Zbornik konference

# Jezikovne tehnologije in digitalna humanistika

*Proceedings of the Conference on*

# *Language Technologies and Digital Humanities*

**24.– 25. september 2020**

**Ljubljana, Slovenija**

**September 24th – 25th 2020**

**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
v. d. direktorja / *Acting Director*

Ljubljana, 2020
First edition

**Spletno mesto konference / Conference website:**
http://www.sdjt.si/jtdh-2020 / http://www.sdjt.si/jtdh-2020/en/

**Publikacija je brezplačno dostopna na: / *Publication is available free of charge at*:**
http://nl.ijs.si/jtdh20/proceedings-sl.html / *http://nl.ijs.si/jtdh20/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=30096131

ISBN 978-961-7104-01-1 (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 (CJVT) ter raziskovalnima infrastrukturama CLARIN.SI in DARIAH-SI, že dvanajstič po vrsti prireja konferenco "Jezikovne tehnologije in digitalna humanistika", ki po uspešni programski širitvi na digitalno humanistiko leta 2016 ohranja povezovalni fokus med disciplinama, hkrati pa si prizadeva postati pomembno srečevališče raziskovalcev v regiji.

Zaradi pandemije covid-19 je letošnja konferenca potekala virtualno, izvedbo v tej obliki pa je podprl CLARIN ERIC. Ker smo želeli zagotoviti, da bi bila konferenca v čim večji meri dostopna vsem zainteresiranim, je bila udeležba za vse brezplačna, program pa smo sestavili iz asinhronih in sinhronih elementov. Zbornik konference in posnetki predstavitev sprejetih prispevkov so bili objavljeni vnaprej, tako, da si jih udeleženci lahko ogledajo, ko jim to najbolj ustreza. Med konferenco smo v konferenčnem okolju Zoom organizirali moderirane tematske diskusije v živo, ki smo jih uvrstili v štiri tematske sklope: Jezikovne in govorne tehnologije, Jezikovni viri, Korpusna analiza ter Digitalna humanistika in pedagogika. Poleg rednega programa smo priredili tudi panel Razvoj slovenščine v digitalnem okolju. Ob zaključku konference smo nagradili najboljši študentski prispevek, 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 21 rednih in 5 študentskih prispevkov. Vse prispevke so pregledali trije recenzenti. 15 prispevkov je napisanih v slovenskem, 11 pa v angleškem jeziku. Skupno število vseh avtorjev prispevkov je 73, od katerih je tretjina tujih (iz Nemčije, Avstrije, Švice in Češke).

Urednika se iz srca zahvaljujeva vsem, ki so prispevali k uspehu konference, ki je zaradi virtualnega formata od vseh zahtevala številne prilagoditve in dodatno predanost: vabljenima predavateljema in avtorjem prispevkov za skrbno pripravljene prispevke in posnetke predstavitev, programskemu odboru za natančno recenzentsko delo, organizacijskemu odboru za izvedbo konference v virtualni obliki, moderatorjem diskusij, da so kljub fizični oddaljenosti med udeleženci spodbudili živahno izmenjavo mnenj, tehničnim urednikom za pripravo spletnega zbornika in CLARIN ERIC-u za podporo s konferenčnim sistemom.

Ljubljana, september 2020

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 12th 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.

Due to the Covid-19 pandemic this year's conference was made virtual with the support of CLARIN ERIC. In order to make the conference as accessible as possible to all participants, conference participation was free of charge, while the program had an asynchronous and a synchronous component. Conference proceedings and recordings of the presentations were published on the conference website in advance so that everyone could view them at their own convenience. Four live moderated thematic discussion sessions were organised: Language and speech technologies, Language resources, Corpus analysis, and Digital humanities and pedagogy. In addition to the scientific programme, we also hosted a panel on the development of Slovene in the digital environment. At the end of the conference, the best student paper was awarded, after which the conference concluded with the annual general meeting of the Slovenian Language Technologies Society.

This year's conference saw presentations from two invited speakers and from authors of 21 regular and 5 student papers. All the papers were reviewed by three reviewers. 15 papers are written in Slovene, and 11 in English. The total number of all authors of the accepted papers is 73, a third of which were from abroad (from Germany, Austria, Switzerland and the Czech Republic).

The editors would like to thank everyone who has contributed to the success of this conference, which, because of its virtual format, required extra flexibility and dedication: 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 for successfully bridging the physical distance between the participants and engaging discussions, the technical editors for preparing the online proceedings and CLARIN ERIC for their support with the conference platform.

Ljubljana, September 2020

Darja Fišer and Tomaž Erjavec

# Programski odbor / *Programme committee*

## Predsedstvo programskega odbora / *Steering committee*

**Darja Fišer,** predsednica / *Chair*
Filozofska fakulteta, Univerza v Ljubljani / *Faculty of Arts, University of Ljubljana*

**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*

**Ajda Pretnar**
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*

**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*

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

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

**Darja Fišer**
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*

**Mario Hibert**
Filozofska fakulteta, Univerza v Sarajevu / *Faculty of Arts, University of Sarajevo*

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

**Ivo Ipšić**
Tehniška fakulteta, Univerza na Reki / *Faculty of Engineering, University of Rijeka*

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

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

**Mojca Kotar**
Univerza v Ljubljani / *University of Ljubljana*

**Simon Krek**
Institut "Jožef Stefan" in Center za jezikovne vire in tehnologije, Univerza v Ljubljani / *Jožef Stefan Institute and Centre for Language Resources and Technologies, University of Ljubljana*

**Cvetana Krstev**
Filozofska fakulteta, Univerza v Beogradu / *Faculty of Philology, University of Belgrade*

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

**Nikola Ljubešić**
Odsek za informacijske in komunikacijske znanosti, Univerza v Zagrebu / *Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb*

**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*

**France Mihelič**
Fakulteta za elektrotehniko, Univerza v Ljubljani / *Faculty of Electrical Engineering, University of Ljubljana*

**Maja Miličević**
Filološka fakulteta, Univerza v Beogradu / *Faculty of Philology, University of Belgrade*

**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*

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

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

**Senja Pollak**
Institut "Jožef Stefan" / *Jožef Stefan Institute*

**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*

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

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

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

**Aleš Vaupotič**
Raziskovalni center za humanistiko, Univerza v Novi Gorici / *Research Centre for Humanities, University of Nova Gorica*

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

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

## Organizacijski odbor / *Organising committee*

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

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

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

**Darja Fišer**
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*

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

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

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

**David Bordon**, tehnična podpora s platformo Zoom / *Technical assistance with the Zoom platform*
Filozofska fakulteta, Univerza v Ljubljani / *Faculty of Arts, University of Ljubljana*

## Organizatorji / *Organizers*

# Urnik / *Timetable*

## Četrtek / Thursday, 24. 9. 2020

| | |
|---|---|
| **13:00-13.30** | **Vabljeno predavanje 1 / Keynote 1:**<br>Vodja sekcije / Session Chair: Filip Dobranić<br><br>**Sara Tonelli**<br>***Avtomatsko prepoznavanje nasilnega govora: preveč digitalne, premalo humanistike?***<br>***Abusive Language Detection: Too much Digital, not enough Humanities?*** |
| **13:30-14.30** | **Sekcija 1 / Session 1:**<br>**Jezikovne in govorne tehnologije / NLP and Speech Technologies**<br>Vodja sekcije / Session Chair: Tanja Samardžić<br><br>**Dolge predstavitve / Long presentations:**<br><br>Anka Supej, Matej Ulčar, Marko Robnik Šikonja, Senja Pollak:<br>*Dimenzija spola v slovenskih vektorskih vložitvah besed: primerjava modelov prek analogij poklicev*<br><br>Andrej Žgank, Mirjam S. Maučec, Gregor Donaj:<br>*Razpoznavalnik tekočega govora UMB Broadcast News 2020: prehod na arhitekturo z nevronskimi mrežami*<br><br>**Kratke predstavitve / Short presentations:**<br><br>Marko Pranjić, Vid Podpečan, Marko Robnik-Šikonja, Senja Pollak:<br>*Evaluation of related news recommendations using document similarity methods*<br><br>Maja Popović, Alberto Poncelas:<br>*Extracting correctly aligned segments from unclean parallel data using character n-gram matching*<br><br>Marko Robnik-Šikonja, Kristjan Reba, Igor Mozetič:<br>*Cross-lingual Transfer of Twitter Sentiment Models Using a Common Vector Space*<br><br>Martin Pernuš, Simon Dobrišek:<br>*Načrtovanje jezikovne komponente generativnega nasprotniškega nevronskega omrežja*<br><br>Darinka Verdonik, Matej Rojc, Izidor Mlakar, Danilo Zimšek, Zdravko Kačič, Albin Bregant, Marko Ferme, Milan Ojsteršek, Tomaž Erjavec:<br>*Raziskovalna infrastruktura projekta RI-SI CLARIN*<br><br>Andraž Pelicon:<br>*Zaznavanje sentimenta v novicah z globokimi nevronskimi mrežami* (Študentski prispevek / Student paper) |
| **15:00-16:00** | **Sekcija 2 / Session 2:**<br>**Jezikovni viri / Language resources**<br>Vodja sekcije / Session Chair: Darinka Verdonik<br><br>**Dolge predstavitve / Long presentations:**<br><br>Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Polona Gantar, Špela Arhar Holdt, Jaka |

| | |
|---|---|
| | Čibej, Janez Brank:<br>*The ssj500k Training Corpus for Slovene Language Processing*<br><br>Špela Arhar Holdt, Senja Pollak, Marko Robnik-Šikonja, Simon Krek:<br>*Referenčni seznam pogostih splošnih besed za slovenščino*<br><br>Lucia Vlášková, Hana Strachoňová:<br>*Selected methodological challenges of sign language lexicography*<br><br>Špela Arhar Holdt, Jaka Čibej:<br>*Rezultati projekta "Slovar sopomenk sodobne slovenščine: od skupnosti za skupnost"*<br><br>**Kratke predstavitve / Short presentations:**<br><br>Dolores Lemmenmeier-Batinić, Nikola Ljubešić, Tanja Samardžić:<br>*XML-Encoding of a spoken Serbian corpus targeting forms of address*<br><br>Karoline Irschara, Claudia Posch, Astrid Grams, Leonhard Gruber, Anna-Lena Huber, Stephanie Mangesius, Gerhard Rampl, Birgit Waldner, Bernhard Glodny:<br>*Challenges in Building a Corpus of Unstructured German Radiology Reports*<br><br>Mojca Stritar Kučuk:<br>*Modul Leto plus – prvi korak do korpusa slovenščine kot tujega jezika*<br><br>Špela Antloga:<br>*Korpus metafor KOMET 1.0* (Študentski prispevek / Student paper) |
| **16:00-17:00** | **Panel:**<br>**RSDO - Razvoj slovenščine v digitalnem okolju / RSDO - The Development of Slovene in the Digital Environment**<br>Vodja sekcije / Session Chair: Špela Arhar Holdt |

## Petek / Friday, 25. 9. 2020

| | |
|---|---|
| **13:00-13:30** | **Vabljeno predavanje 2 / Keynote 2:**<br>Vodja sekcije / Session Chair: Dan Podjed<br><br>**Kaspar Beelen**<br>***Govorjenje v imenu drugih - Zakaj naj digitalno humanistiko zanimajo parlamentarni podatki***<br>***Speaking on behalf of others - Why the digital humanities should care about parliamentary data*** |
| **13:30-14:30** | **Sekcija 3 / Session 3:**<br>**Korpusna analiza / Corpus analysis**<br>Vodja sekcije / Session Chair: Nataša Logar<br><br>**Dolge predstavitve / Long presentations:**<br><br>Kristina Pahor de Maiti, Darja Fišer, Nikola Ljubešić, Tomaž Erjavec:<br>*Grammatical Footprint of Socially Unacceptable Facebook Comments*<br><br>Jakob Lenardič, Darja Fišer: |

| | |
|---|---|
| | *Epistemic modal adverbs in Slovenian academic discourse*<br><br>Zoran Fijavž:<br>*Ambivalence of Queer Visibility in Video-Based Social Media Content* (Študentski prispevek / Student paper)<br><br>Eva Trivunović:<br>*Variante in modifikacije (iz)biblijskih frazemov* (Študentski prispevek / Student paper) |
| **15:00-16:00** | **Sekcija 4 / Session 4:**<br>**Digitalna humanistika in pedagogika / Digital Humanities and Pedagogy**<br>Vodja sekcije / Session Chair: Miran Hladnik<br><br>**Dolge predstavitve / Long presentations:**<br><br>Narvika Bovcon, Aleš Vaupotič:<br>*Literarne pešpoti v povečani resničnosti na področju med Goriško, Tržaško pokrajino in Furlanijo: Konceptualna zasnova, izvedba in trajno vzdrževanje*<br><br>Katja Meden, Ana Cvek:<br>*Nadgradnja Zgodovinarskega indeksa citiranosti*<br><br>Darinka Verdonik, Simona Majhenič, Špela Antloga, Sandi Majninger, Marko Ferme, Kaja Dobrovoljc, Simona Pulko, Mira Krajnc Ivič, Natalija Ulčnik:<br>*Uporaba jezikovnotehnoloških virov in postopkov pri razvoju učnega e-okolja Slovenščina na dlani*<br><br>**Kratke predstavitve / Short presentations:**<br><br>Katja Mihurko Poniž, Narvika Bovcon, Aneta Trajanov, Tina Smrekar:<br>*Digitalizacija kulturne dediščine v dveh projektih programa Študentski inovativni projekti za družbeno korist*<br><br>Andrej Pančur, Neja Blaj Hribar, Mihael Ojsteršek, Mojca Šorn:<br>*Projekt Vojaške žrtve prve svetovne vojne na Slovenskem*<br><br>Magdalena Schlintl, Kerstin Pawluch, Mara Rader, Verena Novak-Geiger:<br>*Working with Digital Devices in Teacher Training Using the Example of the Teaching-Learning-Lab* (Študentski prispevek / Student paper) |
| **16:00-16:15** | **Podelitev nagrade najboljšemu študentskemu prispevku in zaključek konference / Best Student Paper Award & Closing**<br>Vodja sekcije / Session Chair: Ajda Pretnar |
| **16:15-17:00** | **Občni zbor SDJT / Annual meeting SDJT**<br>Vodja sekcije / Session Chair: Simon Dobrišek |

# 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*** …… x

**VABLJENI PRISPEVKI / *INVITED TALKS*** ……. 1

**Abusive Language Detection: Too much Digital, not enough Humanities?**

*Sara Tonelli* …… 1

**Speaking on behalf of others: Why the digital humanities should care about parliamentary data**

*Kaspar Beelen* …… 2

**PRISPEVKI – *PAPERS*** ……. 3

**Rezultati projekta "Slovar sopomenk sodobne slovenščine: od skupnosti za skupnost"**

*Špela Arhar Holdt, Jaka Čibej* …… 3

**Referenčni seznam pogostih splošnih besed za slovenščino**

*Špela Arhar Holdt, Senja Pollak, Marko Robnik-Šikonja, Simon Krek* …… 10

**Literarne pešpoti v povečani resničnosti na področju med Goriško, Tržaško pokrajino in Furlanijo: Konceptualna zasnova, izvedba in trajno vzdrževanje**

*Narvika Bovcon, Aleš Vaupotič* …… 16

**The ssj500k Training Corpus for Slovene Language Processing**

*Simon Krek, Tomaž Erjavec, Kaja Dobrovoljc, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Janez Brank* ..... 24

**Epistemic modal adverbs in Slovenian academic discourse**

*Jakob Lenardič, Darja Fišer* ..... 34

**Nadgradnja Zgodovinarskega indeksa citiranosti**

*Katja Meden, Ana Cvek* ..... 42

**Grammatical Footprint of Socially Unacceptable Facebook Comments**

*Kristina Pahor de Maiti, Darja Fišer, Nikola Ljubešić, Tomaž Erjavec* ..... 48

**Načrtovanje jezikovne komponente generativnega nasprotniškega nevronskega omrežja**

*Martin Pernuš, Simon Dobrišek* ..... 58

**Digitalizacija kulturne dediščine v dveh projektih programa Študentski inovativni projekti za družbeno korist**

*Katja Mihurko Poniž, Narvika Bovcon, Aneta Trajanov, Tina Smrekar* ..... 66

**Extracting correctly aligned segments from unclean parallel data using character n-gram matching**

*Maja Popović, Alberto Poncelas* ..... 74

**Evaluation of related news recommendations using document similarity methods**

*Marko Pranjić, Vid Podpečan, Marko Robnik-Šikonja, Senja Pollak* ..... 81

**Cross-lingual Transfer of Twitter Sentiment Models Using a Common Vector Space**

*Marko Robnik-Šikonja, Kristjan Reba, Igor Mozetič* ..... 87

**Primerjava slovenskih besednih vektorskih vložitev z vidika spola na analogijah poklicev**

*Anka Supej, Matej Ulčar, Marko Robnik Šikonja, Senja Pollak* ..... 93

**Uporaba jezikovnotehnoloških virov in postopkov pri razvoju učnega e-okolja Slovenščina na dlani**

*Darinka Verdonik, Simona Majhenič, Špela Antloga, Sandi Majninger, Marko Ferme, Kaja Dobrovoljc, Simona Pulko, Mira Krajnc Ivič, Natalija Ulčnik* …… 101

