.

Bruce McNair. 2016. *Read Kanji Read: Read the 2,136 Jooyoo Kanji in Two Months Using Phonic Mnemonics (English Edition)*. Kanji Learning Institute.

Kunihiko Ogawa. 1990. *Kana Can Be Easy*. Tokyo: The Japan Times.

Rebecca L. Oxford. 2016. *Teaching and researching language learning strategies: Self-regulation in context*. London: Routledge.

Hiroko Quackenbush, Kiyomi Chujo, Kazuhiko Nagamoto in Shinichiro Tawata. 1989. 50分ひらがな導入法：連想法と色付きカード法の比較 Teaching how to read hiragana in 50 minutes: A comparison of mnemonics and the use of cards with associated colours. 日本語教育 *Journal of Japanese Language Teaching,* 69: 147–162.

Michael Rowley. 1995. *Kana Pict-O-Graphix: Mnemonics for Japanese Hiragana and Katakana*. Albany, CA: Stone Bridge Press.

Filter nezaželene elektronske pošte za akademski svet

Anja Vrečer

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
anja.vrecer@gmail.com

Povzetek

Nezaželena akademska elektronska sporočila so nezaželena sporočila, ki jih prejemajo predvsem profesorji, raziskovalci in drugi akademiki, in jih navadni filtri nezaželene elektronske pošte ne zaznavajo. V prispevku predstavimo izdelavo filtra nezaželene akademske elektronske pošte, pri čemer smo naredili primerjavo različnih metod filtriranja sporočil in različnih tehnik obdelave besedila. Za končni model smo uporabili nevronsko mrežo v kombinaciji z vektorskimi vložitvami besed ter ga povezali z izbranim odjemalcem elektronske pošte, in sicer z Gmailom. Filter smo testirali z 10-kratnim prečnim preverjanjem in dosegli tudi do 98% točnost.

1. Uvod

Elektronska pošta je v zadnjem času postala ena najbolj uporabljenih aplikacij za komunikacijo. Vsakodnevno jo uporablja na milijone ljudi, tako v službi kot v prostem času (Whittaker et al., 2005). Slabost vsesplošne uporabnosti elektronske pošte pa je vse večja količina elektronskih sporočil, ki jih prejemamo. Med njimi je tudi veliko nezaželenih elektronskih sporočil. Prebiranje vseh sporočil nam zato včasih vzame ogromno časa in energije. Ker želimo čim hitreje ločiti nezaželena sporočila od drugih, uporabnih sporočil, imajo mnogi poštni odjemalci že vgrajene filtre nezaželene elektronske pošte. Vendar pa takšni filtri ne zaznajo vseh vrst nezaželene elektronske pošte. V prispevku se osredotočimo na eno takšnih skupin nezaželene elektronske pošte, in sicer na nezaželeno akademsko elektronsko pošto.

Profesorji in drugi akademiki v svoj elektronski nabiralnik stalno dobivajo vabila k objavljanju člankov v različnih revijah, k sodelovanju na konferencah ali ponudbe odprtih delovnih mest. Takšne ponudbe se velikokrat ne navezujejo na prejemnikovo področje raziskovanja ali pa je takšnih ponudb preprosto preveč. Velik problem predstavljajo vabila k prispevanju člankov za manj znane ali predatorske revije. Akademiki, ki se strinjajo z objavo svojega članka v takšnih revijah, tvegajo, da je njihova kariera lahko oškodovana. Takšne revije namreč objavijo vsak članek, ki ga prejmejo in s tem razveljavijo akademsko vrednost objavljenih člankov, akademika pa zaznamujejo kot soavtorja predatorske revije (da Silva et al., 2020). Hkrati se nepazljivemu prejemniku lahko zgodi, da preko sporočila posreduje svoje osebne informacije osebam, ki imajo od tega finančno korist (Lin, 2013).

Ker je vsebina akademskih nezaželenih elektronskih sporočil pogosto precej drugačna od nezaželenih elektronskih sporočil, ki jih zazna večina navadnih filtrov nezaželene pošte, mora prejemnik sam ločevati uporabna in neuporabna sporočila. Raziskovalni prispevek našega dela je razvoj orodja za filtriranje nezaželenih akademskih elektronskih sporočil, ki za klasifikacijski model uporablja nevronsko mrežo in dosega primerljive ali celo boljše rezultate od nekaterih raziskovalcev, ki so se ukvarjali s podobnim problemom.

2. Namen članka

Obstoječi filtri nezaželene akademske elektronske pošte so v veliki večini samo "ročno" napisana pravila, ki izključujejo sporočila določenih prejemnikov ali z določenimi ključnimi besedami. Takšna pravila pa je za uspešno delovanje potrebno stalno posodabljati, saj se pošiljatelji, pa tudi vsebina oziroma besede v teh sporočil, ves čas spreminjajo. Zato smo v sklopu raziskave ustvarili filter nezaželene akademske elektronske pošte, ki temelji na modelu nevronske mreže v kombinaciji z vektorskimi vložitvami besed. Začetni model je naučen na množici 660 nezaželenih akademskih elektronskih sporočil, skupaj z 2.551 drugih sporočil. Model se lahko tudi prilagodi uporabniku, tako da upošteva uporabnikova nezaželena akademska elektronska sporočila v njegovem elektronskem nabiralniku.

3. Sorodna dela

3.1. Nezaželena akademska elektronska pošta

V tem razdelku opišemo ugotovitve o nezaželeni akademski elektronski pošti, povzete po različnih avtorjih. Pri pregledu značilnosti smo upoštevali tudi ugotovitve pri pregledu nezaželene akademske elektronske pošte iz naše testne zbirke sporočil.

Nezaželena vabila. Izkoriščevalske ali predatorske revije so revije, katerih glavni cilj ni širjenje znanja ali upoštevanje akademske kvalitete člankov, ampak nepošten zaslužek. Profesorje in druge akademike skušajo pretentati, da bi z njimi sodelovali, s tem da bi jim plačali za objavo svojih člankov. Glavne lastnosti (Wahyudi, 2017) teh revij so:

- za objavo članka je potrebno plačilo,
- revija se izdaja pogosto,
- za objavo je sprejeto nadpovprečno veliko člankov,
- čas obdelave in pregleda člankov sta nerealno hitra in
- kvaliteta objavljenih člankov je slaba ali zelo neenakomerna.

Leta 2014 je knjižničar iz Univerze v Koloradu, Jeffrey Beall, sestavil dva seznama, in sicer seznam vprašljivih

založnikov in seznam vprašljivih revij. Zapisal je, da obstajajo samo zato, da črpajo denar od avtorjev, ki morajo plačati za to, da so njihovi članki sprejeti v revijo (Wahyudi, 2017). Beallov seznam vprašljivih revij (angl. *Beall's list of predatory journals*) se najpogosteje uporablja pri identifikaciji izkoriščevalskih revij. Obstajajo tudi druge zbirke sumljivih revij, kot na primer *Alexa database* in baza lažnih spletnih strani *Phish Tank database* (Dadkhah et al., 2017). Tudi za pomoč pri identifikaciji pravih, strokovnih revij obstajajo baze, kot je na primer Direktorij odprto-dostopnih revij (angl. *Directory of Open Access Journals*) (Kozak et al., 2016).

Na podoben način so zasnovana tudi vabila na konference. V večini primerov se takšna vabila sploh ne navezujejo na prejemnikovo področje raziskovanja in ne obstajajo za širjenje znanja med podobno mislečimi akademiki, ampak je njihov namen oglaševati svoje revije in služiti (D. Cobey et al., 2017).

Zavajanje. Pri zavajanju oziroma ribarjenju (angl. *phishing attacks*) so spletne strani, na katere elektronsko sporočilo usmerja, ustvarjene z namenom, da prejemnik vanje vnese osebne podatke, kot so številka bančne kartice, gesla in podobno (da Silva et al., 2020). Te spletne strani so narejene tako, da so podobne dejanskim stranem resničnih organizacij, zato prejemnik velikokrat sploh ne ve, da gre za ponaredek (Dadkhah et al., 2017). Zavajajoča elektronska sporočila so torej podvrsta nezaželene elektronske pošte, v kateri se pošiljatelj pretvarja, da je predstavnik neke druge legitimne organizacije z namenom pridobivanja osebnih podatkov (Gupta et al., 2018). Sporočila te vrste so večinoma namenjena določeni skupini ljudi ali določeni organizaciji.

Še en način, kako delujejo zavajajoča elektronska sporočila, je s samo-izvršilno kodo. Ta način deluje tako, da se ob kliku na povezavo izvede skrit program in povzroči škodo na prejemnikovem računalniku z vgraditvijo virusa, ki uniči prejemnikove datoteke ali pa ukrade osebne informacije, gesla in druge podatke iz njega (da Silva et al., 2020).

3.2. Generična struktura nezaželenih akademskih sporočil

Wahyudi (2017) je v svojem članku natančno preučil strukturo nezaželene akademske elektronske pošte, zato v nadaljevanju opišemo glavne ugotovitve iz tega in drugih člankov.

Generična struktura nezaželene akademske elektronske pošte je sestavljena iz pozdrava, napovedi, uvoda, osrednjega dela in zaključka. Velikokrat so uporabljeni laskajoči pozdravi in nazivi, kot sta "ugledni profesor" ali "ste strokovnjak na tem področju" (Grey et al., 2016). V pozdravu sta lahko uporabljena tudi prejemnikova ime in priimek. Sporočilo velikokrat izraža hvaljenje, lažne spodbude in obljublja nagrade ali karierne priložnosti (Dadkhah et al., 2017; Soler in Cooper, 2019). Pošiljatelj velikokrat zatrjuje, da je prebral prejemnikov članek in da to sporočilo ni nezaželena pošta (da Silva et al., 2020). V veliki večini sporočilo govori o splošni temi, ki se ne navezuje na prejemnika (Grey et al., 2016; Moher in Srivastava, 2015). Še ena lastnost nezaželene akademske elektronske pošte je ta, da od prejemnika zahteva odgovor v nerealno kratkem času (Dadkhah et al., 2017). V nekaterih primerih se tudi zgodi, da če prejemnik ne odgovori na prvo sporočilo, sledijo nova (Grey et al., 2016).

Tudi pri pošiljateljih nezaželene akademske elektronske pošte so prisotne nekatere skupne značilnosti. Pošiljatelji se ponavljajo ali pošiljajo ponavljajoča sporočila več prejemnikom naenkrat (da Silva et al., 2020). Včasih je elektronski naslov zakrit, ponarejen ali pa se ne sklada s podpisom na koncu besedila (Soler in Cooper, 2019). Elektronski naslovi, ki niso zakriti imajo večinoma uradno domeno institucije, ki ji ukradejo reference (Dadkhah et al., 2017). Poleg tega je v veliko primerih lažno predstavljena lokacija sedeža pošiljatelja (Kozak et al., 2016). To pomeni, da pošiljatelj v sporočilo napiše drugo lokacijo, kot je dejanska lokacija, iz katere je bilo sporočilo poslano.

Opisane značilnosti so povzete iz ugotovitev različnih študij. Tudi pri pregledovanju nezaželene akademske elektronske pošte, ki smo jo uporabili za učno množico, smo opazili podobne značilnosti. Nekateri filtri nezaželene akademske elektronske pošte sicer upoštevajo najdene skupne lastnosti teh sporočil, vendar pa so to v večini "na roko" napisana pravila, ki jih je za dobro delovanje potrebno stalno spreminjati. Zato v nadaljevanju opišemo razvoj filtra nezaželene elektronske pošte, ki deluje na podlagi klasifikatorja, ki avtomatsko klasificira elektronska sporočila.

4. Razvoj filtra nezaželene pošte

V tem poglavju predstavimo zasnovo filtra nezaželene akademske elektronske pošte. Najprej opišemo učno množico sporočil in tehnike obdelave besedila, ki smo jih uporabili. Zatem predstavimo poenostavljen načrt filtra. Poglavje zaključimo z opisom povezave filtra z izbranim odjemalcem elektronske pošte.

4.1. Učna množica elektronskih sporočil

Učno množico elektronskih sporočil smo pridobili iz dveh različnih virov, saj ni bilo mogoče najti ustrezne zbirke akademskih sporočil, ki bi zajemala tako nezaželena kot tudi druga akademska sporočila. Uporabili smo nezaželena akademska sporočila od profesorjev z Univerze v Ljubljani in druga sporočila s spleta. Skupno smo zbrali 660 sporočil, označenih kot nezaželena akademska elektronska sporočila. Drugo skupino sporočil, ki niso nezaželena, smo našli na spletu, in sicer na spletni strani kaggle (van Lit, 2019). Omenjena spletna zbirka vsebuje nezaželeno in drugo elektronsko pošto, vendar pa ta sporočila nimajo akademske vsebine. Za potrebe izdelave našega sistema smo uporabili le sporočila, ki niso nezaželena. Iz omenjene spletne baze sporočil smo dobili 2.551 sporočil, ki smo jih uporabili kot učne primere sporočil, ki niso nezaželena. Ker je bila množica elektronskih sporočil sestavljena iz sporočil iz različnih virov, je bilo potrebno sporočila pretvoriti v enako obliko, primerno za nadaljnjo obdelavo. Poleg tega je bilo potrebno obdelati besedilo sporočila in ga ustrezno spremeniti. V nadaljevanju opišemo, kako smo se lotili tega problema.

Vsako sporočilo v učni množici smo spremenili v slovar s ključi `Subject` (zadeva), `Sender` (pošiljatelj), `Receiver` (prejemnik), `Date` (datum prejema) in `Body` (telo sporočila). Sporočila v skupini sporočil, ki

niso nezaželena, imajo isti vir in obliko. Zato smo vsa sporočila v tej skupini lahko pretvorili v slovar na isti način. Nezaželena sporočila pa smo dobili iz različnih virov in jih je bilo zato potrebno spremeniti v slovar na različne načine glede na končnico datoteke.

Naslednji korak obdelave sporočil je pretvorba slovarjev sporočil v obliko, primerno za model. Lai (2007) v svojem članku trdi, da je najbolj uporaben del za klasifikacijo nezaželenih sporočil zadeva sporočila in da samo telo sporočila ne klasificira tako dobro, kot če je v kombinaciji z zadevo. To smo tudi preizkusili in se odločili, da tudi mi uporabimo kombinacijo zadeve in telesa sporočila. Poleg tega Méndez et al. (2006) pravijo, da priloga, ki je lahko priložena sporočilu in jo spremenimo v besedilo, doda nepotrebne informacije, ki niso dobre za klasifikacijo. Zato priloge sporočila nismo uporabili.

Sledi opis obdelave besedila sporočil. V zadevi so bile v nekaterih primerih v oglatih oklepajih zapisane oznake sporočila (na primer `INBOX`). Zato smo iz besedila odstranili del, ki je v oglatih oklepajih. Predvsem v množici sporočil, ki niso nezaželena, je veliko sporočil, ki vsebujejo druga sporočila (nizi izmenjujočih odgovorov). Zato smo morali najti takšne dele sporočil in jih odstraniti. To smo naredili tako, da smo odstranili vrstice, ki se začnejo z določenim znakom ali nizom znakov, kot so na primer: "`To:`", "`From:`", "`Wrote:`" itd.

Nato smo zadevo in telo sporočila obdelali na enak način. Najprej smo odstranili velike začetnice in celotno sporočilo spremenili v male črke. Opazili smo, da se v nekaterih nezaželenih sporočilih pojavljajo znaki, ki izgledajo kot črke, vendar so v resnici drugi znaki in jih program zazna kot ločila. Primeri znakov, ki smo jih našli v naši zbirki sporočil, so prikazani na sliki 1. Pošiljatelji nezaželenih akademskih sporočil so s tem očitno želeli preprečiti filtrom nezaželene elektronske pošte razpoznavo nekaterih besed, ki nakazujejo na nezaželeno akademsko elektronsko pošto. Najdene znake smo zamenjali s pravo črko in na koncu besedila dodali značko "`specialchars`". V besedilu smo poiskali klicaje in jih zamenjali z značko "`exclamationmark`". Opazili smo namreč, da izrazito velika raba klicajev lahko nakazuje na to, da je sporočilo nezaželena akademska elektronska pošta. Poleg tega smo poiskali elektronske naslove, povezave in imena mesecev ter jih zamenjali z značkami "`emailwashere`", "`linkwashere`" in "`monthwashere`", saj struktura elektronskega naslova in povezave ter ime meseca niso pomembni. Poleg omenjenega smo iz sporočil odstranili ločila in nepotrebne besede, kot so vezniki, zaimki in vprašalnice. V angleščini se te pogoste in nepotrebne besede imenujejo *stop words*.

Da bi identiteta profesorjev, ki so prispevali nezaželena akademska sporočila za učno množico, ostala skrita, je bilo potrebno iz sporočil odstraniti imena prejemnikov. Poleg tega smo odstranili tudi imena pošiljatelja, saj je tudi ta podatek nepotreben pri klasifikaciji. Ime ali priimek smo zamenjali z značko `receivername` za prejemnika oziroma `sendername` za pošiljatelja.

Na opisani način smo sporočila spremenili iz seznama slovarjev v seznam besedil oziroma zbirko besedil (angl. *corpus*). Sporočila smo nato shranili s pomočjo knji-

Znak	Črka
α	a
a	
b	b
c	c
c	
c	
d	d
e	e
e	

Znak	Črka
l	i
i	
i	
ι	
l	
j	j
μ	m
o	o
p	p

Znak	Črka
ρ	p
r	r
s	s
s	
u	u
u	
υ	
v	v

Slika 1: Primeri znakov iz nezaželenih akademskih sporočil in ustrezne črke, ki so jih pošiljatelji nezaželenega akademskega sporočila zamenjali.

žnice, ki objekt serializira (angl. *serialize*) in s tem spremeni v binarni tok (angl. *byte stream*). Tako shranjenih sporočil ne moremo brati direktno iz datoteke, ampak jih je za branje potrebno pretvoriti nazaj v besedilo.

4.2. Tehnike obdelave sporočil

Štetje ponovitev besed. Najbolj enostavna tehnika obdelave sporočil je štetje ponovitev besed v posameznem sporočilu. Zbirke besedil smo spremenili v matriko, v kateri vsaka vrstica predstavlja sporočilo, stolpec pa besedo. Ker je vseh besed v vseh sporočilih lahko zelo veliko, smo se omejili na 2000 besed. Poskusili smo tudi z odstranitvijo besed, ki se pojavijo v manj kot treh sporočilih, tako kot so to opisali Sakkis in sodelavci (Sakkis et al., 2003).