**Selected methodological challenges of sign language lexicography**

*Lucia Vlášková, Hana Strachoňová* …… 109

**Razpoznavalnik tekočega govora UMB Broadcast News 2020: prehod na arhitekturo z nevronskimi mrežami**

*Andrej Žgank, Mirjam S. Maučec, Gregor Donaj* ……... 117

**RAZŠIRJENI POVZETKI – *EXTENDED ABSTRACTS*** …… 124

**Challenges in Building a Corpus of Unstructured German Radiology Reports**

*Karoline Irschara, Claudia Posch, Astrid Grams, Leonhard Gruber, Anna-Lena Huber, Stephanie Mangesius, Gerhard Rampl, Birgit Waldner, Bernhard Glodny* …… 124

**XML-Encoding of a spoken Serbian corpus targeting forms of address**

*Dolores Lemmenmeier-Batinić, Nikola Ljubešić, Tanja Samardžić* …… 127

**Modul Leto plus – prvi korak do korpusa slovenščine kot tujega jezika**

*Mojca Stritar Kučuk* …… 131

**Projekt Vojaške žrtve prve svetovne vojne na Slovenskem**

*Andrej Pančur, Neja Blaj Hribar, Mihael Ojsteršek, Mojca Šorn* …… 136

**Raziskovalna infrastruktura projekta RI-SI CLARIN**

*Darinka Verdonik, Matej Rojc, Izidor Mlakar, Danilo Zimšek, Zdravko Kačič, Albin Bregant, Marko Ferme, Milan Ojsteršek, Tomaž Erjavec* …… 141

**ŠTUDENTSKI PRISPEVKI – *STUDENT PAPERS*** ..... 144

**Ambivalence of Queer Visibility in Video-Based Social Media Content**

*Zoran Fijavž* ..... 144

**Zaznavanje sentimenta v novicah z globokimi nevronskimi mrežami**

*Andraž Pelicon* ..... 150

**Variante in modifikacije (iz)biblijskih frazemov**

*Eva Trivunović* ..... 158

**ŠTUDENTSKI POVZETKI – *STUDENT ABSTRACTS*** ..... 167

**Korpus metafor KOMET 1.0**

*Špela Antloga* ..... 167

**Working with Digital Devices in Teacher Training Using the Example of the Teaching-Learning-Lab**

*Magdalena Schlintl, Kerstin Pawluch, Mara Rader, Verena Novak-Geiger* ..... 171

# Abusive Language Detection: Too much Digital, not enough Humanities?

**Sara Tonelli**
Fondazione Bruno Kessler
Via Santa Croce, 77
38122 Trento, Italy
satonelli@fbk.eu

## Abstract

Social media messages are often written to attack specific groups of users based on their religion, ethnicity or social status, and they can be particularly threatening to vulnerable users such as teenagers. While it is important to develop tools and algorithms that automatically detect abusive messages, so that they can be deleted and hateful accounts can be blocked, it is also important to understand the shared language used by online communities, to what extent offensiveness can be generalized, and what is the role of dataset annotation in avoiding biases in online abuse detection. This talk gives an overview of current state-of-the-art approaches in abusive language detection and discusses ongoing work to better understand this phenomenon, analysing the role of communities, textual context and stakeholders' involvement.

## Bio

Sara Tonelli is the head of the Digital Humanities research group at Fondazione Bruno Kessler in Trento, and adjunct professor of Language Interfaces at the Dept. of Psychology and Cognitive Science, University of Trento. She got a PhD in Language Sciences from Università Ca' Foscari in Venice in 2010, after which she joined FBK first as a post-doc in the Natural Language Processing group and then as a tenured researcher leading the newly founded DH group. She has been involved in several national and European projects dealing with historical archives, event and temporal information extraction and political stance detection and, more recently, with social media monitoring and hate speech detection. Her research interests are highly interdisciplinary, trying to apply and adapt advanced text analysis approaches to social sciences and historical investigation.

# Speaking on behalf of others
# Why the digital humanities should care about parliamentary data

**Kaspar Beelen**
Alan Turing Institute, London
British Library, 96 Euston Road, London NW1 2DB
kbeelen@turing.ac.uk

## Abstract

Parliaments keep minute records of all the words spoken by politicians during plenary sessions. Over the last decades, these (almost verbatim) proceedings have become increasingly available in digital format, spawning immense historical corpora, which contain millions of words, often spanning multiple centuries of political debate. These data give a unique insight into the language and worldview of members of parliaments, and the constituents they aimed to represent, thus providing a detailed account on almost every issue that moved public opinion at some point in time. Digital parliamentary proceedings constitute an invaluable, but still under-explored collection for digital historians. This talk gives an overview of emerging approaches to parliamentary data that combine computer science and history. It firstly covers how speeches of MPs allow scholars to study different aspects of language, form sentiment, or sophistication, to use of humour. Secondly, it discusses how the computational analysis of parliamentary debates contributes to answering critical historical questions, such as the role of women in national politics or the changing shape of political representation.

## Bio

Kaspar Beelen is a digital historian, who explores the application of machine learning to humanities research. After obtaining his PhD in History (2014) at the University of Antwerp he worked as postdoctoral fellow at the University of Toronto. As researcher on the Digging into Linked Parliamentary Data (Dilipad) project, he published several papers situated at the interface of data science, political science and history, which explored a wide range of topics, including: the representation of women in Westminster, the evolution of public health discourse, and the use of affect in parliamentary language. In 2016, Kaspar moved to the University of Amsterdam where he first worked as a postdoc at the Informatics Institute, and later became assistant professor in Digital Humanities (Media Studies). Since February 2019, he works at the Alan Turing Institute as research associate for the Living with Machines project.

# Rezultati projekta "Slovar sopomenk sodobne slovenščine: od skupnosti za skupnost"

**Špela Arhar Holdt,* Jaka Čibej***

* Center za jezikovne vire in tehnologije, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
spela.arhar@cjvt.si, jaka.cibej@cjvt.si

**Povzetek**

V prispevku predstavljamo rezultate projekta "Slovar sopomenk sodobne slovenščine: od skupnosti za skupnost", ki je izvedel vrsto diseminacijskih aktivnosti z namenom ozaveščanja javnosti o izidu Slovarja sopomenk sodobne slovenščine in o novem konceptu odzivnega slovarja, ki je digitalno zasnovan in v razvoj vključuje tudi uporabniško skupnost. Poleg diseminacije se je projekt osredotočal tudi na razvoj metod za spremljanje eksplicitne in implicitne uporabniške povratne informacije, ki pomaga nadgrajevati odzivne jezikovne vire v smeri večje uporabniške prijaznosti, ter z motivacijo uporabnikov za sodelovanje pri odzivnih slovarjih.

**"The Thesaurus of Modern Slovene: By the Community for the Community": Project Results**

The paper presents the results of a project titled "The Thesaurus of Modern Slovene: By the Community for the Community". The project carried out a range of dissemination activities with the aim of informing the public about the publication of the Thesaurus of Modern Slovene and the new concept of a responsive dictionary, which is digital-born and includes the language user community in its development. Aside from dissemination, the project also focused on developing methods for collecting explicit and implicit user feedback to help upgrade responsive language resources, particularly in terms of user-friendliness, as well as the motivation of users to contribute to the development of responsive dictionaries.

## 1. Uvod

Prispevek opisuje namen in rezultate projekta “Slovar sopomenk sodobne slovenščine: od skupnosti za skupnost“, ki ga je v letih 2018 in 2019 financiralo Ministrstvo za kulturo Republike Slovenije v sklopu Javnega razpisa za (so)financiranje projektov, namenjenih promociji in uveljavljanju slovenskega jezika na področju jezikovnih virov, tehnologij in digitalizacije (JR-Promocija SJ-2018–2019). Projekt v finančnem obsegu 40.000 EUR je prijavila Fakulteta za računalništvo in informatiko Univerze v Ljubljani, izvedba pa je potekala v okviru razvojnih nalog Centra za jezikovne vire in tehnologije UL.

Vsebinsko je bil projekt vezan na Slovar sopomenk sodobne slovenščine, ki je na voljo od pomladi 2018 (https://viri.cjvt.si/sopomenke/). Slovar je v slovaropisni prostor uvedel nov tip jezikovnega vira: odzivni slovar. Slednji je v izhodišču pripravljen z avtomatskimi metodami, njegov nadaljnji razvoj pa poteka stopenjsko, odprto in v sodelovanju z uporabniško skupnostjo.

Primarni namen projekta je bil obvestiti javnost o novem slovarju, prednostih in omejitvah odzivnih slovarjev ter možnostih za uporabniško sodelovanje pri njihovem razvoju, s posebno mislijo na uporabniške skupine, za katere so informacije o sopomenskosti najbolj dragocene. V širšem smislu smo želeli s projektom opozoriti na novosti, ki jih skupnosti prinaša digitalno slovaropisje, na pomen prosto dostopnih in podatkovno odprtih jezikovnih virov, kot tudi vrednost uporabniške povratne informacije za njihov nadaljnji razvoj – ter uporabnike za sodelovanje pri razvoju tudi usposobiti in motivirati.

Pomemben del projekta je bilo zbiranje povratne informacije o novorazvitem odzivnem konceptu. V projektu so bile izvedene različne uporabniške študije: anketa o novostih odzivnega slovarja, uporabniška evalvacija slovarskega vmesnika, analiza avtentičnih jezikovnih težav, ki jih slovar sopomenk lahko naslovi. Razvita je bila metodologija za spremljanje uporabniških aktivnosti med slovarsko rabo in izdelan je bil sopomenski modul za mobilno aplikacijo *Igra besed*, s katero v model odzivnega slovarja vključujemo igrifikacijo.

Rezultati projekta, vključno s projektnimi poročili, so na voljo na spletni strani https://www.cjvt.si/promocija-sopomenk/, niso pa še bili celovito predstavljeni raziskovalni skupnosti. Slovaropisne rešitve, ki omogočajo dostop do odprtih jezikovnih podatkov v čim krajšem času, so ključne za vzpostavitev in vzdrževanje temeljne jezikovne infrastrukture. Odpiranje slovaropisnih delotokov za sodelovanje uporabniške skupnosti je skladno s trendi digitalnega slovaropisja in pričakovanji slovarskih uporabnikov v digitalni dobi. V tem kontekstu je projekt relevanten zlasti zato, ker preizkuša, evalvira in nadgrajuje koncept odzivnega slovarja, ki je možni odgovor na izpostavljene izzive in bo v nadaljevanju apliciran tudi na druge temeljne jezikovne vire za slovenščino.[1]

## 2. Slovar sopomenk sodobne slovenščine kot odzivni slovar

Slovar sopomenk sodobne slovenščine je prvi primer odzivnega slovarja, ki se od drugih jezikovnih virov v določenih vidikih razlikuje in na slovaropisno področje vnaša nekatere novosti. Kot prvo, slovar je povsem digitalen in zasnovan izključno za digitalni medij, kar pomeni dvoje: a) zasnovan ni le kot slovar, temveč tudi kot odprto dostopna (slovarska) baza podatkov, ki jo je mogoče uporabiti za razvoj drugih jezikovnih virov in orodij za

[1] V dveh letih od objave Slovarja sopomenke sodobne slovenščine sta po tem modelu že nastala Kolokacijski slovar sodobne slovenščine (opisano v Kosem et al., 2018, dostopno prek vmesnika na https://viri.cjvt.si/kolokacije/slv/, slovarska baza na CLARIN.SI: http://hdl.handle.net/11356/1250 (Kosem et al., 2019)) in oblikoslovni leksikon Sloleks 2.0 (dostopno prek vmesnika na https://viri.cjvt.si/sloleks/slv/, slovarska baza na repozitoriju CLARIN.SI: http://hdl.handle.net/11356/1230 (Dobrovoljc et al., 2019)), v načrtu pa so tudi novi viri.

obdelavo naravnega jezika; in b) uporabniku je na voljo preko spletnega vmesnika, ki je bil razvit tako za računalnike kot za mobilne naprave.

Kot drugo, slovarska baza je v začetku zgrajena avtomatsko – v primeru Slovarja sopomenk sodobne slovenščine je bila baza zgrajena (Krek et al., 2017) na osnovi baze Velikega angleško-slovenskega slovarja Oxford DZS® (Šorli et al., 2006) in korpusa Gigafida 1.0 (Logar et al., 2012). Gradnja je temeljila na vzpostavljanju povezav med prevodnimi ustreznicami v slovenskem delu Velikega angleško-slovenskega slovarja Oxford DZS® in na računanju relevantnosti teh povezav z algoritmom Personal PageRank (Page et al., 1999). Tako izračunane povezave so bile podlaga za sopojavitvene grafe, ki so služili kot osnova za iztočnice in nabor njihovih sopomenk v Slovarju sopomenk, sopomenke pa so bile naknadno razdeljene še na bolj relevantne (*jedrne sopomenke*, ang. *core synonyms*) in manj relevantne (*bližnje sopomenke*, ang. *near synonyms*) – v vmesniku sta kategoriji prikazani v ločenih razdelkih. Povezave med besedami so bile dodatno uravnotežene s korpusom Gigafida 1.0 in preverjene s Slovarjem slovenskega knjižnega jezika.

Kot tretje, odzivni slovar je objavljen relativno hitro po avtomatski izgradnji. Ustreznost rezultatov avtomatske gradnje z evalvacijo najprej potrdijo jezikoslovci, slovar pa je nato objavljen, čeprav baza kot avtomatsko generiran vir vsebuje določeno mero šuma. Iz tega razloga – in v skladu s sodobnimi trendi v slovaropisju, ki vse bolj vključuje množičenje v gradnjo jezikovnih virov (Čibej et al., 2015; Fišer in Čibej, 2015) – je v razvoj odzivnega slovarja vključena tudi uporabniška skupnost. Pri Slovarju sopomenk lahko uporabniki k izboljšavi slovarja pripomorejo na dva načina: a) s pomočjo razdelka v vmesniku lahko k iztočnici dodajajo manjkajoče sopomenke; b) kot ustrezne ali neustrezne lahko ocenjujejo obstoječe sopomenske kandidate v slovarju (tako uporabniško dodane kot tiste, ki so bile v slovar vključene z avtomatskim postopkom). Uporabniške sopomenke in glasovi se takoj pojavijo v vmesniku, tako pridobljeni podatki pa so upoštevani pri nadgradnjah slovarja. Vse uporabniške aktivnosti v slovarju so tudi sistematično in dosledno beležene v dnevnikih iskanja (Arhar Holdt et al., 2018: 407–408), kar omogoča tudi nadzor nad nekonstruktivnimi vnosi.

Odzivni slovar je odprto dostopen, se redno posodablja in je objavljen v različicah. V vmesniku Slovarja sopomenk je datum zadnje posodobitve izpisan ob iztočnici kot časovni žig, kar omogoča sledljivost. Vse različice slovarja so (in bodo) dostopne pod licenco Creative Commons 4.0 BY-SA in arhivirane v repozitoriju CLARIN.SI (npr. Krek et al., 2018 za Sopomenke 1.0), zato ima skupnost vedno možnost dostopati tudi do podatkov iz prejšnjih različic slovarja. Razvoj odzivnega slovarja tako nikoli ni končan – baza se postopoma veča oz. prečiščuje, slovar pa se z upoštevanjem uporabniškega doprinosa odziva tako na mnenja uporabniške skupnosti kot na spremembe v jeziku in se spreminja v skladu z jezikovno rabo – od tod tudi poimenovanje odzivni slovar.

## 3. Diseminacija slovarja v širši javnosti

Ena od aktivnosti v začetnem delu projekta je bilo obveščanje širše javnosti o izidu Slovarja sopomenk sodobne slovenščine, o novostih, ki jih prinaša odzivni slovar kot nov tip slovarja, posebej pa so bili uporabniki pozvani tudi k sodelovanju pri nadaljnjem razvoju slovarja. Slovar je digitalen in predvideva sodelovanje skupnosti v digitalnem okolju, zato je velik del diseminacije potekal v digitalnem okolju. Ob tem smo poskrbeli tudi za diseminacijo med potencialnimi uporabniki, ki ostajajo zunaj digitalnega promocijskega dometa. Veliko pozornosti je bilo usmerjene v vzpostavljanje dialoga z uporabniki in zbiranje njihove povratne informacije. Diseminacija je potekala v štirih glavnih sklopih: obveščanje po e-pošti in po družbenih omrežjih, priprava promocijskih tiskovin, izvedba diseminacijskih dogodkov in anketa o novostih odzivnega slovarja.

### 3.1. Obveščanje po e-pošti in družbenih omrežjih

Za obveščanje po e-pošti je bila pripravljena obsežna adrema, ki je upoštevala tipologijo uporabniških skupin na področju slovarske rabe (Arhar Holdt, 2015) – vsebovala je vse najrelevantnejše skupine uporabnikov, v okviru katerih smo nato ciljno iskali naslovnike, za katere smo ocenili, da bi jih lahko zaradi poklicnih nalog, izobraževalnih potreb ali prostočasnih interesov Slovar sopomenk še posebej zanimal, npr. prevajalce, lektorje ter zaposlene v razvojnem sektorju, kulturi, izobraževanju in jezikovnih agencijah. Osredotočili smo se tudi na osnovne in srednje šole, fakultete in študentske klube ter različne skupine, kulturna in literarna društva, ki se z jezikom ukvarjajo ljubiteljsko. Slovar je bil diseminiran tudi s pomočjo strokovnih e-poštnih seznamov in skupin na družbenih omrežij, kjer se skupnost organizira okrog jezikovnih vprašanj.

E-poštna sporočila o novem slovarju sopomenk so bila poslana skupno na 1.545 e-poštnih naslovov, od tega 661 osnovnim in srednjim šolam, 223 visokošolskim in univerzitetnim profesorjem, 58 študentskim klubom, 70 centrom, zavodom in agencijam, 67 skupinam Slovencev v tujini, 252 na različne naslove v povezavi s slovensko prevajalsko skupnostjo, 64 kulturnim in literarnim društvom in 150 knjižnicam.

### 3.2. Promocijske tiskovine

Za obveščanje v nedigitalnem svetu so bile pripravljene promocijske tiskovine v sodelovanju z Znanstveno založbo Filozofske fakultete Univerze v Ljubljani za sejem akademske knjige Liber.Ac 2018. Vsa nastala gradiva (plakat, zloženka, knjižna kazala) so dostopna na projektni spletni strani in jih je mogoče ponovno uporabiti za promocijske in izobraževalne namene. V sklopu projekta smo natisnili 30 plakatov, 500 zloženk in 18.500 kazal. Tiskovine smo uporabnikom večinoma poslali po pošti, manjši del pa smo jih razdelili tudi na promocijskih dogodkih. Poslanih je bilo 150 pošiljk splošnim knjižnicam in večini šolskih knjižnic slovenskih osnovnih ter srednjih šol, vsaka pošiljka je vsebovala 100 kazal in informativno zloženko ter poziv prejemnikom, naj gradivo prosto ponudijo obiskovalcem knjižnice.

### 3.3. Diseminacijski dogodki

Pri izvedbi diseminacijskih dogodkov smo skušali pokriti čimbolj raznolike uporabniške skupine (od dijakov do učiteljev, lektorjev in raziskovalcev), pri čemer so imeli prednost tisti uporabniki, ki so sami izkazali interes za sodelovanje, ter tisti, kjer je bilo mogoče slovar predstaviti tudi v obliki delavnic. Izvedenih je bilo 16 dogodkov, med drugim predstavitev Slovarja sopomenk za lektorje

slovenščine v tujini, delavnice za dijake 3. in 4. letnikov, predstavitev na Lingvističnem krožku na Filozofski fakulteti UL, sopomenski turnir na sejmu Liber.ac 2019 in predstavitev dela na Slovarju sopomenk na konferenci eLex 2019. Celoten seznam dogodkov je opisan na spletni strani projekta. Predstavitvena gradiva so na voljo v spletni knjižnici in jih lahko kdorkoli uporabi za nadaljnje (samo)izobraževanje. Štiri diseminacijske dogodke smo posneli in podnaslovili, da so na voljo za ogled tudi po koncu projekta (gl. razdelek 4.2).

### 3.4. Anketa o novostih odzivnega slovarja

Za strukturirano zbiranje povratne informacije uporabnikov smo oblikovali spletno anketo, ki sprašuje o rabi slovarja, njegovi uporabnosti ter ponuja uporabnikom možnost, da opredelijo odnos do novosti odzivnega slovarja. Izvorno so anketo izpolnjevali udeleženci na dogodkih, nato pa smo jo diseminirali tudi širše. Anketo je v celoti izpolnilo 671 uporabnikov, kar je v kontekstu evropskih uporabniških raziskav relativno visoka številka.[2] Rezultate ankete smo tudi statistično analizirali in predstavili na konferenci eReL 2019 v Zagrebu. Anketa, ki je natančneje predstavljena v (Arhar Holdt, 2020), kaže, da so uporabniki naklonjeni do večine novosti, ki jih prinaša odzivni slovar, npr. stalno posodabljanje, digitalni format, kolokacijski podatki, povezave na korpus, uporabniško vključevanje. Kot izboljšavo med drugim predlagajo dodatne iztočnice (tudi nestandardne), dodatek kvalifikatorjev k zaznamovanim iztočnicam in sopomenkam, nekatere izboljšave vmesnika in večjo povezljivost z drugimi jezikovnimi viri.

## 4. Multimedijska izobraževalna gradiva

Multimedijska gradiva služijo za (samo)izobraževanje ciljnih uporabniških skupin tudi po koncu projekta in širijo domet informacij na uporabnike, ki imajo lahko korist od uporabe podnapisov (v slovenščini ali angleščini). V projektu so bila pripravljena video navodila, kako uporabljati slovar, posnetki predavanj o izdelavi slovarja in značilnostih odzivnega koncepta za različne vrste javnosti, in informativno-promocijske poddaje (podcasti), v katerih so avtorji slovarja, raziskovalci, predavatelji in predstavniki različnih uporabniških skupin Slovar sopomenk umestili v širši kontekst svojega poklicnega dela.

### 4.1. Poglavje na Portalu jezikovnih virov

Portal jezikovnih virov (https://viri.trojina.si/) je spletna knjižnica izobraževalno-promocijskih videoposnetkov o načinih uporabe prosto dostopnih spletnih jezikovnih virov in tehnologij.[3] Po uveljavljeni metodologiji (Arhar Holdt et al., 2016) je bilo za portal pripravljenih šest posnetkov, ki razlagajo: a) iskanje po slovarju; b) filtriranje in sortiranje podatkov; c) ocenjevanje sopomenk; č) dodajanje sopomenk; d) dostop do kolokacij in zgledov ter e) izvažanje podatkov. Poleg tega je na voljo tudi posnetek *Pet minut za Sopomenke*, ki predstavlja koncept slovarja, njegovo zasnovo in namen. Vsi posnetki vsebujejo slovenske in angleške podnapise, ki so bodisi integrirani v sam posnetek ali dodani z orodji platforme YouTube.

### 4.2. Posnetki predavanj in prosojnice

V spletni knjižnici projekta so zbrane povezave na posnetke štirih predavanj za različne tipe javnosti: a) za širšo javnost (Iztok Kosem: *Kam se obrniti, ko zmanjka besed?*); b) za šolskega naslovnika (Andreja Pavle Jurman: *Slovar sopomenk sodobne slovenščine in njegova uporabnost pri pouku slovenščine*); c) za študente slovenistike in prevajalstva (Špela Arhar Holdt in Jaka Čibej: *Digitalno slovaropisje: Kdo je slovar?* in č) za znanstveno javnost (Špela Arhar Holdt: *Thesaurus of Modern Slovene: By the Community for the Community*).[4]

Ob posnetih predavanjih so na voljo tudi prosojnice, ki so jih uporabljali predavatelji na posameznem od dogodkov. Na tej osnovi smo kot dodatno izobraževalno gradivo pripravili dopolnjene in odprto dostopne prosojnice o nastanku in sestavi Slovarja sopomenk sodobne slovenščine, ki jih učitelji in predavatelji lahko prenesejo na svoj računalnik in prosto prilagodijo za svoje delo v razredu oz. predavalnici.

### 4.3. Serija oddaj Sopomenki

V projektu smo pripravili serijo desetih poddaj z imenom *Sopomenki*. Z avtorji slovarja in drugimi predstavniki raziskovalne skupnosti smo obravnavali naslednje teme: a) interdisciplinarno sodelovanje računalničarjev in jezikoslovcev (gosta dr. Kaja Dobrovoljc in dr. Marko Robnik Šikonja); b) jezikoslovni pristopi k sopomenskosti (gosta dr. Polona Gantar in dr. Vojko Gorjanc); c) slovaropisje v preteklosti in prihodnosti (gosta dr. Simon Krek in dr. Marko Stabej); č) značilnosti odzivnih slovarjev (gostji dr. Špela Arhar Holdt in Barbara Gorišek); d) uporabniško vključevanje (gosta dr. Iztok Kosem in Jaka Čibej) in e) šibke točke odzivnih slovarjev (gostji dr. Nataša Logar in Eva Pori). V oddajah so sodelovali tudi predstavniki uporabniške skupnosti: f) predavatelj dr. Damjan Popič in študentka prevajalstva Sara Sever; g) učiteljica slovenščine Andreja Pavle Jurman in dijakinja Gimnazije Ledina Neja Šubic; h) prevajalca Andreja Golob in Josh Rocchio ter i) dr. Mojca Stritar Kučuk in Tanja Jerman, ki učita slovenščino kot drugi/tuji jezik. Posnetke je mogoče prek spletne povezave poslušati na različnih napravah ali jih prenesti na lastno napravo za poslušanje brez povezave.

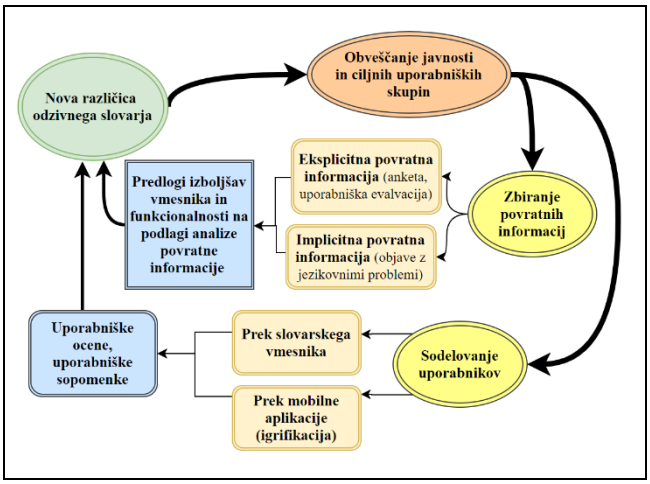
## 5. Spremljanje uporabniške povratne informacije

Projekt je zajemal razvoj in implementacijo postopkov za spremljanje uporabniških mnenj in odzivov na spletu ter metod za spremljanje različnih načinov uporabniškega vključevanja v slovarski razvoj. Uporabniško vključevanje pri razvoju odzivnega slovarja poteka po modelu, ki ga predstavlja Slika 1.

[2] Temeljna raziskava s področja slovarskih uporabnikov v digitalnem okolju, ki sta jo izvedla Koplenig in Müller Spitzer (2014: 83), je denimo vključevala 684 udeležencev.

[3] Portal je bil zasnovan v letih 2014–15 s podporo Ministrstva za kulturo: "Izdelava spletne strani z opisi jezikovnih virov in orodij za slovenščino ter osnovnimi (video)navodili za njihovo uporabo" in "Nadgradnja in popularizacija predstavitvenega portala spletnih jezikovnih virov za slovenščino".

[4] Zadnje predavanje je bilo posneto v okviru mednarodnega kongresa Euralex 2018 in je dostopno na portalu VideoLectures (http://videolectures.net/). Ostale tri posnetke smo zagotovili v sklopu projekta, skupaj s slovenskimi in angleškimi podnapisi.

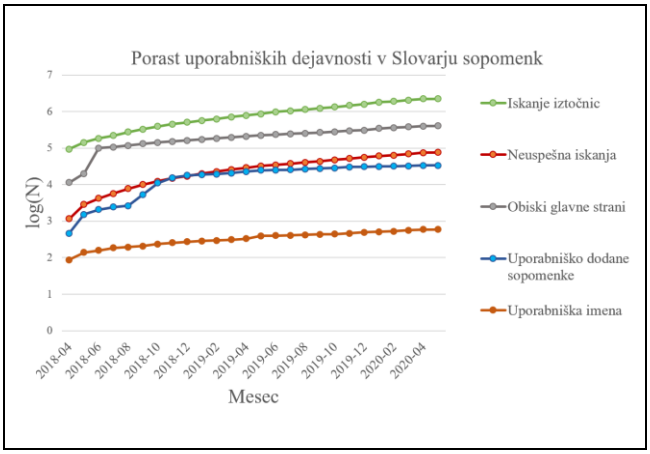



Slika 1. Uporabniško vključevanje v razvoj odzivnega slovarja.

Poleg *eksplicitnih povratnih informacij* uporabnikov, ki smo jih preverili z anketo o novostih odzivnega slovarja (gl. razdelek 3.4), smo v projektu spremljali tudi njihove *implicitne povratne informacije*, in sicer na dva načina: s spremljanjem dnevnikov iskanja in z analizo jezikovnih objav na družbenem omrežju Facebook.

## 5.1. Dnevniki iskanja

V projektu smo razvili metodologijo sistematičnega beleženja dnevnikov iskanja in spremljanja uporabniških aktivnosti med uporabo vmesnika Slovarja sopomenk. Beležijo se vsi podatki, ki jih je mogoče legitimno spremljati, od uporabniškega sodelovanja (ocenjene in dodane sopomenke) do iskanj (uspešnih in neuspešnih), premikov znotraj vmesnika (npr. razvrščanje, filtriranje podatkov, kliki na prikaz kolokatorjev in povezave na korpus Gigafida), izvoza podatkov, deljenja na družbenih omrežjih itd. (za natančnejši pregled gl. Arhar Holdt et al., 2018). Dnevnike izvažamo in analiziramo mesečno. Slika 2 prikazuje rast nekaterih uporabniških dejanj v slovarju od objave do 6. maja 2020.




Slika 2: Porast uporabniških dejavnosti v Slovarju sopomenk sodobne slovenščine.

Iz podatkov je med drugim razvidno, da uporabniki pogosto iščejo tudi iztočnice, ki jih v Slovarju sopomenk sodobne slovenščine še ni. Do 6. maja 2020 je bilo zabeleženih 2.251.093 iskanj, 75.672 neuspešnih iskanj, 406.454 obiskov glavne strani, 597 uporabniških imen, 26.683 uporabniško dodanih sopomenk in 23.730 nenajdenih nizov. Seznam najpogostejših nenajdenih nizov razkrije, katere iztočnice je treba prioritetno dodati pri nadgradnji oz. kako izboljšati iskalno vrstico (npr. avtomatsko dodajanje prostega morfema *se*, če uporabnik vpiše glagol, ki brez morfema ne obstaja, npr. *smejati*).

Opazen je tudi stalen porast uporabniškega sodelovanja, kar gre neposredno pripisati tudi aktivnostim promocijskega projekta. Visok odziv je presegel pričakovanja in priča o naklonjenosti jezikovne skupnosti do slovarske vključevalnosti, obenem pa potrjuje vrednost diseminacijskih projektov za seznanjanje javnosti o možnosti in omejitvah, ki jih prinašajo odzivni slovarji.

Prve analize uporabniško dodanih sopomenk (Čibej, 2019: 13–16) kažejo, da so predlogi v veliki večini dobronamerni in konstruktivni: na uravnoteženem vzorcu 1.662 sopomenk (vzorec zajema vseh 467 uporabnikov, ki so do decembra 2019 v slovar dodali vsaj en vnos) je bilo več kot 70 % sopomenk ustreznih in slogovno nevtralnih, več kot 20 % pa ustreznih in slogovno zaznamovanih (regionalizmi, termini ipd.). Le dobrih 6 % gradiva je neustreznega (lapsusi, zlonamerni vnosi). Opombe v oklepajih (npr. *v športu*; *slabšalno*), ki jih uporabniki pripišejo dodanim sopomenkam, kažejo, da želijo uporabniki dodano besedišče ob vnosu tudi ustrezno kvalificirati. Vsebnost kvalifikatorjev, zlasti možnost opozarjanja na potencialno žaljivo ali celo sovražno besedišče, je bila tudi na drugih mestih projekta izpostavljena kot ključna točka za nadaljnji razvoj slovarja. V prvi slovarski nadgradnji bomo zato prioritetno dodali tovrstne kvalifikatorje, omogočili pa bomo tudi, da svoje vnose s temi oznakami kvalificirajo že sami uporabniki.

Veliko število uporabniških predlogov priča o tem, da lahko jezikovna skupnost pomembno izboljša uporabnost slovarja. Ključna pri odzivnem konceptu pa je tudi kontrola kakovosti uporabniških vnosov. Ti se v vmesniku najprej prikazujejo v ločenem, spodnjem delu slovarskega gesla, ki sugerira jezikoslovno nepregledanost, vmesnik pa uporabnikom omogoča tudi, da na predloge reagirajo (pozitivno ali negativno). Po jezikoslovnem pregledu se uporabniški predlogi lahko umestijo v slovarsko bazo in premestijo v gornji del gesla, v primeru, da so zlonamerni, pa se lahko tudi izbrišejo. Uredniški posegi, ki bodo zagotovili kontinuirano izboljšavo kakovosti slovarskega gradiva, se bodo opravljali ob vsaki slovarski nadgradnji.