Frekvenca besed z inverzno frekvenco v dokumentih (angl. *term frequency-inverse document frequency* - TF-IDF). Frekvenca besed (angl. *term frequency*) enostavno pomeni število besed v posameznem sporočilu. Inverzna frekvenca v dokumentih (angl. *inverse document frequency*) pa predstavlja informativnost besede, torej ali se beseda pogosto ali redko pojavlja v sporočilih (Hakim et al., 2014). TF-IDF besede izračunamo tako, da uporabimo enačbo (1), pri čemer je t têrmin in d dokument oziroma sporočilo. $TF(t, d)$ predstavlja frekvenco besede t v sporočilu d, $IDF(t)$ pa je inverzna frekvenca besede t v dokumentih. Izračuna se jo z enačbo (2), pri čemer je n število vseh sporočil, $DF(t)$ pa število sporočil v katerih se beseda t pojavi vsaj enkrat. V imenovalcu ulomka vrednosti $DF(t)$ dodamo še $+1$, da se izognemo deljenju z nič.

$$TF\text{-}IDF(t, d) = TF(t, d) * IDF(t) \tag{1}$$

$$IDF(t) = log\left(\frac{n}{DF(t) + 1}\right) + 1 \tag{2}$$

Prednost uporabe tehnike TF-IDF je, da se normalizira vpliv besed, ki se v dokumentih pojavljajo zelo pogosto in so zato manj informativne kot besede, ki se pojavijo manjkrat.

Medsebojna informacija. Za izbiro atributov smo preizkusili tudi odstranitev atributov, ki imajo premajhno medsebojno informacijo (angl. *mutual information*). Medse-

bojna informacija dveh naključnih spremenljivk je nenegativna vrednost, ki pove odvisnost med tema spremenljivkama (Kraskov et al., 2004). Z drugimi besedami, medsebojna informacija meri količino informacije, ki jo pridobimo o neki spremenljivki, če imamo podano neko drugo spremenljivko (Witten in Frank, 2000). Večja ko je medsebojna informacija dveh spremenljivk, bolj sta spremenljivki odvisni med sabo. Če pa je medsebojna informacija enaka nič, sta spremenljivki popolnoma neodvisni. Medsebojno informacijo med dvema naključnima spremenljivkama lahko izračunamo z enačbo (3), kjer $I(X;Y)$ predstavlja medsebojno informacijo za spremenljivki X in Y, $H(X)$ predstavlja entropijo spremenljivke X, $H(X|Y)$ pa je pogojna entropija za spremenljivko X, če imamo podano spremenljivko Y. Entropija je enaka povprečni lastni informaciji in prestavlja stopnjo negotovosti oziroma informacije. Izračunamo jo s pomočjo formule (4), kjer so možni rezultati $x_1, ..., x_n$ in $P(x_i)$ verjetnost rezultata x_i. Pogojno entropijo pa izračunamo s formulo (5).

$$I(X;Y) = H(X) - H(X|Y) \tag{3}$$

$$H(X) = -\sum_{i-1}^{n} P(x_i) log P(x_i) \tag{4}$$

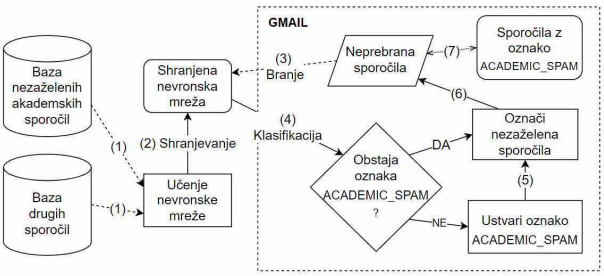
$$H(X|Y) = -\sum_{x \in X, y \in Y} p(x,y) log \frac{p(x,y)}{p(x)} \tag{5}$$

Vektorska vložitev besed (angl. *word to vector embedding*). Vektorska vložitev besed je tehnika predstavitve besed z vektorji, ki ohranjajo pomenske značilnosti besed. To pomeni, da so besede, ki so si pomensko bolj podobne, bolj blizu v vektorskem prostoru (Ghannay et al., 2016). Vektorje besed se sestavi glede na to, katere besede se v stavku nahajajo skupaj, saj se tako najlažje ugotovi pomen besede. Zaradi tega je za učinkovito sestavljanje besednih vektorjev potrebna velika učna množica besedil. Ker je to velikokrat težko pridobiti in ker je učenje vektorjev lahko precej zamudno, na spletu obstajajo baze besed in njihovih vektorjev, ki so naučeni na velikih množicah besedil. Primeri zbirk naučenih vektorjev so na primer `Googlova zbirka` in `zbirka GloVe` (*Global Vectors for Word Representation*) (Pennington, 2014).

Ker je zbirka vektorjev iz baze GloVe naučena na ogromni množici besedil in je prosto dostopna, smo se odločili, da jo bomo uporabili v našem sistemu. Na voljo ima več zbirk vektorjev iz različnih virov in velikosti. Zbirke smo preizkusili in ocenili njihovo uspešnost. Preizkusili smo tudi različno maksimalno število besed v posameznem sporočilu in maksimalno število unikatnih besed. V končnem sistemu smo uporabili 100-dimenzionalne vektorje, omejitev 2.000 besed na sporočilo in omejitev 500.000 različnih besed.

4.3. Zasnova filtra nezaželenih akademskih sporočil

Zgradili smo programsko rešitev, sestavljeno iz dveh programov. Prvi, glavni program, je namenjen klasifikaciji neprebranih sporočil. Drugi program pa je namenjen posodobitvi nevronske mreže glede na uporabnikova označena sporočila. Pri gradnji programske rešitve smo preizkusili več klasifikatorjev, in sicer smo preizkusili naivni Bayes, naključni gozd, metodo podpornih vektorjev, logistično regresijo in različne nevronske mreže. V končnem sistemu smo uporabili nevronsko mrežo, saj so bili rezultati testiranja pri tem klasifikacijskem modelu najboljši. Za odjemalec elektronske pošte pa smo izbrali Gmail (Google, 2022).

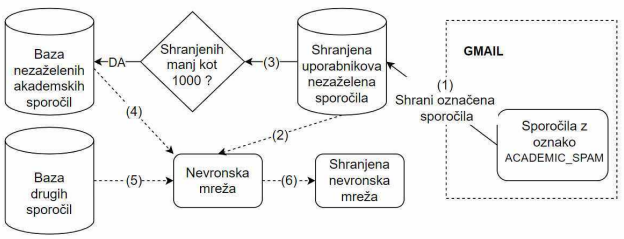


Slika 2: Načrt delovanja sistema ob prvem učenju nevronske mreže in ob klasifikaciji neprebranih sporočil.

Slika 2 prikazuje delovanje sistema ob zagonu programa za klasifikacijo neprebranih sporočil. Najprej program preveri, ali je nevronska mreža že shranjena na disku. Če ni, se izvede začetno učenje nevronske mreže. Za učenje nevronske mreže so potrebni označeni učni podatki, kar so v našem primeru nezaželena akademska elektronska sporočila in druga elektronska sporočila. Ker je vir sporočil lahko različen, se elektronska sporočila pretvori v enako obliko in obdela tako, da se odstrani nepotrebne atribute sporočila. Ta korak je na sliki označen s številko (1). Sledi učenje nevronske mreže in shranjevanje na disk (2). Nevronska mreža je tako ob naslednjem zagonu pripravljena na klasifikacijo in ni potrebno pri vsakem zagonu čakati na učenje nevronske mreže. Naslednji korak programa je branje neprebranih sporočil iz elektronskega nabiralnika (3). Prebrana sporočila se obdelajo na enak način kot pri koraku (1). Shranjena nevronska mreža nato klasificira neprebrana sporočila. Če je katero izmed neprebranih sporočil klasificirano kot nezaželena akademska elektronska pošta, program preveri, ali v uporabnikovem elektronskem nabiralniku že obstajajo sporočila z oznako ACADEMIC_SPAM. Če ne, program za uporabnika ustvari novo oznako ACADEMIC_SPAM in označi ustrezna sporočila (5). Če oznaka že obstaja, program samo označi ustrezna sporočila s to oznako. Oznaka se nato prikaže na neprebranih sporočilih v uporabnikovem elektronskem nabiralniku (6), hkrati pa nastane oziroma se dopolnjuje tudi mapa sporočil z oznako ACADEMIC_SPAM (7).

Slika 3 prikazuje drugi program, ki je namenjen posodobitvi nevronske mreže, tako da se čim bolj prilagodi uporabniku. Posodobitev deluje samo, če ima uporabnik v svojem elektronskem nabiralniku sporočila, označena z oznako ACADEMIC_SPAM.

Program najprej prebere sporočila, ki so označena z oznako ACADEMIC_SPAM. Nato ta sporočila doda k shranjenim uporabniškim sporočilom ali pa ustvari novo datoteko s shranjenimi uporabnikovimi nezaželenimi akademskimi sporočili (1). Ta sporočila se potem uporabi kot del učne množice pri učenju nevronske mreže (2). Če je shra-



Slika 3: Načrt delovanja sistema ob posodobitvi nevronske mreže.

njenih uporabnikovih sporočil več kot 1000, se sporočila razvrsti po datumu prejema in se jih izbere le zadnjih 1000. Če pa je shranjenih uporabnikovih sporočil manj kot 1000 (3), se množica nezaželenih akademskih sporočil dopolni s sporočili iz baze nezaželenih akademskih sporočil (4). Poleg nezaželenih akademskih sporočil nevronska mreža za učenje potrebuje tudi množico drugih sporočil. Te se pridobijo iz baze drugih sporočil (5). Nevronska mreža se nato nauči na podanih učnih podatkih in posodobljena mreža se shrani na disk (6), kjer je na voljo za naslednjo klasifikacijo neprebranih sporočil.