## 5.2. Objave na družbenem omrežju Facebook

Izdelali smo skripto, s katero lahko iz vnaprej določenih jezikovnih uporabniških skupin na Facebooku izluščimo objave in komentarje nanje ter nekatere metapodatke (npr. čas objave, število komentarjev na objavo, število všečkov in ostalih odzivov), skladno z zakonodajo pa skripta poskrbi za varovanje osebnih podatkov, tako da anonimizira uporabniška imena. Skripta je jezikovno neodvisna in odprto dostopno na voljo drugim zainteresiranim raziskovalcem (gl. Čibej in Arhar Holdt, 2018).

S programom smo izluščili podatke iz Facebookovih skupin *Za vsaj približno pravilno rabo slovenščine* ter *Društvo ljubiteljskih pravopisarjev in slovničarjev*. Vzorec 500 najnovejših objav smo ročno uredili in tako dobili 501

jezikovno vprašanje. Vprašanja smo kategorizirali, da bi ugotovili najpogostejše uporabniške jezikovne zadrege, ki jih Slovar sopomenk (skupaj z ostalimi odzivnimi slovarji) lahko naslovi (Čibej, 2019: 10–13).

Največ vprašanj (158 oz. slabih 31 % vzorca), ki so relevantna za Slovar sopomenk, spada v scenarij, pri katerem uporabnik sprašuje za mnenje drugih uporabnikov pri odločanju med dvema različicama besed ali besednih zvez. Na ravni sopomenskih besed in razlike med njimi je uporabnikom v tem primeru slovar lahko v pomoč, saj omogoča primerjavo sopomenk v kontekstu. Ugotovitev nakazuje, da je treba pri nadgradnji še nekoliko več pozornosti nameniti primerjavi sopomenk s kolokatorji. V trenutni različici so kolokatorji, s katerimi je mogoče primerjati sopomenke, izluščeni avtomatsko, navedeni pa so samo najbolj tipični po besednih vrstah. Izboljšati bi jih bilo mogoče z evalvacijo, pri kateri bi uporabniki npr. izrazili, katere besedne vrste kolokatorjev se jim zdijo najbolj intuitivne za primerjavo (npr. kombinacije prislov + pridevnik, pridevnik + samostalnik) in podobno.

## 6. Motiviranje uporabnikov za sodelovanje

Za uporabniško sodelovanje, ki je pomemben del koncepta odzivnega slovarja (Slika 1), je treba skupnost ustrezno motivirati. Vmesnik Slovarja sopomenk sodobne slovenščine vsebuje lestvico uporabnikov glede na število vnesenih sopomenk, v zavihku Skupnost pa je vloga sodelovanja pri nadaljnjem razvoju slovarja tudi eksplicitno izpostavljena. Če motivacijo za sodelovanje delimo na družbeno, psihološko in ekonomsko (Lew, 2013), se vmesnik dotika predvsem družbene ravni: naslavlja uporabniško željo sodelovati v skupnosti in v njej doseči določen status. V okviru različnih možnosti, ki jih ponuja množičenje na področju slovaropisja (gl. tudi Čibej et al., 2015; Fišer in Čibej, 2015), je trenutna rešitev precej osnovna. Pomemben del projekta je bil zato identifikacija dodatnih motivacijskih možnosti, ki bi raziskovanje sopomenskosti in sodelovanje pri razvoju slovarja naredili uporabnikom privlačnejše in dolgoročneje zanimivo. Želene prilagoditve smo na eni strani iskali s pomočjo uporabniških evalvacij slovarskega vmesnika, na drugi strani pa v smeri igrifikacije, ki bi v šolskem okolju lahko služila tudi didaktičnim namenom.

### 6.1. Uporabniška evalvacija slovarskega vmesnika

V sklopu projekta smo izvedli uporabniško evalvacijo slovarskega vmesnika, pri kateri je sodelovalo dvajset študentov jezikoslovnih smeri: deset s področja slovenistike in deset s področja prevajalstva. Evalvacije so potekale po metodi glasnega razmišljanja (ang. *think aloud protocol*), ki ji je sledil vodeni intervju z uporabniki (Štingl et al., 2018). Teme intervjuja so pokrile konceptualna vprašanja (npr. odnos do neprečiščenosti avtomatsko pridobljenih podatkov, nedokončanosti in spremenljivosti slovarja) kot tudi specifične lastnosti in funkcionalnosti vmesnika (npr. postavitev in poimenovanje gumbov, navigacija in preklapljanje med različnimi vrstami podatkov, opaznost in razumljivost vmesniških možnosti). Fokus je bil na vprašanjih, kaj uporabniki v zvezi s slovarjem dojemajo kot moteče in kaj bi jih za sodelovanje pri razvoju slovarja posebej motiviralo.

Izsledke predstavlja poročilo, ki je dostopno v spletni knjižnici (Arhar Holdt, 2019). Glavni pomisleki evalvatorjev so bili vezani na nejasno kontrolo uporabniških vnosov, tako nad tem, kdo predloge v slovar (lahko) dodaja, kot nad tem, kaj v slovarju po uredniškem pregledu in potrditvi (lahko) ostane. Ker uredniški delotoki v trenutnem vmesniku niso jasno predstavljeni, si sodelujoči niso znali razlagati, kako bo nadgradnja slovarja potekala. Na drugi strani so izpostavili pomanjkljivosti mehanike uporabniškega sodelovanja: moteče je, da je pri ocenjevanju avtomatsko pridobljenega gradiva sopomenki mogoče obenem pripisati tako pozitiven kot negativen glas; kot drugo, vmesnik uporabniku ne omogoča, da izbriše lasten predlog, zato v slovarju na uredniški poseg čakajo tudi lapsusi, ki bi jih lahko odstranili uporabniki sami. Našteti in še drugi predlogi evalvatorjev so strnjeni v seznam nalog za vmesniško nadgradnjo, ki vsebuje tudi določene izboljšave slovarske baze, npr. opredelitev besedne vrste pri slovarskih iztočnicah in s tem pripravo ločenih slovarskih vnosov za enakopisne iztočnice.

Evalvacija je pokazala, da elementi tekmovalnosti, ki jih v vmesnik vnaša tekmovalna lestvica, na sodelujoče nimajo posebej spodbudnega vpliva. Motivacija, kot so jo opredelili sodelujoči evalvatorji, je predvsem psihološka: pomagati drugim, pomagati odpraviti specifične probleme slovarske baze. V družbenem smislu bi jih motiviralo, če bi sodelovanje pri razvoju slovarja lahko uporabili za referenco v okviru študijskih nalog ali kot vnos v življenjepisu.

### 6.2. Sopomenski modul za Igro besed

Dobro izhodišče za motivacijo skupnosti ponuja igrifikacija. Tudi na področju razvoja jezikovne infrastrukture so že bile preizkušene t. i. igre z namenom (ang. *Games with a Purpose, GWAPs*), npr. za označevanje jezikovnih podatkov Phrase Detectives (Poesio et al., 2013), JeuxDeMots (Lafourcade, 2007) ali ZombiLingo (Guillaume, 2016). Ideja tovrstnih iger je, da motivirajo z igralnostjo, igranje pa je zasnovano tako, da obenem prispeva k določenemu uporabnemu cilju.

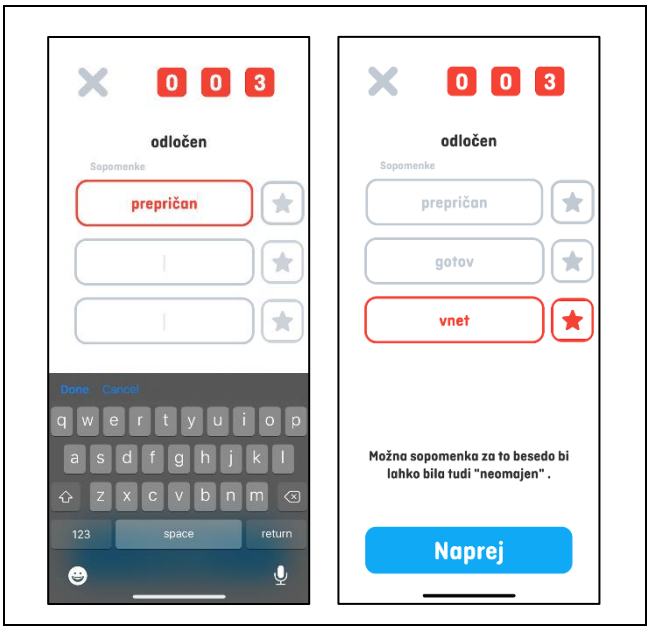
Igra z namenom, ki dopolnjuje model odzivnega slovarja, je mobilna aplikacija "Igra besed: kolokacije". Aplikacija, ki je nastala v projektu "Promocija jezikovne igralne aplikacije za mobilne naprave" (2018−2019),[5] se osredotoča na prepoznavo in dopolnjevanje slovenskih kolokacij. Vpisovanje, razvrščanje in selekcioniranje kolokacijskega gradiva, ki ga v obliki izzivov ponuja igra, posredno pomaga urejati bazo Kolokacijskega slovarja sodobne slovenščine (Kosem et al., 2018).

V projektu "Slovar sopomenk sodobne slovenščine: od skupnosti za skupnost" je bil razvit nov modul igre, ki se osredotoča na sopomenske podatke – modul "Sopomenke". Pri izbiri in pripravi nalog je bil ključen razmislek o didaktični vrednosti uporabniškega sodelovanja za spoznavanje koncepta sopomenskosti, usvajanje besedišča in razvijanje jezikovnega metaaparata. Igra Sopomenke posledično združuje množičenje in jezikovno didaktiko ter je kot taka primerna tudi za uporabo v razredu.

[5] Igra besed je bila prvotno razvita kot spletna aplikacija, dostopna na https://www.igra-besed.si/. Njeno pripravo je sofinanciralo Ministrstvo za kulturo RS v letu 2013.

Igra ponudi iztočnico, na katero igralka ali igralec vpiše do tri sopomenke. Vnosi se primerjajo z bazo Slovarja sopomenk in ustrezno točkujejo. Didaktično vrednost ima prikaz dodatne sopomenke iz baze (Slika 3), pa tudi progresivno stopnjevanje težavnosti igre po stopnjah, ki na začetku vsebujejo iztočnice z bogatim naborom sopomenk (npr. *hud, živahen, odličen*), kasneje pa iztočnice z redkejšimi sopomenkami (npr. *šumeč, utrujen, zvezda*). Učitelj ima možnost, da zahtevnost določene stopnje preveri vnaprej in za delo pri določeni šolski uri izbere težavnost, ki je ustrezna določeni stopnji ali skupini učencev.

Slika 3 prikazuje dva mobilna ekrana. Igralec se nahaja na stopnji 003 (skupno igra ponuja 500 stopenj s po 10 sopomenkami). Za iztočnico *odločen* igralec vnese sopomenke *prepričan, gotov* in *vnet*, od katerih je le zadnja v slovarski bazi, zato je vnos ocenjen z eno zvezdico. Kot "didaktično sopomenko" igra ponudi besedo *neomajen*. Z novimi sopomenkami se bo posodabljala tudi sama igra, s čimer bo izboljšana igralnost in uporabniška izkušnja z ocenjevanjem vnesenih besed.




Slika 3. Sopomenski modul v Igri besed.

## 7. Sklep in bodoče delo

Projekt "Slovar sopomenk sodobne slovenščine: od skupnosti za skupnost" prinaša rezultate, ki so uporabni za nadaljnje (samo)izobraževanje uporabniške skupnosti, kot tudi izsledke, kako skupnost razume in sprejema koncept odzivnega slovarja, katere novosti vidijo kot prednost in kaj se jim zdi v povezavi z novim konceptom problematično. Rezultati na drugi strani osvetljujejo, kakšne adaptacije bi bilo treba zagotoviti za izboljšavo uporabniške izkušnje z vmesnikom Slovarja sopomenk sodobne slovenščine. Predlogi uporabnikov so bili vključeni v načrt za nadgradnjo slovarja v različico 2.0, ki je že v pripravi.

Projektna izkušnja potrjuje, da je motivacija uporabnikov za sodelovanje pri nadaljnjem razvoju slovarja pomemben del odzivnega koncepta. Težave slovarja in širše odzivnega koncepta, ki jih uporabniki izpostavljajo kot demotivacijske, bodo naslovljene pri nadgradnji Slovarja sopomenk, kot tudi pri nadaljnjem razvoju odzivnih virov. Na drugi strani vpeljava igrifikacije v odzivni model ponuja nove, še ne preverjene možnosti za sodelovanje skupnosti. V naslednjem koraku zato želimo oceniti učinkovitost tovrstnega sodelovanja in kvaliteto podatkovnega doprinosa, igralni modul pa bi radi tudi evalvirali s pomočjo uporabnikov in ga ustrezno izboljšali.

Koncept odzivnega slovarja je odgovor na potrebe novega časa, vendar bo pri njegovem razvoju tudi v bodoče nujno upoštevati uporabniško povratno informacijo. Model uporabniškega vključevanja, ki je bil razvit v projektu, je mogoče videti kot podstat za nadaljnji, kontinuirani razvoj v smer vedno boljšega rezultata, pri razvoju katerega – za razliko od preteklih praks – slovaropisje uporabnikov ne želi le "ciljati", ampak jih po mnenju, potrebah in védenju aktivno sprašuje ter povratno informacijo v čim večji meri tudi upošteva.

## 8. Zahvala

Projekt, opisan v prispevku, je med letoma 2018 in 2019 financiralo Ministrstvo za kulturo Republike Slovenije. Raziskovalni program št. P6-0411 (Jezikovni viri in tehnologije za slovenščino) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Raziskave množičenja in uporabniškega vključevanja v projektu so potekale v povezavi z vsebinami akcije COST CA160105 eNetCollect (European Network for Combining Language Learning with Crowdsourcing Techniques).

Zahvaljujeva se vsem, ki so prispevali k projektnim rezultatom: dr. Kaji Dobrovoljc, dr. Poloni Gantar, mag. Teji Goli, Andreji Golob, Barbari Gorišek, dr. Vojku Gorjancu, Tanji Jerman, Leonu Noetu Jovanu, dr. Iztoku Kosmu, dr. Simonu Kreku, dr. Cyprianu Laskowskemu, dr. Nataši Logar, Andreji Pavle Jurman, dr. Damjanu Popiču, Evi Pori, Juretu Preglau, Davidu Rajšterju, Joshu Rocchiu, dr. Marku Robniku Šikonji, Sari Sever, dr. Marku Stabeju, dr. Mojci Stritar Kučuk, Mihu Šajini, Neji Šubic, Petri Štingl, Lenki Trdina in Gašperju Uršiču.

## 9. Literatura in viri

Špela Arhar Holdt. 2015. Uporabniške raziskave za potrebe slovenskega slovaropisja: prvi koraki. V: Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 136–149. Ljubljana: Znanstvena založba Filozofske fakultete.

Špela Arhar Holdt, Kaja Dobrovoljc in Iztok Kosem. 2016. Predstavitveni portal spletnih jezikovnih virov za slovenščino. V: *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2016*, str. 27–31.

Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Bojan Klemenc, Iztok Kosem, Simon Krek, Cyprian Laskowski in Marko Robnik Šikonja. 2018. V: *Proceedings of the 18th EURALEX International Congress: lexicography in global contexts*, str. 401–410. Ljubljana: Ljubljana University Press.

Špela Arhar Holdt. 2019. *Specifikacije za adaptacije slovarskega vmesnika. Poročilo projekta Slovar sopomenk sodobne slovenščine: Od skupnosti za skupnost*. Ljubljana: CJVT. https://www.cjvt.si/promocija-sopomenk/knjiznica/.

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): v tisku.

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*, str. 70–83. Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton: Lexical Computing.

Jaka Čibej in Špela Arhar Holdt. 2018. Researching Dictionary Needs of Language Users Through Social Media: A Semi-Automatic Approach. V: *Proceedings of the 18th EURALEX International Congress: lexicography in global contexts*, str. 67–76. Ljubljana: Ljubljana University Press.

Jaka Čibej. 2019. *Končno poročilo o spremljanju uporabniških odzivov, mnenj in načinov uporabniškega vključevanja. Projekt Slovar sopomenk sodobne slovenščine: Od skupnosti za skupnost*. Ljubljana: CJVT. https://www.cjvt.si/promocija-sopomenk/knjiznica/

Kaja Dobrovoljc et al. 2019. Morphological lexicon Sloleks 2.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1230.

Darja Fišer in Jaka Čibej. 2015. Potencial množičenja v sodobni leksikografiji. V: Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 542–564. Ljubljana: Znanstvena založba Filozofske fakultete.

Bruno Guillaume, Karën Fort in Nicolas Lefebvre. 2016. Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax. V: *Proceedings of the International Conference on Computational Li*nguistics (COLING). https://hal.inria.fr/hal-01378980/.