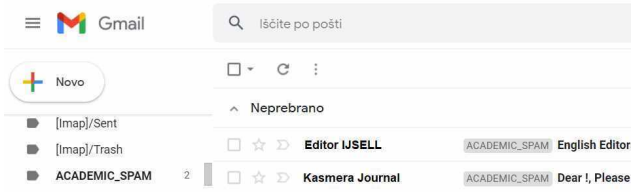
4.4. Povezava z odjemalcem elektronske pošte

Filter nezaželene akademske elektronske pošte smo povezali z brezplačno e-poštno storitvijo, ki jo ponuja Google, in sicer Gmail. Za povezavo tega spletnega odjemalca elektronske pošte s programom smo uporabili Gmail API. To je aplikacijski programski vmesnik, ki temelji na arhitekturi REST (angl. *RESTful API*) (Developers, 2021). Arhitektura REST (angl. *representational state transfer*) je arhitektura za izmenjavo podatkov med spletnimi storitvami, kjer je vsak vir dostopen z enoličnim identifikatorjem vira URL. Uporablja se za dostop do Gmail elektronskih nabiralnikov in pošiljanje elektronskih sporočil preko programa.

Program smo z Gmail API-jem povezali s pomočjo računalniškega okolja `Google Cloud`. Tam smo ustvarili nov projekt in v njem omogočili Gmail API ter dodali avtorizacijo in avtentikacijo za program. Uporabili smo `API Keys` in `OAuth 2.0 Client IDs` za omogočanje `Gmail APIja` v programu.

Če je povezovanje z `Gmail APIjem` uspešno, je program pripravljen na branje uporabnikovih neprebranih sporočil. V primeru, da neprebrana sporočila ne obstajajo, se izpiše sporočilo: `"No messages found."` in program se zaključi. V nasprotnem primeru pa se iz podatkov, pridobljenih z `Gmail APIjem`, generira slovar s ključi `Subject` (zadeva), `Sender` (pošiljatelj), `Receiver` (prejemnik), `Date` (datum prejema) in `Body` (telo sporočila). Sporočila v obliki slovarja je nato potrebno preurediti v obliko, primerno za klasifikator, podobno kot smo to naredili za sporočila v učni množici (glej razdelek 4.1.). Tako smo namesto seznamov slovarjev dobili seznam obdelanih besedil. Ta seznam smo nato s pomočjo shranjenih vektorjev spremenili v seznam vektorjev in ga pretvorili v matriko.

Naslednji korak je nalaganje shranjenega klasifikatorja in klasifikacija neprebranih sporočil. Če klasifikator označi katerega izmed sporočil kot nezaželeno akademsko pošto, se izvede del programa za posodobitev oznak. Najprej program preko `Gmail APIja` prebere vse oznake, ki obstajajo v uporabnikovem elektronskem nabiralniku, in preveri, ali je katera med njimi `ACADEMIC_SPAM`. Če oznaka že obstaja, se sporočilom, ki jih je klasifikator označil kot nezaželena, doda ta oznaka. Če oznaka še ne obstaja, pa se ustvari nova oznaka `ACADEMIC_SPAM`.



Slika 4: Izsek elektronskega nabiralnika v Gmailu, kjer sta bili dve neprebrani sporočili klasificirani kot nezažalena akademska pošta.

Rezultat zagona programa in klasifikacije neprebranih sporočil je oznaka `ACADEMIC_SPAM`, ki se prikaže na ustreznih sporočilih. Na sliki 4 je prikazan primer takšne klasifikacije v Gmailu. Pred zadevo sporočil, klasificiranih kot nezaželena akademska sporočila, se pojavi oznaka `ACADEMIC_SPAM`. Hkrati pa lahko na levi strani v seznamu vseh oznak opazimo oznako `ACADEMIC_SPAM`, kjer lahko najdemo vsa sporočila, ki so bila v preteklosti označena kot nezaželena akademska sporočila.

5. Testiranje in rezultati

V zadnjem poglavju predstavljamo način testiranja preizkušenih modelov klasifikacije in obdelave elektronskih sporočil. Primerjamo rezultate in prikažemo rezultate algoritma SHAP, ki poišče besede, ki so najbolj pripomogle h klasifikaciji nezaželenih akademskih sporočil.

5.1. Način testiranja

Za testiranje uspešnosti smo uporabili 10-kratno prečno preverjanje (angl. *10-fold cross-validation*). Pri tej metodi učno množico razdelimo na 10 približno enako velikih množic in za vsak model naredimo 10 ponovitev testiranja. V vsaki iteraciji vzamemo za testno množico eno izmed množic, ostale množice pa združimo v učno množico.

Na tak način bolj natančno preverimo uspešnost modelov, kot če bi iz množice naključno izbrali 10% primerov in le enkrat testirali model. Pri 10-kratnem prečnem preverjanju je namreč vsak primer v množici enkrat uporabljen kot testni. Tako lahko na koncu izračunamo povprečje in standardno deviacijo rezultatov iz vseh ponovitev testiranja ter dobimo bolj realne rezultate. Poleg tega smo lahko zaradi večkratne ponovitve testiranja za primerjavo modelov uporabili tudi statistične teste.

5.2. Rezultati

Del rezultatov testiranja različnih modelov je prikazan v tabeli 1. Preizkusili smo različne tehnike obdelave besedila, v tabeli pa so prikazani rezultati ob uporabi tehnike TF-IDF z odstranitvijo besed, ki se pojavijo v manj kot treh sporočilih in besed, ki imajo medsebojno informacijo manjšo kot 0.01 pri prvih petih modelih ter

vektorsko vložitvijo besed pri zadnjem modelu nevronske mreže. Uporabili smo celotno množico sporočil, in sicer 660 nezaželenih akademskih sporočil in 2.551 drugih sporočil. Kot lahko vidimo, so rezultati že pri teh modelih precej dobri, saj so pravilno klasificirana skoraj vsa sporočila iz testne množice sporočil.

Tabela 1: Povprečne vrednosti in standardna deviacija testiranja z 10-kratnim prečnim preverjanjem.

Model	Točnost	F1	AUC
Naivni Bayes	88.49% ± 2.40%	77.29% ± 5.27%	0.91 ± 0.02
Naključni gozd	98.32% ± 0.32%	95.88% ± 0.79%	0.96 ± 0.01
SVM	98.65% ± 0.68%	96.62% ± 1.79%	0.97 ± 0.01
Logistična regresija	98.82% ± 0.58%	97.02% ± 1.55%	0.97 ± 0.01
Nevronska mreža	98.98% ± 0.42%	97.49% ± 1.10%	0.98 ± 0.01
Nevronska mreža z GloVe	98.69% ± 0.62%	96.79% ± 1.47%	0.97 ± 0.01

Tabela 2: Povprečni rangi uspešnosti modelov glede na vrednost AUC.

Naivni Bayes	Naključni gozd	SVM	Logistična regresija	Nevronska mreža
5	2.9	2.65	2.15	2.3

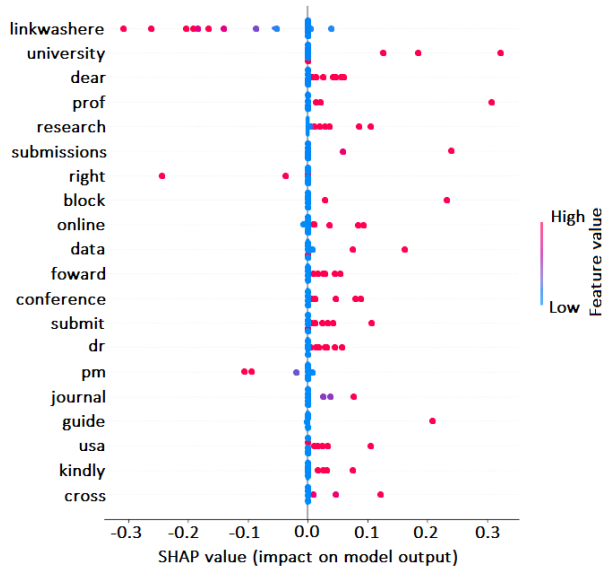
Za primerjavo modelov smo uporabili Friedmanov test (Friedman, 1937). Natančno razlago uporabe tega testa opisuje Demšar (2006). Najprej smo primerjali skupino klasifikacijskih modelov, na katerih smo uporabili prej omenjene tehnike obdelave besedila in so v tabeli na prvih petih mestih. S Friedmanovim testom pri $\alpha = 0.05$ na AUC smo preverili, ali lahko za kateri par modelov rečemo, da je eden izmed njiju izrazito boljši od drugega. Povprečni rangi uspešnosti modelov glede na AUC so razvidni v tabeli 2. Izračunali smo kritično razdaljo $CD = 1.93$ in jo primerjali z razlikami povprečnih vrstnih redov uspešnosti modelov ter ugotovili, da so vsi modeli izrazito boljši za klasifikacijo nezaželenih akademskih sporočil kot naivni Bayes. Za ostale pare modelov s Friedmanovim testom tega nismo mogli dokazati.

Čeprav so bili rezultati že pri teh modelih precej dobri, smo se vseeno odločili implementirati še različne modele nevronskih mrež v kombinaciji z vektorskimi vložitvami besed. Zgradili smo več različnih nevronskih mrež in jih med seboj primerjali. V tabeli 2 je na zadnjem mestu prikazan rezultat ene izmed teh nevronskih mrež. Čeprav je rezultat nekoliko slabši od zgoraj opisanih modelov, smo v končnem sistemu vseeno uporabili to nevronsko mrežo z vektorskimi vložitvami besed. Ta metoda namreč upošteva pomene besed in ne samo njihov zapis, tako kot ostale metode obdelave besedil.

Rezultati našega testiranja so nekoliko boljši od rezultatov nekaterih raziskovalcev, ki so se ukvarjali s podobnim problemom. Sicer nismo našli primerov, v katerih bi raziskovalci skušali klasificirati nezaželeno akademsko elektronsko pošto, vseeno pa lahko do neke mere primerjamo naše rezultate z rezultati navadnih filtrov nezaželenih elektronskih sporočil. Koprinska et al. (2007) so na eni izmed testnih množic z modelom naključnega gozda dosegli točnost 96.03%, natančnost 95.62%, priklic 95.62% in F1 mero 94.16%. Za obdelavo sporočil so uporabili posebno metodo izbire atributov, in sicer varianco frekvence têrma (angl. *term frequency variance*). Ostali modeli v njihovem primeru niso bili tako uspešni. Lai (2007) je v članku opisal preizkus modelov Naivnega Bayesa, k-najbližjih sosedov, SVM in kombinacijo TF-IDF z metodo SVM. Za najbolj uspešen model se je izkazala kombinacija TF-IDF z metodo SVM. S to metodo so v enem primeru dosegli točnost 93.43%.

5.3. Razlaga klasifikacije z algoritmom SHAP

Razumljivost in enostavna razlaga modela sta izjemno pomembni za interpretacijo rezultatov in možnost nadgradnje modela. To je velikokrat razlog, da se nekateri raziskovalci odločijo za uporabo enostavnih (linearnih) modelov namesto kompleksnejših, ki jih je težko razumeti. Vendar pa je zaradi naraščajoče količine podatkov, ki jih želimo obdelati, nujno, da uporabljamo tudi slednje. Za to obstajajo algoritmi, ki nam jih pomagajo razumeti in interpretirati rezultat njihove klasifikacije. Eden takšnih algoritmov je algoritem SHAP (Lundberg in Lee, 2017).



Slika 5: Slika prikazuje besede, ki najbolj vplivajo na rezultat klasifikacije nevronske mreže. Besede, ki imajo več pik na desni strani, so prispevale k temu, da je bilo sporočilo klasificirano kot nezaželeno.

Algoritem SHAP (*SHapley Additive exPlanations*) oziroma Shapleyjeve aditivne razlage je algoritem, ki za podane primere razloži, zakaj jih je model klasificiral tako, kot jih je. Z drugimi besedami, algoritem SHAP nam pove, kako posamezen atribut vpliva na napoved modela. V primeru klasifikacije sporočil s tem algoritmom torej lahko ugotovimo, katere besede najbolj vplivajo na rezultat klasifikacije. Na sliki 5 je prikazan rezultat algoritma SHAP na eni izmed ustvarjenih nevronskih mrež. Zaradi zahtevnosti algoritma smo uporabili manjšo podmnožico testnih

sporočil. Graf na sliki od spodaj navzgor prikazuje katere besede naj bi najbolj vplivale na klasifikacijo nezaželenih sporočil. Beseda `linkwashere`, s katero smo nadomestili vse url povezave, očitno najbolje nakazuje na to, da sporočilo ni nezaželeno. Besede, ki močno nakazujejo na to, da je sporočilo nezaželeno akademsko elektronsko sporočilo pa so `university` (univerza), `dear` (dragi oz. spoštovani), `prof` (kratica za profesor), `research` (raziskava) in `submissions` (oddaje).

6. Zaključek

Za cilj smo si zadali izdelavo filtra nezaželene akademske elektronske pošte, ki bi med neprebranimi elektronskimi sporočili v uporabnikovem elektronskem nabiralniku, čim bolj učinkovito poiskal nezaželena akademska elektronska sporočila in jih označil. Za dosego tega cilja smo morali preučiti strukturo in skupne značilnosti nezaželene akademske elektronske pošte ter preiskati obstoječe načine filtriranja nezaželene elektronske pošte. S testiranjem smo določili, da je model nevronske mreže najbolj učinkovit pri filtriranju nezaželene akademske elektronske pošte, zato smo ga tudi uporabili v končnem sistemu.

Ugotovili smo, da obstaja zelo malo rešitev za filtriranje nezaželene elektronske pošte, katerih osrednji cilj bi bil filtriranje nezaželenih akademskih elektronskih sporočil. Velika večina teh rešitev uporablja le prepoznavanje znanih pošiljateljev nezaželenih akademskih sporočil, vendar pa je za učinkovitost tega načina filtriranja potrebno stalno posodabljanje seznama. Zato smo implementirali sistem, ki neprebrana elektronska sporočila klasificira kot nezaželeno akademsko elektronsko pošto, glede na pomen besed v sporočilih. To smo dosegli z vektorsko vložitvijo besed v kombinaciji z modelom nevronske mreže. Poleg tega smo izdelali program, ki lahko klasifikacijski model posodobi glede na uporabnikovo nezaželeno akademsko elektronsko pošto. Na tak način se model lahko prilagodi uporabnikovemu elektronskemu nabiralniku in še bolj natančno označuje nezaželena akademska elektronska sporočila.

Ena izmed večjih pomanjkljivosti opisane rešitve je nadomestitev akademskih sporočil, ki niso nezaželena, z navadnimi nezaželenimi sporočili. Zaradi varovanja osebnih podatkov namreč nismo mogli uporabiti sporočil profesorjev, pa tudi na spletu ni bilo mogoče najti zbirk s takšnimi akademskimi sporočili. Tudi profesorji in drugi akademiki sicer dobivajo takšna navadna sporočila in so zato tudi ta sporočila do neke mere ustrezna za učno množico. Vseeno pa bi bilo potrebno preveriti, da klasifikator zaradi pomanjkanja akademskih sporočil, ki niso nezaželena, ne označi kar vseh akademskih sporočil, kot nezaželena.

Sistem bi lahko izboljšali še tako, da bi ob posodobitvi modela upoštevali ne samo uporabnikova nezaželena akademska sporočila, ampak tudi druga sporočila. Poleg tega sistem trenutno dobro deluje le za angleška sporočila, saj je naša množica učnih sporočil bila sestavljena le iz angleških sporočil. Možna izboljšava bi torej lahko bila prepoznavanje jezikov in prilagajanje filtra nanje. Dodali bi lahko tudi uporabniški vmesnik, ki bi uporabniku olajšal uporabo sistema.

7. Zahvala

Zahvaljujem se prof. dr. Zoranu Bosniću za vodenje, nasvete in mentorstvo med raziskavo ter profesorjem Fakultete za računalništvo in informatiko, Univerze v Ljubljani, ki so prispevali nezaželena akademska elektronska sporočila za učno množico sporočil.

8. Literatura

Kelly D. Cobey, Miguel de Costa e Silva, Sasha Mazzarello, Carol Stober, Brian Hutton, David Moher in Mark Clemons. 2017. Is this conference for real? navigating presumed predatory conference invitations. *Journal of oncology practice*, 13(7):410–413.

Jaime A Teixeira da Silva, Aceil Al-Khatib in Panagiotis Tsigaris. 2020. Spam emails in academia: issues and costs. *Scientometrics*, 122(2):1171–1188.

Mehdi Dadkhah, Glenn Borchardt in Tomasz Maliszewski. 2017. Fraud in academic publishing: researchers under cyber-attacks. *The American journal of medicine*, 130(1):27–30.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

Google Developers. 2021. Gmail api overview. `https://developers.google.com/gmail/api/guides`.

Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

Sahar Ghannay Ghannay, Benoit Favre, Yannick Esteve in Nathalie Camelin. 2016. Word embedding evaluation and combination. V: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, str. 300–305. European Language Resources Association (ELRA).

Google. 2022. Gmail: Brezplačna, zasebna in varna e-pošta. `https://www.google.com/intl/sl/gmail/about/`, pridobljeno: 2022-01-08.

Andrew Grey, Mark J. Bolland, Nicola Dalbeth, Greg Gamble in Lynn Sadler. 2016. We read spam a lot: prospective cohort study of unsolicited and unwanted academic invitations. *BMJ*, 355.

Brij B. Gupta, Nalin AG Arachchilage in Kostas E. Psannis. 2018. Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Telecommunication Systems*, 67(2):247–267.

Ari Aulia Hakim, Alva Erwin, Kho I Eng, Maulahikmah Galinium in Wahyu Muliady. 2014. Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (tf-idf) approach. V: *2014 6th international conference on information technology and electrical engineering (ICITEE)*, str. 1–4. IEEE.

Irena Koprinska, Josiah Poon, James Clark in Jason Chan. 2007. Learning to classify e-mail. *Information Sciences*, 177(10):2167–2187.

Marcin Kozak, Olesia Iefremova in James Hartley. 2016. Spamming in scholarly publishing: A case study. *Jour-*

nal of the Association for Information Science and Technology, 67(8):2009–2015.