Alexander Koplenig in Caroline Müller-Spitzer. 2014. The first two international studies on online dictionaries – background information. V: Caroline Müller-Spitzer, ur., *Using Online Dictionaries*. Berlin, Boston: De Gruyter Mouton.

Iztok Kosem, Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej in Cyprian Laskowski. 2018. Collocations dictionary of modern Slovene. V: *Proceedings of the 18th EURALEX International Congress: lexicography in global contexts*, str. 989–997. Ljubljana: Ljubljana University Press.

Iztok Kosem et al. 2019. Collocations Dictionary of Modern Slovene KSSS 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1250.

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: *Proceedings of eLex 2017: Lexicography from Scratch*, str. 93–109. Leiden, Netherlands.

Simon Krek et al. 2018. Thesaurus of Modern Slovene 1.0. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1166.

Mathieu Lafourcade. 2007. Making people play for Lexical Acquisition with the JeuxDeMots prototype. V: *SNLP'07: 7th International Symposium on Natura*l Language Processing, str. 7. Pattaya, Chonburi, Thailand.

Robert Lew. 2013. User-generated content (UGC) in online English dictionaries. V: Andrea Abel, Annette Klosa, ur., *Ihr Beitrag bitte! – Der Nutzerbeitrag im Wörterbuchprozess (OPAL – Online publizierte Arbeiten zur Linguistik)*, str. 9–30. Mannheim: Institut für Deutsche Sprache.

Nataša Logar, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Lawrence Page, Sergey Brin, Rajeev Motwani in Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web. Technical Report*. Stanford InfoLab.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo in Luca Ducceschi, 2013. Phrase Detectives: Utilizing Collective Intelligence for Internet-scale Language Resource Creation. V: *ACM Transactions on Interactive Intelligent Systems 3*, 3:1–3:44.

Mojca Šorli, Katarina Grabnar, Simon Krek in Tomato Košir. 2006. Oxford-DZS comprehensive English-Slovenian dictionary. V: *Proceedings of the XII EURALEX International Congress*, str. 631–637. Edizioni dell`Orso: Universita di Torino: Academia della Crusca.

Petra Štingl, Apolonija Gantar in Špela Arhar Holdt. 2018. Research on the User Experience of the Thesaurus of Modern Slovene. V: *Proceedings of MEi:CogSci Conference 2018, Bratislava*, str. 98.

# Referenčni seznam pogostih splošnih besed za slovenščino

**Špela Arhar Holdt,[♦*] Senja Pollak[‡], Marko Robnik Šikonja,[*] Simon Krek[†♦*]**

[♦]Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, SI-1000 Ljubljana
Spela.ArharHoldt@ff.uni-lj.si
*Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, SI-1000 Ljubljana
Marko.RobnikSikonja@fri.uni-lj.si

[‡] Laboratorij za tehnologije znanja
[†]Laboratorij za umetno inteligenco
Institut Jožef Stefan
Jamova 39, SI-1000 Ljubljana
{senja.pollak,simon.krek}@ijs.si

**Povzetek**

Prispevek predstavlja pripravo in vsebino referenčnega seznama pogostih splošnih besed za slovenščino. Seznam smo pripravili s prekrivanjem najpogostejših 10.000 lem iz štirih slovenskih besedilnih korpusov: uravnoteženega referenčnega korpusa pisne slovenščine Kres, referenčnega korpusa govorjene slovenščine GOS, korpusa računalniško posredovane komunikacije Janes ter korpusa šolske pisne produkcije Šolar 2.0. Kandidatke za seznam so tiste pojavnice, ki se pojavljajo v vseh naštetih korpusih. V prispevku opišemo pripravo seznama vključno s čiščenjem, nato uspešnost pristopa ocenimo s pomočjo analize besedišča na b-, ki je v seznam vključeno, v primerjavi s tistim, ki ostane pod pragom vključitve, in s pregledom pokritosti besedišča izbranih tematskih sklopov, ki se pogosto pojavljajo v jezikovni didaktiki (poimenovanja za dele telese in hrano). Končni seznam, ki vsebuje 4.768 pogostih splošnih lem, bo raziskovalni skupnosti na voljo na repozitoriju CLARIN.SI.

**Reference List of Slovene Frequent Common Words**

We present the reference list of Slovene most frequent common words. The list was prepared by selecting vocabulary at the intersection of the most frequent 10,000 lemmas of four Slovene text corpora: the balanced reference corpus of written Slovene Kres, the reference corpus of spoken Slovene GOS, the corpus of computer-mediated communication Janes and the corpus of school written production Šolar 2.0. The paper describes the preparation of the list, including manual cleaning of results, then evaluates the approach through a detailed analysis of the vocabulary starting with b-. We compare the vocabulary, which is included in the list, compared to the vocabulary below the selected threshold, and analyse the vocabulary coverage of selected thematic fields, which are frequent in language didactics (names for body parts and food). The final list containing 4,768 common general lemmas, will be made available to the research community at the CLARIN.SI repository.

## 1. Uvod

Referenčni seznami temeljnega, jedrnega in pogostega besedišča se uporabljajo za različne naloge na področjih obdelave naravnega jezika, teoretičnega, uporabnega in eksperimentalnega jezikoslovja ter sorodnih disciplin, denimo za pripravo učnih ciljev in gradiv na področju jezikovnega učenja, diagnostiko učnih primanjkljajev na področju branja in pisanja, strojno ocenjevanje berljivosti besedil, pisanje besedil za lahko branje in podobno.

Metodologija priprave tovrstnih seznamov je odvisna od njihove namembnosti in razpoložljivih jezikovnih virov za določen jezik. V prispevku predstavljamo seznam pogostega splošnega besedišča, ki smo ga pripravili iz štirih slovenskih korpusov: uravnoteženega referenčnega korpusa pisne slovenščine Kres (Logar et al., 2012), referenčnega korpusa govorjene slovenščine GOS (Verdonik & Zwitter Vitez, 2011), korpusa računalniško posredovane komunikacije Janes (Fišer et al., 2020) ter korpusa šolske pisne produkcije Šolar 2.0 (Kosem et al, 2016).

Seznam je nastal pod okriljem projekta Za kakovost slovenskih učbenikov,[1] katerega osrednji cilj je razvoj kazalnikov kakovosti učbenikov za praktično uporabnost v procesu potrjevanja učbenikov in njihove evalvacije. V okviru projekta je bil implementiran v orodje za oceno berljivosti slovenskih besedil (Škvorc et al., 2019). Seznam smo zgradili in v več iteracijah pregledali in prečistili, kot pojasnjuje pričujoči prispevek. Predstavljeni seznam bo po objavi prispevka raziskovalni skupnosti dostopen na repozitoriju CLARIN.SI.[2]

## 2. Besedni seznami za različne namene

Glede na konceptualno in metodološko ozadje je mogoče govoriti o seznamih temeljnega, jedrnega, splošnega, pogostega, preprostega (in še kakšnega) besedišča. Na videz podobna poimenovanja nakazujejo pomembne razlike, ki jih je pri interpretaciji in uporabi rezultatov treba upoštevati.

Poimenovanja *enostavne, preproste, lahke, razumljive besede* nekoliko zavajajoče sugerirajo, da je bila pri sestavi seznama preverjena in upoštevana jezikovnorecepcijska izkušnja, kar (pri korpusnem in tudi drugih pristopih) običajno ne drži. Raziskave razumevanja jezika so kompleksne in zahtevajo upoštevanje številnih znotraj- in zunajjezikovnih dejavnikov (Ferbežar & Stabej, 2008), zato so v praksi izredno redke. Dejstvo, da je besedišče na seznamih izluščeno iz konteksta in predstavljeno v osnovni (slovarski, lematizirani) obliki, vprašanje razumljivosti in njenega preverjanja še dodatno zaplete. V tem smislu je pri besednih seznamih ustrezneje govoriti o *predvideni enostavnosti*, ki jo lahko opredeljujejo značilke na ravni

[1] Kratko ime projekta: Kauč, projekta stran: http://kauc.si/.

[2] Na http://hdl.handle.net/11356/1346.

oblike (npr. kratkost besede, odsotnost grafično podobnih oz. lahko zamenljivih črk, morfološka in fonološka regularnost ter enoznačnost besede) in pomena (npr. enopomenskost in polnopomenskost besede, konkretnost pomena, prisotnost denotata v vsakodnevnem življenju, besedotvorna osnovnost in transparentnost besede). Pri pripravi seznama, ki ga predstavljamo v prispevku, značilke predvidene enostavnosti niso upoštevane, mogoče pa ga je v to smer nadgraditi naknadno.[3]

Formule berljivosti za angleščino pogosto uporabljajo seznama, ki so ju pripravili Dale in Chall (1948) ali Spache (1953). Seznama vsebujeta besede, ki so glede na testiranja v razredu *poznane* večini šolske populacije na določeni stopnji šolanja: Dale-Chall zajema 3.000 besed, ki so (bile) znane vsaj 80 % ameriških učencev četrtega razreda osnovne šole, Spache pa navaja okrog 1.000 besed, znanih učencem prve triade osnovne šole. Seznama sta bila v desetletjih uporabe kritično ovrednotena z več vidikov, na tem mestu pa je ključno izpostaviti predvsem, da se tovrstni seznami od tistih, ki so osnovani na referenčnih korpusih, lahko precej razlikujejo, saj se besedišče, ki ga pri usvajanju jezika spoznamo zgodaj, v kasnejši jezikovni rabi ne pojavlja nujno zelo pogosto.[4] Vseeno pa je mogoče korpusno osnovani seznam pogostega splošnega besedišča naknadno oceniti tudi z vidika poznanosti besedišča na določeni stopnji šolanja, tako pri govorkah in govorcih slovenščine kot prvega jezika kot tistih, ki slovenščino spoznavajo kot drugi ali tuji jezik.

Na drugi strani poimenovanja *jedrne, temeljne, osnovne besede* nakazujejo, da seznam zajema besedišče, ki je z vidika določene naloge (npr. učenja jezika v določeni vrsti jezikovne rabe) prioritetno. »Temeljnost« besedišča je relevantna tudi pri izboljševanju učinkovitosti pojasnjevalnega oz. definicijskega diskurza v učbenikih, učnih gradivih, slovarjih in podobnem. Korpusni pristop za izdelavo seznamov temeljnih besed je že bil preizkušen, npr. v projektu Kelly,[5] pod okriljem katerega so bile pripravljene kartice za učenje devetih (tujih) jezikov na različnih nivojih SEJO (Splošni jezikovni okvir, ang. CEFR). Domet pristopa opisujeta Johansson Kokkinakis & Volodina (2011): seznam temeljnega besedišča sodobne švedščine je bil pripravljen na osnovi korpusa SweWAC (Swedish Web-Acquired Corpus), iz katerega so izvozili in ročno pregledali 9.000 najpogostejših pojavnic. Kot izziv se je pokazalo nestandardno besedišče, besedne variante, lastna imena ter težave z lematizacijo in oblikoskladenjskim označevanjem. Pričakovano (ker gre za spletni korpus) je pregled rezultatov razkril prisotnost trendovskega, tujejezičnega in na (politično) zgodovino vezanega besedišča, medtem ko je za didaktične namene umanjkalo besedišče tematskih sklopov, ki se pri učenju jezika najbolj tipično pojavljajo. Seznam denimo ni vseboval vseh imen za dneve v tednu, sorodstvena razmerja, hrano, dele telesa in podobno.[6] Seznam je bil posledično ročno dopolnjen, dodatno pomoč pri iskanju 'temeljnosti' pa je nudila tudi primerjava z rezultati drugih sodelujočih jezikov.

Z upoštevanjem različnih pristopov k izdelavi seznamov v tem prispevku govorimo o *pogostem splošnem besedišču*. Osnova za seznam so frekvenčni spiski: seznami korpusnih lem, urejeni padajoče glede na pogostost v obravnavanem korpusu. Korpusni pristop pogostost jezikovnih pojavov v korpusu razume kot njihovo tipičnost za tisto vrsto jezikovne rabe, ki jo korpus kot premišljeno pripravljeni besedilni vzorec reprezentira. Pri tem je treba upoštevati, da niso vse besede enako pogoste v vseh vključenih korpusih, saj so ti različnih velikosti in sestave. Splošnost vključenega besedišča dosežemo s prekrivanjem podatkov iz več različnih korpusov: kot splošne razumemo besede, ki se pojavljajo tako v pisnem kot govornem prenosniku, tako pri šolajoči se populaciji kot pri odraslih piscih in (kar se tiče odraslih jezikovnih uporabnikov) tako pri predvideni vsakdanji recepciji kot produkciji.

## 3. Metodologija

Pristop temelji na predpostavki, da bodo besede na presečišču različnih besedilnih korpusov predstavljale najbolj splošno besedišče. Uporabili smo korpuse Janes, Šolar in GOS, ki so na voljo na repozitoriju NoSketch Engine,[7] in korpus Kres v lokalni inštalaciji programa Sketch Engine[8] ter iz vsakega izvozili 10.000 najpogostejših lempos (leme s pripisano besednovrstno oznako). Kot kandidate za splošno besedišče smo izbrali besede, ki se pojavljajo med najpogostejšimi besedami v vseh štirih korpusih, tako pripravljeni izhodiščni seznam pa smo nato pregledali in prečistili. Statistike velikosti korpusov so podane v Tabeli 1.

### 3.1. Uporabljeni korpusi

#### 3.1.1. Kres

Korpus Kres (Logar et al., 2012) je vzorčeni uravnoteženi podkorpus referenčnega korpusa[9] pisne slovenščine Gigafida (ibid.). Ima skoraj 100 milijonov besed (in več kot 120 milijonov pojavnic) in vsebuje časopisje, revije, stvarna besedila, leposlovje in internetne vsebine. Do korpusa smo dostopali preko lokalne inštalacije programa Sketch Engine in uporabili različico, objavljeno 01/03/2017.

#### 3.1.2. Janes

Korpus Janes (Fišer et al., 2020) je korpus slovenskih spletnih uporabniško generiranih vsebin. Zajema besedila (v velikosti okoli 191 milijonov besed) iz forumov, tvitov in blogov ter uporabniških komentarjev. Janes kot referenčni vir za slovensko računalniško posredovano komunikacijo vsebuje tudi pisna besedila v nestandardni slovenščini. Do korpusa smo dostopali preko platforme NoSketch Engine in uporabili različico Janes (družbena omrežja) v1.0, objavljeno 10/28/2017.

#### 3.1.3. Šolar Clear 2.0

Korpus Šolar 2.0 (Kosem et al., 2016) vsebuje pisna besedila, ki so jih učenci zadnje triade in dijaki različnih srednjih šol samostojno tvorili pri pouku. Največji delež besedil predstavljajo eseji oziroma spisi, nastali pri pouku slovenščine, v korpusu pa najdemo še pisne izdelke, nastale

[3] Denimo za potrebe lahkega branja, ki ga v slovenskem prostoru razvija zavod Risa (http://www.risa.si/).
[4] Za primer npr. besede *baa, babies, banjo, bedbug, bedtime, bonnet, bow-wow, buggy* s seznama Dale-Chall.
[5] Keywords for Language Learning for Young and Adults Alike (2009–2012), https://spraakbanken.gu.se/en/projects/kelly.
[6] Tipični primeri manjkajočega besedišča, ki jih navajata avtorici, so *orange, elbow* in *alphabet* (Johansson Kokkinakis & Volodina 2011: 138). K vprašanju manjkajočega besedišča se vračamo v poglavju 4.4.
[7] Dostopno kot del raziskovalne infrastrukture CLARIN.SI: https://www.clarin.si/noske/index-en.html.
[8] Lokalna inštalacija Centra za jezikovne vire in tehnologije Univerze v Ljubljani na https://sketch.cjvt.si/.
[9] Kres (in ne celotne Gigafide) smo uporabili zaradi boljše uravnoteženosti zajetih besedilnih zvrsti.

pri drugih učnih urah, in teste. Pri izdelavi seznama smo uporabili različico korpusa Šolar 2.0 Clear, ki ne vsebuje oznak za učiteljske popravke. Do korpusa smo dostopali preko platforme NoSketch Engine in uporabili različico, objavljeno 11/07/2019.

#### 3.1.4. GOS

Korpus GOS (Verdonik & Zwitter Vitez, 2011) je referenčni korpus govorjene slovenščine. Zajema 120 ur posnetkov govorjene slovenščine, kar ustreza korpusu transkripcij v velikosti malo več kot milijon besed. Posnetke tvorijo različni tipi govora, od zasebnih pogovorov (med prijatelji, z družino) do delovnih sestankov, svetovanj, pogovora ob storitvah, televizijskega govora, šolskega diskurza ipd. Do korpusa smo dostopali preko platforme NoSketch Engine in uporabili različico, objavljeno 04/12/2019.

| | Št. besed | Št. stavkov | Št. besedil |
|---|---|---|---|
| Kres | 97.135.649 | 7.599.063 | 21.456 |
| Janes | 191.292.328 | 27.903.937 | 12.864.041 |
| Šolar | 1.625.118 | 125.712 | 5.485 |
| GOS | 1.033.024 | 122.961 | 287 |

Tabela 1: Obseg uporabljenih korpusov.