Alexander Kraskov, Harald Stögbauer in Peter Grassberger. 2004. Estimating mutual information. *Physical review E*, 69(6):066138.

Chih-Chin Lai. 2007. An empirical study of three machine learning methods for spam filtering. *Knowledge-Based Systems*, 20(3):249–254.

Songqing Lin. 2013. Why serious academic fraud occurs in China. *Learned Publishing*, 26(1):24–27.

Scott M Lundberg in Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

José Ramon Méndez, Florentino Fdez-Riverola, Fernando Díaz, Eva Lorenzo Iglesias in Juan Manuel Corchado. 2006. A comparative performance study of feature selection methods for the anti-spam filtering domain. V: *Industrial Conference on Data Mining*, str. 106–120. Springer.

David Moher in Anubhav Srivastava. 2015. You are invited to submit.... *BMC medicine*, 13(1):1–4.

Jeffrey Pennington. 2014. Glove: Global vectors for word representation. `https://nlp.stanford.edu/projects/glove/`, pridobljeno: 2022-07-15.

Georgios Sakkis, Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Constantine D. Spyropoulos in Panagiotis Stamatopoulos. 2003. A memory-based approach to anti-spam filtering for mailing lists. *Information retrieval*, 6(1):49–73.

Josep Soler in Andrew Cooper. 2019. Unexpected emails to submit your work: Spam or legitimate offers? the implications for novice english l2 writers. *Publications*, 7(1):7.

Wessel van Lit. 2019. Email spam Kaggle. https://www.kaggle.com/veleon/ham-and-spam-dataset.

Ribut Wahyudi. 2017. The generic structure of the call for papers of predatory journals: A social semiotic perspective. V: *Text-based research and teaching*, str. 117–136. Springer.

Steve Whittaker, Victoria Bellotti in Paul Moody. 2005. Introduction to this special issue on revisiting and reinventing e-mail. *Human–Computer Interaction*, 20(1-2):1–9.

Ian H Witten in Eibe Frank. 2000. *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann.

Preparing a Corpus and a Question Answering System for Slovene

Matjaž Zupanič*, Maj Zirkelbach*, Uroš Šmajdek*, Meta Jazbinšek†

*Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, SI-1000 Ljubljana
{mz4689, mz5153, us6796}@student.uni-lj.si

†Department of Translation Studies, Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana
mj6953@student.uni-lj.si

Abstract

Lack of proper training data is one of the key issues when developing natural language processing models based on less-resourced languages, such as Slovene. In this paper we discuss machine translation as a solution to this issue, with the focus on question answering (QA). We use the SQuAD 2.0 dataset, which we have translated using eTranslation machine translator. To improve the reliability of translations, we translate the answers together with the context instead of separately, reducing the rate at which answers were not found in the context from 56% to 7%. For comparison, we also perform manual post-editing of the small subset of machine translations. We then compare these datasets utilizing various transformer-based QA models and observe the differences between the datasets and different model configurations. The results have shown little distinction between monolingual and larger multilingual models: monolingual SloBERTa scored 64.9% exact matches on the machine translated dataset and 72.6% exact matches on human translated one, whereas multilingual RemBERT scored 64.2% exact matches on the machine translated dataset and 71.9% exact matches on human translated one. Additionally, using machine translated dataset in the evaluation produces notably worse results then the human translated dataset. Qualitative analysis of the translations has shown that mistakes often occur when the sentences are longer and have more complicated syntax.

1. Introduction

One of the goals in artificial intelligence is to build intelligent systems that would be able to interact with humans and help them. One of such tasks is reading the web and then answer complex questions about any topic over given content. These question-answering (QA) systems could have a big impact on the way that we access information. Furthermore, open-domain question answering is a benchmark task in the development of Artificial Intelligence, since understanding text and being able to answer questions about it is something that we generally associate with intelligence.

Recently, pre-trained Contextual Embeddings (PCE) models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and A Lite BERT (ALBERT) (Lan et al., 2020) have attracted lots of attention due to their great performance in a wide range of NLP tasks.

Multilingual question answering tasks typically assume that answers exist in the same language as the question. Yet in practice, many languages face both information scarcity—where languages have few reference articles—and information asymmetry—where questions reference concepts from other cultures. Due to the sizes of modern corpora, performing human translations is generally infeasible, therefore we often employ machine translations instead. Machine translation however is for the most part incapable of interpreting nuances of specific languages such as culturally specific vocabulary or for example the use of articles, indication of grammatical number or gender and conjugation endings when comparing English and Slovene.

In this work we present a method for a construct of a machine translated dataset from SQuAD 2.0 (Rajpurkar et al., 2018) and evaluate its quality using various modern QA models. Additionally, we benchmark its effectiveness by performing manual post-editing on a subset of the translated dataset and comparing the results.

The main contributions of our work are:

- a pipeline for translation of English question answering dataset;
- a Slovene monolingual model SloBERTa, fine-tuned on machine translated data and three different fine-tuned multilingual QA models, M-BERT, XLM-R and CroSloEngual BERT, all on machine translated and both original and machine translated data; and
- comparison of human and machine translated data in terms of question answering performance.

In Section 2 we present the related work. In Section 3 we present our dataset, the process of translation and post-edition, and evaluate the quality of the translation. In Section 4 we give a brief overview of the models used in the evaluation. In Section 5 we present the evaluation and discuss the results in Section 6. In Section 7 we present the conclusions and give possible extensions and enhancements for future work.

2. Related work

Early question answering systems, such as LUNAR (Woods and WA, 1977), date back to the 60's and the 70's. They were characterised by a core database and a set of rules, both handwritten by experts of the chosen domain. Over time, with the development of large online text

repositories and increasing computer performance, the focus shifted from such rule-based system to using machine learning and statistical approaches, like Bayesian classifiers and Support Vector Machines. An example of this kind of system that was able to perform question answering on Slovene language was presented by Čeh et al. (Čeh and Ojsteršek, 2009) in 2009.

Another major revolution in the field of question answering and natural language processing in general was the advent of deep learning approaches and self-attention. One of the most popular approaches of this kind is BERT (Devlin et al., 2018), a transformer model introduced in 2019. Since then it has inspired many other transformed based models, for instance RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and T5 (Raffel et al., 2020) , xlm and XLNet (Yang et al., 2019).

Such models also have the advantage of being able to recognise multiple languages, giving rise to multilingual models and model variants, such as M-BERT, XLM-R (Conneau et al., 2019), mT5 (Xue et al., 2021) and RemBERT (Chung et al., 2020). Nevertheless, the training requires large amounts of training data, which many languages lack, leading to varying performance between different languages. They have also shown to perform worse than monolingual models (Martin et al., 2020; Virtanen et al., 2019). As such Ulčar et al. (Ulčar and Robnik-Šikonja, 2020) made an effort to strike a middle ground between the performance of monolingual and versatility of multilingual models by reducing the number of languages in multilingual model to three; two similar less-resourced languages from the same language family and English. This resulted in two trilingual models FinEst BERT and CroSloEngual BERT al. (Ulčar and Robnik-Šikonja, 2020).

In 2020, a Slovene monolingual RoBERTa-based model SloBERTa (Ulčar and Robnik-Šikonja, 2021) was introduced. It was trained on 5 different corpora, totaling 3.41 billion words. The latest version of the model is SloBERTa 2.0, augmenting the original model by more than doubling the number of training iterations. The authors evaluated its performance on named-entity recognition, part-of-speech tagging, dependency parsing, sentiment analysis and word analogy, but not on question answering.

While the described advancements of natural language processing models already offer us a partial solution for the lack of language-specific training corpora, namely the ability to train the model on a language where large corpora are present (e.g. English), the models still require language-specific fine-tuning, for which a sizable corpora is needed. In our work we present a potential solution, by using the machine-translation methods to translate smaller corpora to Slovene and use it to fine-tune and evaluate the results.

3. Dataset description and methodology

Stanford Question Answering Dataset (SQuAD 2.0) (Rajpurkar et al., 2018) is a reading comprehension dataset. It is based on a set of articles on Wikipedia which cover a variety of topics, from historical, pharmaceutical, and religious texts to texts about the European Union. Every question in the dataset is a segment of text or span from the corresponding reading passage. It consists of over 100,000 question-answer pairs extracted from over 500 articles.

The reason to use Squad 2.0 over 1.0 is that it consists of twice as much data and contains unanswerable questions.

3.1. Machine Translation

To translate the dataset into Slovenian we used the eTranslation webservice (Commission, 2020). Due to the web service being primarily designed to translate webpages and short documents in docx or pdf format, our translation pipeline design was as follows:

1. Convert the corpus in html format.
2. Split html file into smaller chunks. We found that 4 MB chunks work best, as larger chunks were often unable to be translated.
3. Send chunks to the translation service.
4. Use the original corpus file to compose the translated document in the original format.

Since the basic translation yielded quite underwhelming results, we employed two different methods to improve the results. The first was to correct the answers by breaking down both the answer and the context into lemmas and search for the answer sequence of lemmas in context sequence of lemmas. To accomplish this, CLASSLA (CLARIN Knowledge Centre for South Slavic languages) library (Ljubešić and Dobrovoljc, 2019) was used. If a match was found, we replaced the bad answer with the original text, forming the lemma sequence in the context. The second method was to embed the answers in the context before translation.