### 3.2. Priprava seznama

#### 3.2.1. Uporaba lempos

Za vse korpuse prejšnje sekcije smo izvozili frekvenčne spiske najpogostejših 10.000 besed.[10] Pri izvozu smo se odločili za oznake lempos (lemma in oznaka o besedni vrsti, npr. *belina-s*), ki omogoča razlikovanje med homonimi različnih besednih vrst (npr. samostalnik *lev* ali pridevnik *lev*). Homonimi enake besedne vrste na seznamu niso ločeni (npr. *klop* kot žival ali *klop* kot pohištvo). Prednost uporabe lempos je, da vsaj deloma loči homonime, pomanjkljivost pa, da so zaradi razlik pri oblikoskladenjskem označevanju nekateri podatki v seznamih po nepotrebnem razpršeni (npr. beseda *edin*, ki se lahko pojavlja označena kot pridevnik ali samostalnik).

#### 3.2.2. Uporaba prilagojene frekvence ARF

Za pogostost uporabljamo mero prilagojene frekvence ARF (angl. *average reduced frequency*).[11] To je varianta frekvenčnega seznama, ki ne šteje pojavitev iste besede, ki se pojavljajo skupaj, npr. v istem dokumentu. Mera ARF je bila razvita posebej za pripravo seznamov, kjer je pomembna splošnost besedišča, kot razložita Savický & Hlaváčová (2002), ki formulo tudi natančneje razložita. Za vsak frekvenčni spisek posebej smo frekvenco ARF relativizirali glede na velikosti uporabljenih korpusov.

#### 3.2.3. Združevanje frekvenčnih spiskov

Frekvenčne spiske smo združili v izhodiščni seznam besed, v katerega smo izpisali lempos, relativizirane ARF (rARF) za posamezen korpus, poleg tega pa še absolutno število korpusov, v katerih se beseda pojavlja (npr. 4 za prisotnost v frekvenčnih spiskih vseh štirih korpusov, 1 za prisotnost v posameznem korpusu). Prav tako smo izračunali *skupno relativizirano ARF*, definirano kot kvocient vsote relativnih frekvenc (prilagojena frekvenca/št. besed) v vseh seznamih in števila seznamov (v našem primeru je število seznamov 4). Podatke v obsegu 17.756 besed smo uredili padajoče po tej vrednosti. Tabela 2 na naslednji strani prikazuje format izhodiščnega seznama pred čiščenjem.

### 3.3. Pregled in čiščenje seznama

Na končni seznam smo vključili besede, ki so vseh štirih korpusih, z izjemo lastnih imen in tujejezičnih lem. Prečiščeni seznam obsega 4.768 lem.

#### 3.3.1. Lastna imena

Seznam smo najprej filtrirali, da je vseboval samo rezultate, ki se pojavljajo na vseh štirih korpusih. Nato smo odstranili 89 lem, ki se pojavljajo v zapisu z veliko začetnico. Med odstranjenimi besedami so predvsem zemljepisna lastna imena oz. njihovi deli (*Slovenija, Ljubljana, Sobota, Big*) in nekaj pridevnikov (*Prešernov, Škofji, Nobelov*). Opozoriti je treba, da se v frekvenčnih seznamih pojavljajo kratična poimenovanja tako v lematizaciji z velikimi (*TV, CD*) kot tudi samimi malimi črkami (*tv, cd*), kar pomeni, da so ti podatki razpršeni in potrebuje njihovo vključevanje dodaten razmislek.

#### 3.3.2. Neznane/tujejezične leme

S seznama smo odstranili 3 leme, ki so bile strojno označene kot neznane/tujejezične, in sicer *i, and, in*. Tovrstnih lem v končnem koraku priprave seznama ni veliko, ker se redko pojavljajo v vseh štirih korpusih. Vse ostale besedne vrste (glede na pripisane oznake) smo na seznamu ohranili (prim. Tabela 3). Glede na izkušnjo Johansson Kokkinakis & Volodine (2011) bi bilo mogoče odstraniti števnike (razen izbranih najbolj osnovnih) ali tudi medmete, členke in druge nepolnopomenske besedne vrste. Odločili smo se, da ohranimo vse besedišče in morebitno selekcijo po besednih vrstah prepustimo uporabnikom seznama glede na konkretni namen uporabe.

#### 3.3.3. Ocena uporabe ARF

Izračun skupne povprečne relativne frekvence smo v metodologijo vključili za lažje razvrščanje končnih rezultatov, vendar smo ugotovili, da dokaj dobro nakazuje tudi relevantnost podatkov. Med prvimi 1.000 besedami izhodiščnega seznama, urejenega padajoče glede na to mero, je denimo samo 27 besed, ki niso v vseh štirih obravnavanih korpusih. Med temi prednjačijo tujejezične leme (*for, is, it*), medmeti (*eee, eem, mmm*) in besede, ki jih označevalniki uvrščajo v različne besedne vrste (*fajn, samo*). Za dodatno analizo in razmislek o potencialni naknadni vključitvi so zanimive predvsem pogovorne besede (*okej, ful, fajn*) in besede, ki v enem od korpusov niso prisotne (npr. samostalnik *komentar*, ki se v korpusu Šolar 2.0 sicer pojavi, vendar ne z dovolj visoko ARF, da bi prišel na seznam prvih 10.000).

[10] V pilotnih eksperimentih smo testirali tako 5.000 kot 10.000 in ugotovili, da z izvozom 10.000 dobimo večji nabor besed, ki pa še zmeraj ustreza potrebam seznama splošnega besedišča.

[11] Slovenski prevod v orodju NoSketch Engine "povprečna relativna frekvenca" je dvoumen, saj ne gre za relativne frekvence, ampak prilagojene oz. zmanjšanje frekvence. Zato v prispevku uporabljamo angleško različico poimenovanja.

| lempos | rARF_Janes | rARF_GOS | rARF_Kres | rARF_Šolar | rARF_skupna | št_korpusov |
|---|---|---|---|---|---|---|
| zanimanje-s | 0,00106 | 0,00058 | 0,00324 | 0,00271 | 0,00190 | 4 |
| kupec-s | 0,00206 | 0,00077 | 0,00298 | 0,00178 | 0,00190 | 4 |
| nogomet-s | 0,00222 | 0,00077 | 0,00207 | 0,00252 | 0,00190 | 4 |
| forum-s | 0,00575 | 0,00068 | 0,00117 | None | 0,00190 | 3 |
| slednji-p | 0,00211 | None | 0,00356 | 0,00191 | 0,00190 | 3 |
| prilagoditi-g | 0,00117 | 0,00126 | 0,00202 | 0,00314 | 0,00190 | 4 |

Tabela 2: Združeni seznam besedišča iz štirih korpusov.

# 4. Rezultati

Kot rezultate našega postopka navajamo najprej statistiko splošnih besed po besednih vrstah, nato besede na b- podrobneje analiziramo v smislu vključenosti in nazadnje ocenimo pokritost besedišča za potrebe jezikovne didaktike.

## 4.1. Splošni opis seznama

Seznam pogostih splošnih besed obsega 4.768 lem kot prikazuje Tabela 3.

| Besedna vrsta | Število | Primeri besed |
|---|---|---|
| samostalnik | 2.038 | leto, človek, čas, dan |
| glagol | 1.174 | biti, imeti, vedeti, iti |
| pridevnik | 914 | dober, velik, sam, nov |
| prislov | 427 | tako, lahko, zdaj, kako |
| zaimek | 59 | se, ta, on, jaz, ves, ti |
| predlog | 43 | v, na, z, za, po, od, iz |
| veznik | 39 | in, da, pa, ki, kot, če, ko |
| členek | 37 | ne, tudi, ja, še, že, samo |
| števnik | 30 | en, drug, dva, prvi, trije |
| medmet | 7 | aha, ej, hm, joj, ah, oh |

Tabela 3: Število besed po besednih vrstah.

## 4.2. Vključene besede na b-

Za primer zajetega in tudi izpuščenega besedišča navajamo besede, ki se začnejo na črko *b*-. Na seznamu je 112 takih besed, od *biti* do *bonbon*. Če seznam razdelimo na štiri dele, dobimo naslednje rezultate (besede so navedene po besednih vrstah, ki jih ločuje podpičje):

- **1. rang** – prvih 1000 besed, med njimi 21 besed na b-: *beseda, bistvo, besedilo, bog, brat, barva, boj, bolezen, branje, banka, bolečina*; *biti, brati, bati, boriti, boleti*; *bel, bogat; bolj, blizu* (prislov)*; brez.*
- **2. rang** – drugih 1000 besed, med njimi 40 na b-: *bližina, bitje, babica, b, bolnišnica, bralec, bolnik, blok, bogastvo, blago, bivanje, bitka, breg, balkon, baba, blagajna, bazen, breme, bolnica, baza, bar, bomba; braniti, bežati, bližati, bojevati, brigati; bližnji, bolan, bivši, božji, bodoč, blizek, brezplačen, bistven, beseden, boleč; bistveno; blizu* (predlog)*; bodisi.*
- **3. rang** – tretjih 1000 besed, med njimi 26 na b-: *borec, bok, blato, božič, bratranec, borba, beda, boja, banana, brazda, borza, bit, biser, bogataš, bedak; brisati, bivati, brskati; barven, blag, beden, blagoven, brezposeln, bolniški, bos; brezplačno.*
- **4. rang** – preostalih 771 besed, med njimi 25 na b-: *bik, baraba, biologija, bob, boben, brk, biblija, buča, bratec, blagoslov, brezdomec, bonbon; bruhati, božati, besediti, brcniti, begati; bojen, bencinski, bled, besen, bister, bronast, blejski; bogato.*

Navedeni nabor besed dobro odraža specifike izbrane metodologije. Ker je seznam enobeseden, ne vključuje zaimka *se* ob glagolskih nedoločnikih (npr. *braniti – braniti se*). Zaradi upoštevanja besednovrstne oznake se v seznamu določene besede lahko ponovijo (npr. *blizu* kot prislov in predlog). Nekatere leme so na prvi pogled nenavadne, ker se v rabi uporabljajo v stopnjevani (*blizek – najbližji*) ali določni obliki (*beseden – besedni; bojen – bojni*).

Več dvomov zbujajo besede, pri katerih so verjetne lematizacijske težave, npr. glagol *besediti*, ki je morda na seznamu pristal na račun napačno lematiziranih samostalnikov *beseda* in *besedilo*. Razmisliti je, ali bi s seznama odstranili posamezne črke (npr. *b*), ki so v besedilih tipično uporabljene za naštevanje. Nekoliko arbitrarno se zdi tudi, da so s seznama odstranjeni lastnoimenski samostalniki, ostajajo pa iz njih izpeljani pridevniki (*blejski*). Kljub navedenim pomislekom se zdi predstavljeni nabor ustrezen in skladen s pričakovanji.

## 4.3. Nevključene besede na b-

Končni seznam ne vsebuje 348 besed na b-, ki se med najpogostejših 10.000 pojavijo samo v delu obravnavanih korpusov. Navajamo jih glede na to, v katerih korpusih se pojavljajo, in sicer do 15 primerov pri vsaki skupini (tudi tukaj podpičje ločuje skupine besed po besednih vrstah).

- **Janes, GOS, Kres** – 34 besed, na primer: *baterija, bencin, bučka, bruto, bilanca, beton, bojazen, bife; blokirati, beležiti, bosti; britanski, bio, betonski; baje.*
- **Janes, Kres, Šolar** – 15 besed: *beg, bojevnik, brezposelnost, bes, blaginja, burja, boter, barje; blesteti; biološki, buden, blaten, božanski, bohinjski, bralen.*
- **Janes, GOS, Šolar** – 14 besed: *bajta, bus, babi, budalo, bunda, bran, budilka, bonton; briti; brezvezen, blond, brihten, bodeč; bla.*
- **Kres, GOS, Šolar** – 8 besed: *boginja, bivališče, besedica; bivalen, bujen, beneški; brž, blago* (prislov).
- **Janes, GOS** – 23 besed, na primer: *biznis, bicikel, burek, brzina, baby, blog; barvati, bluziti, brcati; butast, banalen; blazno, bedno, brezveze; bojda.*
- **Kres, Šolar** – 14 besed: *baron, beljakovina, bazilika, bojišče, bilka, bliščč, brazgotina; bleščati; bleščeč, buren, brezskrben, bolezenski, baročen, bežen.*
- **GOS, Kres** – 13 besed: *bas, bala, balet, bukev, bližnjica, breza, blazinica, brošura; bremeniti; brezžičen, botaničen; blizko, belo.*

- **Janes, Kres** – 12 besed: *begunec, banda, bruhanje, branilec, brisanje, borovnica, brezno, bat; belgijski, bučen, blažen, berlinski.*
- **GOS, Šolar** – 8 besed: *bol, brlog, balada, bič, božanstvo; bojevit, babičin; burno.*
- **Janes, Šolar** – 7 besed: *blagor; bogateti, blatiti; bog* (pridevnik), *blazen, briljanten, bran.*
- **Samo Šolar** – 66 besed, na primer: *brdavs, bajka, berač, bodalo, baronica; beračiti, bičati, bogatiti, bliskati; boječ, brezsrčen, borben; brezglavo, bežno, brezupno.*
- **Samo GOS** – 64 besed, na primer: *bot, bek, bulšit, batina; butniti, baviti, bingljati, buljiti, beliti; bazičen, betežen, bliskovit, babji; banalno, butasto.*
- **Samo Janes** – 47 besed, na primer: *bejba, birokracija, butelj, bolha, burka; basati, butati, bankrotirati, bentiti; begunski, bizaren, bebav, blokiran; bolno, brutalno.*
- **Samo Kres** – 23 besed, na primer: *bralka, beljak, barvilo, bluza, blisk, belina; botrovati, blažiti; borzen, brezhiben, baleten, barvit, bombažen; brezhibno; bržkone.*

Med primeri, ki na seznam niso bili vključeni, je mogoče najti besede, ki bi tja po intuiciji (lahko) sodile. Za vključitev se zdijo zlasti relevante besede, ki se pojavljajo v treh od štirih korpusov (par Janes-Kres v kombinaciji s tretjim korpusom), medtem ko korpusni pari in pojavitve v posameznih korpusih na zanimiv način odslikavajo značilnosti jezikovne rabe, ki jo korpusi reprezentirajo. Rezultati kažejo, da izbrana metodologija te specifike (pogovornost na eni strani ali literarnost na drugi) precej uspešno nevtralizira.

Dopolnitev obstoječega seznama bi lahko iskali s tem, da se pri obeh manjših korpusih (Šolar in GOS) upoštevajo daljši izhodiščni nabori besed. Besede, kot so denimo *bencin, baterija, beton,* se v korpusu Šolar pojavljajo, vendar se zaradi nizke pogostnosti oz. uporabljene mere ARF, ki že v izhodišču nizko pogostnost še niža, ne prebijejo na seznam prvih 10.000. Odprto za nadaljnje delo ostaja tudi vprašanje besedotvorno sorodnih besed, od katerih je ena vključena na končni seznam, druga pa ne (*bel – belo; bolan – bolno; blato – blaten; barva – barvati, barvit;* tudi glagolski vidski pari *brcniti – brcati*).

## 4.4. Uporabnost za didaktične namene

V tem razdelku predstavljamo osnovno oceno seznama za potrebe jezikovne didaktike, konkretno učenja slovenščine kot drugega ali tujega jezika. Glede na spoznanja Johansson Kokkinakis in Volodine (2011) smo za oceno pokritosti uporabili tri tematske sklope: imena za dneve tedna ter mesece, dele telesa ter hrano in pijačo.

### 4.4.1. Imena za dneve v tednu in mesece

Pregled seznama pogostih splošnih besed pokaže, da so vanj vključena vsa imena za dneve v tednu in mesece. To je preprost indikator, da je uporabljena metoda uspešna. Besede za dneve in mesece se pojavljajo na različnih mestih seznama, na kar vpliva tudi morebitna homonimnost besede (*sreda*) ali njena enakopisnost z lastnim imenom (*petek*). To velja tudi za druga dva pregledana tematska sklopa in priča o tem, da mesta na seznamu ne gre enoznačno povezovati s stopnjo, na kateri naj bi učeči se besedišče spoznali.

### 4.4.2. Deli telesa

V seznamu smo nato ročno poiskali vse samostalnike, relevantne za učno enoto, ki se posveča delom telesa. Izbira je subjektivna in precej široka (npr. *križ, celica, žila*, ki ne nastopajo v tipičnih učnih gradivih na to temo), vendar dobro pokaže okvire seznama. 56 samostalnikov navajamo padajoče glede na njihovo mesto v seznamu: prvi samostalnik, *roka*, se pojavlja na 207. mestu, zadnji, *veka*, pa na 4.731. mestu seznama: *roka, glava, jezik, srce, noga, telo, obraz, las, prst, kri, koža, usta, organ, uho, nos, hrbet, sklep, solza, zob, možgani, oko, živec, koleno, vrat, čelo, križ, rit, kost, celica, ustnica, peta, mišica, trebuh, želodec, lice, dlan, prsi, bok, rama, grlo, naročje, noht, brada, žila, hrbtenica, jetra, stopalo, palec, gleženj, brk, obrv, ledvica, dojka, ud, veka*.

### 4.4.3. Hrana in pijača

Podobno kot deli telesa se tudi hrana in pijača pojavlja kot eden od najbolj tipičnih tematskih sklopov za učenje tujih jezikov. Z ročnim pregledom smo na seznamu našli 61 samostalnikov (ponovno jih vključujemo precej široko). Za preglednejši vtis tokrat besedišče prikazujemo glede na vsebino: (a) splošno – *hrana, pijača, prehrana, jed, živilo, obrok, hod, juha, priloga, solata, sladica, pecivo, zajtrk, malica, kosilo, večerja;* (b) pijača – *voda, kava, čaj, mleko, sok, alkohol, vino, pivo;* (c) hrana – *kruh, krompir, testenina, riž, pica, sendvič, riba, školjka, meso, jajce, goba, sol, sladkor, olje, moka, žito, med, testo, čokolada, smetana, sladoled, torta, piškot, bonbon, sadje, sadež, jabolko, banana, jagoda, hruška, pomaranča, zelenjava, koruza, zelje, paradižnik, buča, fižol.*

### 4.4.4. Manjkajoče besedišče

Kot je razvidno iz navedenih podatkov, je na seznam zajetega precej relevantnega besedišča, kar gre v določeni meri pripisati tudi frazemom, v katerih se besede za dele telesa in hrano pojavljajo. Po pričakovanjih pa je domet omejen. Besede *pljuča, trup, komolec, maslo, sir, limona* so denimo prisotne na izhodiščnem seznamu, vendar ne v vseh korpusih: samostalnik *pljuča* manjka v podatkih korpusa GOS; *trup* manjka v podatkih korpusa Janes in GOS; vsi ostali primeri pa manjkajo v podatkih korpusa Šolar. Kot je bilo razloženo, to ne pomeni, da se besede v teh korpusih ne pojavijo. Prilagoditev metodologije, da bi pri korpusih GOS in Šolar upoštevala širši nabor izhodiščnega besedišča, bi pokritost izboljšala. V vsakem primeru pa bi bilo za didaktične namene tematsko besedišče treba dopolnjevati tudi na druge načine, kar je skladno z namenom seznama in obstoječimi praksami.

# 5. Sklep in nadaljnje delo

Seznam pogostega splošnega besedišča vsebuje 4.768 lem, opremljenih z besednovrstno oznako in podatkom o prilagojeni frekvenci v vsakem od štirih vključenih korpusov: Kres, GOS, Janes in Šolar. Seznam smo pripravili iz razpoložljivih jezikovnih virov s precej preprosto metodologijo, vendar analize kažejo, da je z vidika vsebnosti primerljiv podobnim rezultatom za druge jezike in ponuja dobro izhodišče za različne namene rabe. Od možnih nadaljnjih korakov za nadgrajevanje seznama z izbranim korpusnim pristopom bi bilo zanimivo preizkusiti zajem dodatnega besedišča s širitvijo izhodiščnega nabora besed iz manjših korpusov (GOS, Šolar), saj izbrana mera ARF že tako nizke frekvence v teh korpusih dodatno reducira in potisne pod prag vključitve.

Rezultati analize kolokacij v korpusih GOS, Šolar in Janes, ki so bile izluščene primerjalno glede na korpus Kres (Pollak & Arhar Holdt, 2015; Rozman et al., 2018; Pollak et al., 2019), pričajo o tem, da bi bila v tem prispevku predstavljena metodologija z ustreznimi prilagoditvami uporabna tudi za pridobivanje večbesednih enot. Na drugi strani bi bilo trenutni seznam lem mogoče dopolnjevati v smer (pogostih splošnih) oblik, saj je za številne namene (npr. strojno oceno berljivosti besedil) identifikacija redkih besednih oblik relevantnejša kot informacija na ravni leme.

Seznam je v nadaljevanju mogoče obogatiti, da bo uporaben kot nabor enostavnega, poznanega ali za učenje temeljnega besedišča: opredeliti značilke predvidene enostavnosti in slednje tudi preizkusiti na strukturiranem vzorcu ciljnih uporabniških skupin, z različnimi pristopi preveriti poznanost besed na različnih stopnjah šolanja, dodatno opremiti in dopolniti seznam z vidika učnih ciljev in enot (ločeno za pouk slovenščine kot prvega in kot drugega/tujega jezika). V nadaljnje delo je mogoče vključiti tudi primerjave z drugimi korpusi, npr. besediščem šolskih učbenikov (Kosem et al., 2019), seznami besed, ki se pojavljajo v lahkem branju (http://www.risa.si/) in podobno.

## 6. Zahvala

Raziskavo je sofinanciral Evropski socialni sklad ter Republika Slovenija, Ministrstvo za izobraževanje, znanost in šport skozi projekt Za kakovost slovenskih učbenikov (Kauč). Raziskovalna programa št. P6-0411 (Jezikovni viri in tehnologije za slovenski jezik) ter št. P2-0103 (Tehnologije znanja), je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna. Delno je bilo delo sofinancirano tudi s strani okvirnega programa Evropske unije za raziskave in inovacije Obzorje 2020 projekt EMBEDDIA (št. proj. 825153). Zahvaljujemo se tudi anonimnim recenzentom za predloge in komentarje.

## 7. Literatura

Edgar Dale in Jeanne S. Chall. 1948. A Formula for Predicting Readability. *Educational research bulletin*, 27(1): 11–28. https://www.jstor.org/stable/1473169.

Ina Ferbežar in Marko Stabej. 2008. Razumeti razumevanje. *Jezik in slovstvo*, 53(1): [15]–31.

Darja Fišer, Nikola Ljubešić in Tomaž Erjavec. 2020. The Janes project: language resources and tools for Slovene user generated content. *Lang Resources & Evaluation* 54: 223–246.https://doi.org/10.1007/s10579-018-9425-z

Iztok Kosem, Eva Pori in Špela Arhar Holdt. 2019. *Keywords and n-grams from a textbook corpus*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1215.

Nataša Logar, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina; Fakulteta za družbene vede.

Senja Pollak in Špela Arhar Holdt. 2015. Identifying corpus-specific collocations: the case of spoken Slovene. V: *Natural language processing, corpus linguistics, lexicography: proceedings*, str. 117–125. RAM-Verlag.

Senja Pollak, Polona Gantar in Špela Arhar Holdt. 2019. What's new on the internetz? Extraction and lexical categorisation of collocations in computer-mediated Slovene. *International journal of lexicography*, 32(2): 184–206, doi: 10.1093/ijl/ecy026.

Tadeja Rozman, Špela Arhar Holdt, Senja Pollak in Iztok Kosem. 2018. Kolokacije v korpusu Šolar. *Jezik in slovstvo*, 63(2-3): [117]–128.

Petr Savický in Jaroslava Hlaváčová. 2002. Measures of word commonness. *Journal of Quantitative Linguistics*, 9: 215–231.

George Spache. 1953. A New Readability Formula for Primary-Grade Reading Materials. *The Elementary School Journal*, 53(7): 410–413. https://doi.org/10.1086/458513.

Tadej Škvorc, Simon Krek, Senja Pollak, Špela Arhar Holdt in Marko Robnik Šikonja. 2019. Predicting Slovene text complexity using readability measures. *Prispevki za novejšo zgodovino*, 59(1): 198–220. http://ojs.inz.si/pnz/article/view/323.

Darinka Verdonik in Ana Zwitter Vitez. 2011. *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

# Literarne pešpoti v povečani resničnosti na področju med Goriško, Tržaško pokrajino in Furlanijo: Konceptualna zasnova, izvedba in trajno vzdrževanje

**Narvika Bovcon,* Aleš Vaupotič†**

* Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
narvika.bovcon@fri.uni-lj.si
† Fakulteta za humanistiko, Univerza v Novi Gorici
Vipavska cesta 13, 5000 Nova Gorica
ales.vaupotic@ung.si

**Povzetek**

Članek predstavi digitalno humanistični projekt izdelave literarnih pešpoti v povečani resničnosti. Na geolokacijah v Gorici, Trstu in Devinu stojijo 3D modeli virtualnih objektov, ki interpretirajo z njimi povezane odlomke iz literarnih del iz regije. Opisani so na eni strani uporabniška izkušnja in po drugi razvoj ter vzdrževanje projekta v mediju povečane resničnosti.

**Literary itineraries in augmented reality in Gorizia, Trieste and Friuli region: Concept, implementation and sustainability**

The paper presents a digital humanities project of building literary itineraries in augmented reality. Interpretations of regional literary texts are reconstructed as 3-D virtual models that are connected to geolocations in Gorizia, Trieste and Duino. On the one hand the aspects of user experience are discussed, and on the other the development and maintenance of an augmented reality project.

## 1. Povečana resničnost kot medij

Povečana[1] resničnost povezuje sodobne (omrežene) mobilne zaslonske naprave s človeškim telesom v fizičnem življenjskem prostoru prek uskladitve perspektivičnega pogleda realnočasovne kamere s 3D izrisovanjem objektov v isti perspektivi. Za to mora naprava poznati svojo orientacijo v fizičnem prostoru – včasih tudi geolokacijo –, da lahko doda digitalno zapisane točke in volumne v pogled na prostor pred kamero. Realni prostor postane geometrično kodiran v aparatu, v zapis v tej kodi pa je mogoče algoritmično posegati z dodatki, t. i. povečavami (*augments*; Pavlik, 2016).[2] Gledano z uporabniškega konca: oseba v prostoru vedno vidi in dojema ogromno količino informacij pred sabo in okoli sebe, vendar pa diskurz povečane resničnosti predpostavlja, da to še ni vse, ampak da je v svetu pri opazovalki še nekaj, česar ne vidi. In ta nevidni in neposredno neobčuteni del resničnosti je tudi pomemben, saj v nasprotnem primeru povečana resničnost ne bi bila smiselna – bila bi zgolj (igriva)[3] uresničitev naključnih možnosti tehnike. Za primerjavo, iz medijske zgodovine vemo, kako je film iz lahkotne sejemske atrakcije postal dominantni jezik za refleksijo sveta v 20. stoletju (Auerbach, 1998), pri tem pa se je seveda kot medij korenito spreminjal (Bonitzer, 1985). Tudi povečana resničnost nosi tak potencial.

Seveda je mogoče prek kontinuuma (Milgram in Kishino, 1994) med zgolj prikazom predmeta na ekranu računalnika in stvarjo iz okolice, ki jo lahko primemo, razlikovati cel spekter mešane resničnosti (Bovcon et al., 2013), vendar se zdi, da je povečana resničnost posebna med 3D prikazi računalniških podatkov. Igranje računalniške videoigre pred zaslonom, podatkovna očala, ki zakrijejo pogled na okolico, ali kateri od potopitvenih sistemov za virtualno resničnost uporabnico izvzamejo iz življenja tukaj in zdaj ter prestavijo v vzporedni svet; podobno kot pri branju romana, ko v domišljiji gradimo fiktivna prizorišča, ali v sobi s 360º panoramo[4] ali v kinu, ko smo negibni kot v Platonovi votlini, ne zavedamo pa se resničnega sveta okoli nas. Nasprotni pol je računalniška manipulacija realnih objektov. Animirane lutke, animatronika, so seveda računalniško vodene, vendar pa jih včasih ne dojemamo kot računalniško umetnost, ampak samó kot animirano plastiko, ker učinki telesnosti pretehtajo.

Pri projektih v povečani resničnosti se zgodi nekaj posebnega. Zavedamo se, da poleg okolice, ki se daje prek čutov, tu, kjer se trenutno nahajamo, obstaja še nekaj, kar

[1] *Augmented*, tudi »obogatena«, »dopolnjena«, »nadgrajena«. »Razširjena« resničnost je manj posrečen prevod, saj se pokriva z razširjenimi mediji iz neoavantgard in razširjenim kinom (prim. razstavo v Tate Modern, https://youtu.be/lDJqA6jOXYw). T. Caudell naj bi izraz uporabil prvič leta 1992 (Pavlik, 2016).

[2] Povečana resničnost praviloma dodaja, 'povečuje', 'krepi intenzivnost' izkušnje. V času svetovne samoizolacije spomladi 2020 se je razkrila potreba po t. i. virtualnih ozadjih v profesionalni komunikaciji, ki jih omogočajo nekatere telekonferenčne platforme; ta ozadja, ki izrežejo govorca iz domačega okolja, nasprotno zmanjšujejo količino podatkov, ki so del videoslike, ohranja pa se za ta sporazumevalni kanal nujno potrebna mimika in gestika govorke. Virtualno ozadje tukaj torej ne dodaja informacij – v nasprotju z zabavno in zasebno neformalno rabo videoklepeta. Virtualno ozadje je prenosni, nevtralen oz. kontekstno primeren prostor – v času, ko pisarne niso dostopne –, kot je prenosljivo in funkcionalno npr. primerno oblačilo.

[3] Pomembna stopnja pri širši uveljavitvi povečane resničnosti je bila igra *Pokémon Go*, ki je leta 2016 prek fizičnega sveta položila še svet igre domišljijskih bitij iz franšize *Pokémon*. Z vidika tega besedila je tukaj relevanten koncept igrifikacije, uporabe igralnega doživljaja v komunikaciji, npr. v učenju, vendar se zdi, da ni nujno povezan s komunikacijo v mediju povečane resničnosti. V opisanem projektu npr. nismo izdelali t. i. resne igre.

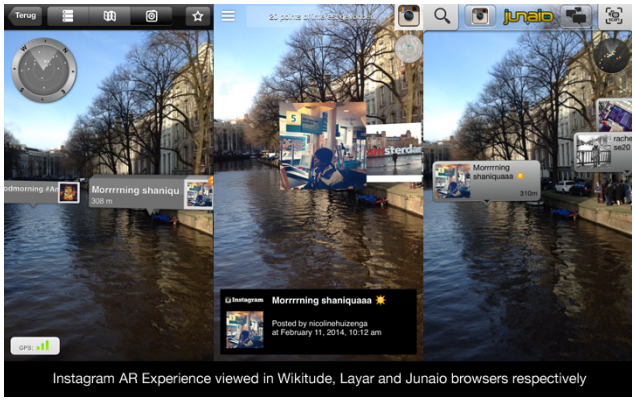
[4] Panorama Mesdag, 1881 (*The Panorama Phenomenon*, 2006).

jo na bistven način dopolnjuje. Poleg projektov na mobilnih kokpitih sodobnih nomadov (Strehovec, 2007), pravzaprav povečane resničnosti v najširšem pomenu, je pomembno povezano področje t. i. lokativna umetnost, kjer je prostor dogajanja tudi del, gradivo umetniškega dela.[5] Vendar povečana resničnost vedno predpostavlja resničnost kot dinamično okolje prav tukaj in zdaj, ki se mu dodajajo interpretacije, zato je potrebna tudi oblika telekomunikacije, ki od drugod vnaša (digitalne) dodatke. Projekt skupine Blast Theory *Ulrike and Eamon Compliant* je nastal za Beneški bienale leta 2009 in je nekakšna gledališka uprizoritev v javnem prostoru, na trgih, mostovih in uličicah Benetk, posebnost pa je to, da igralci govorijo z individualno gledalko-uporabnico po mobilnem telefonu (brez vsakih dodatnih funkcij). Znajdeš se v spektakularnem mestu, kjer ti nekdo po telefonu govori, kaj se okoli tebe dogaja in kaj moraš narediti, kam se premakniti. Ko se telefonska linija prekine, ostaneš delno v zgodbi, z liki, o katerih je bilo govora po telefonu, vmesnika pa ne prepoznaš več – si opazovana? Sprašuješ se, ali obstaja še dodatna gledališka mašinerija? Ni je … izkušnja se sklene v sobi, kjer se preleviš v udeleženko zamejenega ambienta, sobe kot protoodra, ki ni več del odprte resničnosti.[6]

Navedeni primeri se le po naključju osredotočajo na svet umetnosti. Nekateri avtomobili imajo prikazovalnike povečane resničnosti na vetrobranu[7] – so vmes med naglavnim prikazovalnikom in svet prekrivajočim vmesnikom. In kot del interneta stvarí avtomobili sámi – morda zahrbtno – sporočajo v medmrežje, kje smo npr. parkirali.

Obstaja več izvedbeno-tehničnih oblik povečane resničnosti. Delijo se v dva tipa: (a) tiste, ki prek računalniškega vida prepoznavajo orientacijo ploskve-okolja in nanj dodajajo vsebine, npr. na liste v knjigi ali reviji (prim. Bovcon in Vaupotič, 2018; slika 1), na posnetke sebkov …; in (b) tiste, ki uporabljajo žiroskope in pospeškometre, da izračunajo kot snemanja;[8] dve podskupini drugega tipa se ločita po tem, ali so (b.1) povečave v okolici naprave ali pa (b.2) na absolutno določeni geolokaciji – v tem primeru mora prikazovalni vmesnik uporabljati tudi merjenje svojih GPS koordinat.