To evaluate the quality of different translations, we measured how many answers can be directly found in their respective context, as they cannot be used in QA models otherwise. The results can be seen in Table 1. Resulting number of valid questions, compared with the original, are presented in Table 2.

Basic	LC	CE	LC+CE
44%	66%	93%	94%

Table 1: Results for basic translation, lemma correction (LC), and context embedded (CE) translation of SQuAD 2.0 dataset. The percentages represent the number of answers that can be directly found in the respective context.

Dataset	Subset	AQ	IQ	Total
Original	Train	86,821	43,498	130,319
	Test	5,928	5,945	11,873
Machine Trans.	Train	81,884	43,498	125,382
	Test	5,735	5,945	11,680

Table 2: Number of questions in original SQuAD 2.0 dataset and our machine translated dataset. *AQ* denotes the number of answerable questions, *IQ* the number of impossible questions.

3.2. Post-editing of Machine Translation

Due to limited human resources, post-editing was done on small number of automatically translated excerpts that were chosen randomly. The provided excerpts included original paragraphs or contexts, questions and answers, as well as their machine translations, which were to be corrected by a translation student. This was done in two steps: creating a project in the online translation tool Memsource with translation memory in tmx format, generated from machine translations, and revision or post-editing of the segments. Editing was first done on the paragraphs and then on questions and answers, since the answers had to match the text in the paragraph. The editing was minimal, which means that the focus was not on stylistic improvement, but mostly on correcting the grammatical errors, wrong meanings and very unusual syntax, to make the translation comprehensible. As mentioned above, the topics of original texts are diverse and very technical, covering different domains such as religion, history, politics, mathematics and chemistry.

In total, there were 30 manually corrected contexts with accompanying 142 answerable and 143 unanswerable questions. The number of different segment types and of post-editing changes can be seen in Table 3.

Segment content	**S**	**NS**	**CS**	**FS**
Context	30	0	30	100%
Answerable question	142	38	104	73.2%
Answer	435	225	210	48.3%
Impossible question	143	43	100	69.9%
Total number	750	306	444	59.2%

Table 3: Post-editing numerical data. *S* denotes the number of segments, *NS* the number of non-corrected segments, *CS* the number of corrected segments and *FS* the fraction of corrected segments.

3.3. Post-editing Analysis

The numbers seen in Table 3 are not fully representative, since some corrections of the mistakes of machine translation are more severe than others and in some segments, there is a much greater number of corrections than in others. For instance, the corrections, including one of a severe semantic mistake, can be seen in this example:

1. Original: The Northern Chinese were ranked higher and Southern Chinese were ranked lower because southern China withstood and fought to the last before caving in.
2. Machine translation: Severna Kitajci so bili uvrščeni višje in južna Kitajci so bili uvrščeni nižje, ker je južna Kitajska zdržala in se borila do zadnjega pred jamarstvom.
3. Post-edited machine translation: Severni Kitajci so bili uvrščeni višje in južni Kitajci so bili uvrščeni nižje, ker se je južna Kitajska pred predajo upirala in se borila do zadnjega.

Answerable and impossible questions have a similar percentage of segments with corrections. This percentage is quite high because machine translation provided incoherent results. In this segments, the changes in post-editing are also more notable, because they affect the overall understanding for potential readers. This can be seen in the following examples:

Original

1. Who did Kublai make the ruler of Korea?
2. Who was Al-Banna's assassination a retaliation for the prior assassination of?
3. What plants create most electric power?

Machine translation

1. Kdo je Kublai postal vladar Koreje?
2. Kdo je bil Al-Bannin umor maščevanja zaradi predhodnega umora?
3. Katere rastline ustvarjajo največ električne energije?

Post-edited machine translation

1. Koga je Kublajkan nastavil za vladarja Koreje?
2. Al-Bannov umor je bil maščevanje za čigav predhodni umor?
3. Katere naprave ustvarjajo največ električne energije?

The segments with answers have the largest number of non-corrected segments because they are shorter. Nevertheless, the percentage of corrected questions is still high if we take into account that the answers represent 58% of all segments. The mistakes in the answers were in the most part already corrected in the contexts. More severe mistakes include semantic mistakes (e.g. plants translated as 'rastline', not 'naprave') and completely wrong answers (e.g. empty segment instead of 'Fermilab' or 'in' instead of '1,388'). Some frequent mistakes also occured in translations of the names of movements, books, projects or other names (e.g. 'Bricks for Varšava' was left untranslated and was changed to 'Zidaki za Varšavo'). There were some punctuation errors, but the most interesting are grammatical mistakes, especially when the wrong grammatical case, gender or number is used. Even if these mistakes were corrected in the context, the answers had to be in the exact same form, so many answers do not sound coherent, which is of course not the case for English, where the conjugation does not change the words as much (e.g. 'Which part of China had people ranked higher in the class system?' — 'Northern' — 'V katerem delu Kitajske so bili ljudje višje v razrednem sistemu?' — 'Severni' (from the example of a sentence in the context mentioned above)). On the other part, some corrected segments were identical even though the source was different due to the use of articles in English language (e.g. 'North Sea' and 'the North Sea' were both translated as 'Severno morje').

It should also be noted that the database SQuAD 2.0 is not entirely reliable. From the batch of randomly sampled 142 test question and answer groups, there were 14 occurrences where at least one of the given answers was not correct (e.g. 'Advanced Steam movement' instead of 'pollution' as an answer to 'Along with fuel sources, what concern has contributed to the development of the Advanced Steam movement?').

4. Models

In this section we present each of the five models that were used in the evaluation.

4.1. XLM-R

XLM-R (XLM-RoBERTa) (Conneau et al., 2019) is a pre-trained cross-lingual language model based on xlm (Lample and Conneau, 2019). The 'RoBERTa' part of the name comes from its training routine that is the same as the monolingual RoBERTa model, specifically, that the sole training objective is the MLM (masked language mode). There is no next sentence prediction (as in BERT) or Sentence Order Prediction (as in ALBERT). XLM-R shows the possibility of training one model for many languages while not sacrificing per-language performance. It is trained on 2.5 TB of CommonCrawl data in 100 languages.

4.2. M-BERT

M-BERT (Multilingual Bert) (Devlin et al., 2018) is a pre-trained cross-lingual language model as it's name suggest. It is based on BERT (Devlin et al., 2018). The pre-trained model is trained on 104 languages with large amount of data from Wikipedia, using a masked language modeling (MLM) objective. On Hugging Face, there is only a base model with 12 hidden transformer layers available, large model with 24 hidden transformer layers was not uploaded and we were not able to test it.

4.3. RemBERT

RemBERT (Chung et al., 2020) is a model, pre-trained on 110 languages, using a masked language modeling (MLM) objective. It's difference with mBERT is that the input and output embeddings are not tied. Instead, RemBERT uses small input embeddings and larger output embeddings. This makes the model more efficient since the output embeddings are discarded during fine-tuning.

4.4. SloBERTa

SloBERTa (Ulčar and Robnik-Šikonja, 2021) is a Slovene monolingual large pre-trained masked language model. It is closely related to French Camembert model, which is similar to base RoBERTa model, but uses a different tokenization model. Since the model requires a large dataset for training, it was trained on 5 combined datasets. It outperformed existing Slovene models.

4.5. CroSloEngual BERT

It is a trilingual model based on BERT and trained for Slovene, Croatian and English language. It was trained with 5.9 billion tokens from these languages. For those languages it performs better than multilingual BERT, which is expected, since studies showed that monolingual models perform better than large multilingual models (Virtanen et al., 2019).

5. Results

This section is divided into two parts. First we evaluate automatic machine translations and then we evaluate performance of choosen QA models (XLM-R-large, M-Bert-base, CroSloEngual BERT, RemBERT, SloBERTa 2.0). All tests were performed on i5 10400f system with RTX 3070 GPU 8 GB VRAM. For larger models we used RTX 3060 12 GB.

To compare the performance between the English, machine translated Slovene and human translated Slovene versions of the SQuAD 2.0 dataset, we used 5 different question answering models: mBERT, XLM-R, RemBERT, SloBERTa 2.0, CroSloEngual BERT. The evaluation was done in three steps:

1. Performance evaluation of different models and fine-tuning configuration on the English dataset, as a benchmark for the evaluation of the Slovene results.
2. Performance evaluation of different models and fine-tuning configuration on the Slovene dataset, translated using computer only, to evaluate the quality of machine translation.
3. Performance evaluation of different models and fine-tuning configuration on the Slovene subset which was translated by a human, and same subset both in English and translated using computer, to evaluate the benefits of human translation.

Before the evaluation, we removed all punctuation, leading and trailing white spaces and articles from both ground truth and prediction. Both of them were also set in the lower case. Parameters used for fine-tuning are presented in Table 4.

Metrics used for the evaluation match the official ones for SQuAD2.0 evaluation and were as follows:

- **Exact** - The fraction of predictions matched at least of one the correct answers exactly.
- **F1** - The average overlap between prediction and ground truth, defined as an average of F1 scores for individual questions. F1 score of an individual question is computed as a harmonic mean of the precision and recall, where precision was defined as $\frac{T_M}{T_P}$, and recall as $\frac{T_M}{T_{GT}}$, where T_M represents the matching tokens between prediction and ground truth, T_P number of tokens in prediction and T_{GT} number of tokens in ground truth. A token is defined as a word, separated by a white space.

The results of the non-translated SQuAD 2.0 and machine translated dataset can be seen in Table 5. The results of the human translated subset and its English and computer translated counterparts can be seen in Table 6. Additionally, we provide some examples of correct predictions with wrong answers in Table 7 and some of correct answers with wrong predictions in Table 8.

Model Name	B	MS	LR	E
XLM-R-large	4	256	1e-5	3
M-BERT-base	8	320	3e-5	3
CroSloEngual BERT	4	256	1e-5	3
RemBERT	4	256	1e-5	3
SloBERTa 2.0	16	320	3e-5	3

Table 4: Parameters used to fine-tune the evaluated models. *B* denotes the number of batches used during fine-tuning, *MS* the maximum sequence length, *LR* the learning rate and *E* the number of epochs.

Model name	Fine-Tuning Language	Original		Machine Translation	
		Exact	F1	Exact	F1
xlmR-large	Eng	81.8%	84.9%	64.3%	72.3%
xlmR-large	Slo	75.0%	79.2%	65.3%	72.4%
xlmR-large	Eng & Slo	74.4%	78.5%	65.9%	73.4%
M-BERT-base	Eng	75.6%	78.9%	55.4%	61.3%
M-BERT-base	Slo	62.4%	67.2%	60.4%	67.0%
M-BERT-base	Eng & Slo	70.7%	75.0%	60.5%	67.3%
CroSloEngual BERT	Eng	72.8%	76.3%	56.3%	63.6%
CroSloEngual BERT	Slo	63.6%	68.2%	58.4%	65.4%
CroSloEngual BERT	Eng & Slo	68.8%	73.0%	58.1%	65.7%
RemBERT	Eng	**84.5%**	**87.5%**	**67.1%**	**73.8%**
SloBERTa 2.0	Slo	60.6%	64.7%	**66.7%**	**73.9%**

Table 5: Comparison of the results of various models and their fine-tuning configurations on the English SQuAD 2.0 evaluation dataset and Slovene machine translated SQuAD 2.0 evaluation dataset. The English dataset only contains the questions preset in its Slovene counterpart. Specific parameters used in fine-tuning are presented in Table 4.

Model name	Fine-Tuning Language	Original		Machine Translation		Human Translation	
		Exact	F1	Exact	F1	Exact	F1
xlmR-large	Eng	80.0%	82.9%	61.1%	68.5%	71.6%	75.9%
xlmR-large	Slo	69.1%	72.9%	61.4%	69.1%	69.8%	74.8%
xlmR-large	Eng & Slo	68.8%	73.4%	64.6%	72.4%	70.5%	75.7%
M-BERT-base	Eng	71.9%	74.9%	52.6%	57.7%	57.5%	60.3%
M-BERT-base	Slo	56.1%	60.4%	58.6%	64.5%	60.4%	66.2%
M-BERT-base	Eng & Slo	64.9%	68.8%	55.8%	61.2%	63.5%	68.6%
CroSloEngual BERT	Eng	73.3%	75.5%	53.0%	60.8%	62.1%	65.7%
CroSloEngual BERT	Slo	59.6%	63.1%	51.6%	58.8%	60.7%	66.0%
CroSloEngual BERT	Eng & Slo	68.1%	70.6%	58.9%	66.3%	64.6%	71.0%
RemBERT	Eng	**84.9%**	**87.2%**	64.2%	71.4%	71.9%	76.9%
SloBERTa 2.0	Slo	59.3%	65.0%	**64.9%**	**72.2%**	**72.6%**	**78.0%**

Table 6: Comparison of the results of various models and their fine-tuning configurations on the Human Translated subset of SQuAD 2.0, and the subsets containing same question from original English dataset and the machine translated dataset. Specific parameters used in fine-tuning are presented in Table 4.

#	Dataset	Question	Answer	Prediction
1	ENG	How many of Warsaw's inhabitants spoke Polish in 1933?	833,500	833,500
	MT	Koliko prebivalcev Varšave je leta 1933 govorilo poljsko?	prebivalcev	833.500
	HT	Koliko prebivalcev Varšave je leta 1933 govorilo poljski jezik?	833.500	833.500
2	ENG	Who recorded "Walking in Fresno?"	Bob Gallion	Bob Gallion
	MT	Kdo je posnel „Walking in Fresno?“	je Bob	Bob Gallion
	HT	Kdo je posnel ≫Walking in Fresno≪?	Bob Gallion	Bob Gallion

Table 7: Examples of correct predictions with wrong answers. *ENG* denotes the English dataset, *MT* one translated by a computer and *HT* one translated by a human.

#	Dataset	Question	Answer	Prediction
1	ENG	Where did Korea border Kublai's territory?	northeast	northeast
	MT	Kje je Koreja mejila na Kublajevo ozemlje?	severovzhodno	zahodno
	HT	Kje je Koreja mejila na Kublajkanovo ozemlje?	severovzhodno	severovzhodno
2	ENG	How many miles, once completed, will the the Lewis S. Eaton trail cover?	22	22
	MT	Koliko kilometrov, ko bo končano, bo pokrivalo Lewis S. Eaton?	22	35
	HT	Koliko kilometrov bo dolga pot Lewisa S. Eatona, ko bo končana?	22	35

Table 8: Examples of correct answers with wrong predictions. *ENG* denotes the English dataset, *MT* one translated by a computer and *HT* one translated by a human.

6. Discussion

6.1. Quantitative Analysis

From the results in Table 5, we can see that RemBERT and SloBERTa 2.0 gave the best results on the dataset translated by a computer. While the result for SloBERTa was expected, as monolingual models tend to perform better than multilingual ones, RemBERT managed to outperform its multilingual competitors while only being fine-tuned on the English dataset. We would attribute this simply to the better design of the model. Although both models had a very similar performance, we would like to point out that RemBERT model is a much larger model and was pre-trained on a significantly larger dataset. Similar results were also observed when comparing the results on the smaller subset of questions that were translated by a human, as seen in Table 6.

In Table 6 we can see models consistently performing better on the human translated data, suggesting that the machine translation provided by eTranslation webservice comes short of providing adequate set for proper evaluation in the Slovene language. We can also see that while the models fine-tuned using machine translated dataset do perform better when evaluated on the machine translated data, this does not hold true for evaluations on human translated data.

We have also observed that fine-tuning the model on the English dataset first, and then on the Slovene, yields better results on the smaller models, M-BERT-base and CroSloEngular BERT, as compared to fine-tuning on either language.

6.2. Qualitative Analysis

While there are many correct predictions of the answers in the machine translated dataset, it is clear that a great number of predictions still does not answer the question correctly. This is because the machine translation of the sentences in the context is not grammatically and stylistically correct, does not convey the right meaning and thus the model has more problems finding the answer. The correct predictions are mostly the ones where the answer to the question is short and the words are not conjugated, i.e. numbers and names, even though there are some exceptions. The same is true for human post-edited translation, but improvement of some answers is already visible from only a few representative examples in Table 7 and Table 8.

7. Conclusion

In this work we present a machine translated SQuAD 2.0 dataset and evaluate it on the following question answering (QA) models: XLM-R-large, M-BERT-base, RemBERT, CroSloEngual BERT and SloBERTa 2.0. Additionally, we also perform human post-editing on a subset of SQuAD 2.0 translations in order to better ascertain the quality of machine translations. The results show that using machine translated data for evaluation led to notably worse results as compared to the one translated by a human. Moreover, we noticed that while multilingual models fine-tuned using machine translated data performed better than ones fine-tuned on English data when given a task of answering the machine translated question, the situation was in most cases reversed when given a task of answering human translated questions. This leads us to conclude that machine translation, at least one available on via eTranslation (Commission, 2020) service, is not particularly suitable for training multilingual models. Of all the models, SloBERTa 2.0 produced the best results on both machine and human translated data, while the RemBERT gave comparable results even when only fine-tuned on the English dataset.

The testing procedure could be easily improved by employing stronger hardware. RemBERT could for example be fine-tuned on the Slovene dataset, which would allow for its better evaluation. Additionally, we were unable to ascertain the optimal parameters for fine-tuning as performing multiple fine-tunings for each language would be unfeasible. Some restrictions of the project are limited time for post-editing and only one translator who is not an expert in the topics of various technical texts, and the method of minimal editing that can result in mediocre translation. The experiment could be expanded by including a larger subset of human translated or revised data, more datasets, such as Natural Questions (Kwiatkowski et al., 2019), and different machine translation services, such as DeepL.

8. Acknowledgments

We would like to to thank our mentors, Slavko Žitnik and Špela Vintar, for providing us with directions, feedback and advice.

9. References

Ines Čeh and Milan Ojsteršek. 2009. Developing a question answering system for the Slovene language. *WSEAS Transaction on Information science and applications*, (9).

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *CoRR*, abs/2010.12821.

European Commission. 2020. CEF Digital eTranslation. `https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Unsupervised cross-lingual representation learning at scale. arXiv:1911.02116.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In: *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

Nikola Ljubešić and Kaja Dobrovoljc. 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*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In: *Proceedings of the 58th Annual Meeting of the Association for Computa- tional Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest BERT and CroSloEngual BERT. In: *International Conference on Text, Speech, and Dialogue*, pages 104–111. Springer.

Matej Ulčar and Marko Robnik-Šikonja. 2021. SloBERTa: Slovene monolingual large pretrained masked language model.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Fnnish. *arXiv:1912.07076*.

William A Woods and WOODS WA. 1977. Lunar rocks in natural english: Explorations in natural language question answering.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pretrained text-to-text transformer. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.