Slika 1. Knjiga umetnika *À mon seul désir* Vanje Merviča, video se tako rekoč lista; platforma Layar (2017).[9]




Slika 2. Instagram v povečani resničnosti; prikazovalniki.

Podtip b.2 je najtesneje povezan s konkretnim realnim prostorom. Informacije so postavljene na geografske koordinate. Kot se je treba odpraviti na izlet, da lahko naravo ali mesto doživimo, tako je treba zares oditi na določeno fizično lokacijo, da jo izkusimo v prav tam povečani resničnosti. Objava iz leta 2014 omenja integracijo fotografij z Instagrama – tistih z objavljeno geolokacijsko informacijo – v brskalnike za povečano resničnost Wikitude, Junaio in Layar (slika 2) ter ob tem odpira pomembno vprašanje množice prikazovalnih platform (*Wikitude Blog*, 2014). Izkušnja ogleda tega projekta na platformi Layar je, da tako rekoč pokriva ves svet – zaradi razširjenosti Instagrama –, to pa pomeni, da marsikje dejansko deluje, tj. se nekaj prikaže. Hkrati je kot plast resničnosti dovolj zanimiva – kar je v veliki meri posledica t. i. instagramizma,

[5] Prim. podnaslov Pixxelpointa v Novi Gorici leta 2004: *Kreativne skupnosti in lokativni mediji*. Ni vsak projekt, ki upošteva geolokacijo, tudi povečana resničnost, npr. če so podatki uporabljeni za izdelavo kart in prikazovalnikov (projekt *Milk*, Esther Polak in Leva Auzina, 2004, http://milkproject.net.) O povezavah povečane resničnosti z geografskimi informacijskimi sistemi (GIS) prim. Hedley, 2017; Hugues et al., 2011.

[6] Posebno vprašanje je zvočna povečana resničnost. V izkušnjo okolice so dodani npr. le zvoki, pogled pa ni omejen na ekran, ki je pri ogledu likovnih del seveda moteč. Omenjeni projekt Blast Theory v za nas relevantnem segmentu ni vključeval videa (niti računalniškega procesiranja). Vprašanje je morda treba zastaviti, ali video nujno vključuje ekran ali pa morda večpredstavnostni vmesnik lahko deluje tudi brez vizualne komponente?

[7] Steklo optično poveže pogled skozi in odsev projekcije dodatkov, kot na prosojnem naglavnem prikazovalniku. Druga možnost povečane resničnosti gre prek video ekrana, danes na telefonu (Azuma, 1997).

[8] Hedley (2017) ločuje med »otipljivo« (karte, ploskve prostorov) in »mobilno« povečano resničnostjo. Ampak tudi premikanje mobilnih telefonov je otipljivo, pomembneje se zdi, ali je digitalna informacija del prostora (torej b) ali pa premičnega predmeta, ki ji določa lego na njem (a). Problematičen primer povečane resničnosti so analize dogajanja v športnih TV prenosih v živo: nanjo uporabnica ne more aktivno vplivati, ni tukaj, čeprav se dogaja zdaj, saj dodatne črte na igrišču niso manipulabilne (prim. umetniško refleksijo Haruna Farockija *Deep Play*, 2007).

[9] http://www.ung.si/sl/raziskave/raziskovalni-center-za-humanistiko/a_mon_seul_desir

skrbnosti izdelave slik in drugih posebnosti te fotografske kulture (Manovich, 2018).

V letu 2020 je povečana resničnost najpogosteje v obliki aplikacije za iOS ali Android. Velika podjetja, kot je Apple, prevzemajo manjša, Metaio (brskalnik Junaio), pri tem pa obstoječe uporabnice ostanejo brez podpore.[10] Številna podjetja se intenzivno ukvarjajo z možnostmi trženja povečane resničnosti. Poleg tega je ogled povečane resničnosti leta 2011 ali 2019 zelo različno doživetje zaradi strojne opreme, ki danes omogoča bolj tekoč prikaz (Bovcon et al., 2013). Bolj pomembno pa je, da je bil brskalnik Layar okoli leta 2011 včasih celo tovarniško naložen na telefone.[11] V nasprotju z zgodnejšimi unikatnimi tehničnimi rešitvami, kot je *Zlato tele* Jeffreya Shawa (1994), je bilo od 2010 vedno bolj verjetno, da bo kdor koli lahko pogledal povečano resničnost, ne da bi potreboval dodatno strojno opremo in posebne tehnične napotke. To je seveda ključno, saj ni pomembna naprava, ampak resničnost, naš življenjski svet, njegovo prepomenjenje pa šibi, če se ukvarjamo z napravo-vmesnikom.

## 2. Video – refleksivni medij humanistike

Olafur Eliasson, umetnik, ki se mdr. poglobljeno ukvarja s svetlobo, je v maju 2020 (na platfomi Acute Art Deniela Birnbauma) predstavil *Wunderkammer*, serijo predmetov-instalacij v povečani resničnosti, ki vstopajo v okolice uporabnikov. *Redki mormon* je 3D animirana ilustracija ptice – gre za preverjeno strategijo na področju vizualnih učinkov, tj. kopiranje narave, ki je bolj čudežna od izmišljije. Sonce je osrednja tema Eliassonovega dela – npr. *Sončni prijatelj*, in podobno kot *Zavetje severnega sija* se njegovi žarki razlijejo prek vidne okolice (slika 3).

Slika 3. *Solar Friend*, Olafur Eliasson.

Povečana resničnost je pravzaprav video zaprtega krogotoka.[12] Videopodobo spreminjamo, premikamo, se je dotikamo, vedno odzivajoč se na sliko na ekranu. Pri Eliassonu lahko obračamo *Srečni kamen* po zaslonu-prostoru. Odvija se uporabničina reakcija na sliko; s sliko sámo – ne z vnaprejšnjo vsebino slike kot pri filmu – stopamo v dialog, pogoj za to pa je, da tehnosliko takoj vidimo (s filmskim trakom to ni bilo mogoče, saj je bilo treba zaustaviti fotokemično reakcijo in posnetek pripraviti za ogled). Vilém Flusser ugotavlja, da je monitor jedro medija videa, imenuje ga zrcalo s spominom (2002; Vaupotič 2014). Ravno to pa velja tudi za povečano resničnost – posnetek kamere je kot zrcalo, takojšen, mehansko nevtralen, celo vsebinsko trivialen, saj kar je na posnetku-zrcalu, vidimo tudi neposredno; in to podobo s svojim gibanjem spreminjamo. Dodatki izvirajo iz kulturnega spomina, ki so ga odkrili in ga zastopajo avtorji povečane resničnosti.

Flusser odločno zavrne rabo videa samó za ogled filmov: »pri tem [je] karakter videa povsem spregledan« in sklene: »kako to, da video nikoli ni postal filozofska metoda [in] ne služi samo in izključno filozofiji, zakaj se že davno ne vržejo vse filozofske knjige proč in se ukvarjamo z videom« (2009). Povečana resničnost kaže potencial, da postane orodje humanistike, kot so muzeji, arheološka najdišča in učne poti. V okvirih digitalne humanistike pravih zadržkov ni več, je pa potrebna metodološka odprtost. Človek se potopi v intenzivno resničnost in v njej ločuje več drugo na drugo naloženih in zmešanih plasti kulture, družbe in narave. Razumeti jih je mogoče prek pojma kronotopa Mihaila Bahtina; to je semiotična enota prostora, časa in smisla, pa tudi, v drugem pomenu, zgoščanje pomenov in vrednot v prostoru (Vaupotič, 2002). »Naloga je v tem, da bi *stvarno* okolje, ki mehanično deluje na osebnost, prisilili, da bi spregovorilo,« zapiše Bahtin (1999). Kronotopičnost kot vrednota, intenzivni nagovor konkretnega okolja, je rezultat povečane resničnosti, dialoški soobstoj več interpretacij-osmislitev v istem prostoru.

Slika 4. *Topografije spominov*, Gorica in Nova Gorica; posnetek dogodka na trgu iz 1938 sprožen s QR kodo.

## 3. Zapisi kulturne večplastnosti v dvomestju Gorica-Nova Gorica

Plastenje resničnosti je dejstvo, različne skupnosti in različni posamezniki doživljajo iste prostore različno. Ta doživetja zapisujemo in obeležujemo s spomeniki, informacijskimi tablami, tudi grafiti ipd. QR kode so tudi same grafično zanimive – do nedavnega ne zares uporabne, zadnjih nekaj let pa jih ves čas z medmrežjem povezani telefoni preberejo brez dodatnih programov. V Gorici na trgu Travnik najdemo informacijski totem iz projekta *Topografije spominov: Razpršeni muzej 20. stoletja* (2013; slika 4).[13] QR koda odpre spletni video, kjer zagledamo današnji zaspani in umirjeni Gorici kontrastne dogodke na trgu iz preteklosti in slišimo pričevanje udeleženca iz offa. To pa ni edini projekt, ki zapisuje plasti resničnosti v dvomestju Goric. *Let'sGo! Gorizia* (2015) prav tako temelji na QR kodah pritrjenih na tla; obstaja več plasti informacijskih tabel; *A(l)mare in bici* (2018), prenosljive table za kolesarje v regiji; ne nazadnje tudi nekaj *Spotikavcev*

[10] https://techcrunch.com/2015/05/28/apple-metaio

[11] https://www.fastcompany.com/1332828/augmented-reality-here-layar-goes-global-android

[12] Npr. Peter Weibel, *Opazovanje opazovanja: Nedoločenost* (1976). Primer iz kulturne dediščine je soba v Il Vittoriale degli Italiani (posestvo Gabriela d'Annunzia) ob Gardskem jezeru: na več ekranih na stenah prostora se sinhrono predvaja isti video, tako da obiskovalci lahko gledajo video in odzive drug drugega.

[13] http://confine.todm.it

(1991–), del mednarodnega projekta Gunterja Demniga o nacionalsocialističnih deportacijah.

## 4. Literarne pešpoti kot raziskava in zapis večplastne resničnosti

To besedilo je refleksija projekta, v katerem smo izdelali povečano resničnost na področju med Goriško, Tržaško pokrajino in Furlanijo.[14] Nastale so tri literarne pešpoti[15] – projekti v povečani resničnosti, ki temeljijo na geolokacijah zanimivih točk, dostopni so torej le hkrati z doživljanjem fizične okolice. Naš namen je bil okrepiti občutljivost uporabnic za sobivanje na stičišču jezikov, kultur in literatur. Za nastanek geografsko določene povečane resničnosti je potrebno zbrati in oblikovati vsebine, ki smiselno dopolnjujejo dojemanje življenjskega prostora; jih shraniti v podatkovno zbirko; in zasnovati uporabniško izkušnjo prek (čim lažje) dosegljivih tehničnih rešitev.

### 4.1. Literatura na stičišču v furlanščini, italijanščini, narečjih, nemščini, slovenščini …

#### 4.1.1. Berila v prevodih

Projekt je razdeljen na tri povezane segmente: literarnovednega, arhivskega in uporabniškovmesniškega. Prvega je vodila Ana Toroš, nastali sta dve berili v več jezikih o literaturi na stičišču današnje Slovenije in Italije. Digitalna humanistika vključuje specialistična znanja, kar pomeni, da tak projekt ni izvedljiv brez sodelovanja strokovnjaka na področju literature v regiji (Toroš, 2019). Tudi temeljna struktura novomedijskega objekta kot povezava množice vmesnikov (oz. prilagodljivega, variabilnega vmesnika) in podatkovne zbirke zahteva arhiv, ki je ločen od vmesnika (bodisi na ravni vmesnika kot posredujoče metaforične podobe bodisi tehnične rešitve; Manovich, 2001). Povečana resničnost je novomedijski objekt, poglobljeno poznavanje kompleksnih kulturno-jezikovnih povezav regije, kjer je postavljena, pa je lahko eden od njegovih dveh temeljev, tj. arhiv. Okvirni projekt EDUKA 2 je vzpostavljal čezmejno izobraževalno infrastrukturo v furlanski, slovenski in italijanski skupnosti, del projekta pa je predstavljal sklop *Skupne učne enote o manjšinski literaturi ter literaturi obmejnega območja*. Nastali sta dve berili, eno namenjeno osnovnim šolam in drugo za srednje šole; prvo se osredotoča na poezijo in goriško-furlanski del področja raziskave, drugo pa na pripovedno prozo in na Trst.[16] Z jezikovnega vidika je pomembno, da je literarna kultura prostora obstajala in še nastaja vsaj v štirih jezikih, zato so bili učbeniki prevedeni v italijanščino in slovenščino za Tržaško pokrajino ter še dodatno v furlanski jezik za Goriško in Furlanijo, kjer je ta jezik v uporabi.

Slika 5. Kip Jamesa Joycea na mostu v Trstu; prek vmesnika literarne pešpoti na platformi Layar; v sredini je povečava, ki se nanaša na delo Borisa Pahorja (1959)

#### 4.1.2. Kanonska literarna dela v Trstu in Devinu

Drugačen vstop v literarne usedline ob območjih kulturnih žarišč[17] je mogoč prek prizme literarnega kanona, ki omogoča orientacijo v tistem, kar predvsem kot problem (ali goethejevski projekt; Birus, 2004) označuje izraz svetovna književnost. Tak je dostopen širšemu krogu literarnih znanstvenikov, zato smo prek tega vidika lahko pilotno razvili prvo pešpot. V tem kontekstu Furlanija in Trst nista brez zanimivih poudarkov, saj so nekatera najpomembnejša dela evropske literature nastajala tukaj. Npr. James Joyce je svoj slavni roman *Ulikses* (1922) pisal tudi v Trstu – predstavljajmo si, da je med natančnim opisovanjem hoje po Dublinskih ulicah pravzaprav stal na tržaških (slika 5) … to ponuja zanimivo bralno izhodišče.

[14] Projekt EDUKA 2, glej Zahvale (http://www.eduka2.eu).

[15] Prim. tudi druge slovenske literarne pešpoti, tiskane in spletne vodnike, kot sta *Literarne poti Ljubljane* in transverzala Slovenska pisateljska pot (zemljevid, vodnik po literarnih pokrajinah Slovenije, spominske table in klopi, razgledna počivališča, spominska knjižica z žigi, kratki biografski filmi, organizacija dogodkov Društva slovenskih pisateljev, načrtovan je bil spletni portal). Na *Wikipediji* je navedenih 38 slovenskih literarnih poti (https://sl.wikipedia.org/wiki/Slovenske_literarne_poti), številne poti in točke so vnesene v Geopedijo (glej Vsebina, Dodatni sloji: Literarne poti, http://www.geopedia.si/#T105_L16042, in Slovenska pisateljska pot …). Del našega projekta se prostorsko sklada z Rilkejevo potjo, vendar je brez rekreativno-pohodniškega vidika; okrepljeno je osveščanje o nehomogenosti in dialoškosti življenjskega okolja ter poziv k branju del.

[16] http://sabotin.ung.si/~avaupotic/eduka2/berila

[17] Prim. koncept t. i. marginocentričnih mest, tj. večetničnih vozliščnih mest, kot je npr. Trst (Cornis-Pope in Neubauer, 2002).

Mogoče je reči, da je za dojemanje tudi nekaterih nemško pišočih avtorjev posebej primeren prav prostor med Trstom in Gorico. Npr. Rainer Maria Rilke je prve tri od *Devinskih elegij* (1923) med 21. 1. in marcem 1912 pisal na današnji Rilkejevi poti, ki se ob spektakularnem razgledu vije po pečinah med Sesljanom in Devinom. Zato smo eno od treh pešpoti postavili tja (slika 6). Poleg treh 3D skulptur iz začetnih verzov elegij, postavljenih nad obalo in morje, je to pešpot s strani kopnega dopolnjeval tudi interpretacijski izziv – referenca na delo Theodorja W. Adorna *Žargon pravšnjosti* (1964), kjer ta umetnostni teoretik in kritik poveže elegije s tako rekoč zlobnostjo zlagane avtentičnosti.

Cilj te pešpoti je bil spodbuditi uporabnico, da se peš sprehodi po naravi, se ustavi in bodisi kar na prostem ali pa kasneje v miru prebere Rilkejeve pesmi, pa tudi da naredi prvi kritični korak v refleksiji bralnega doživljaja, kot je vpisan v zgodovino recepcije in ki je zdaj že standardni del literarne vede. Videti je, da povečana resničnost v tem primeru ni literarnoraziskovalna na ravni novih interpretacij samih kanonskih del, ampak na telesno-doživljajski povezavi Rilkejeve izkušnje pisanja in sestavljanja pesmi z današnjo bralno-uporabniško na istem mestu prek (končnih verzij) besedil, ki so (bila) pred očmi Rilkeja in nas.

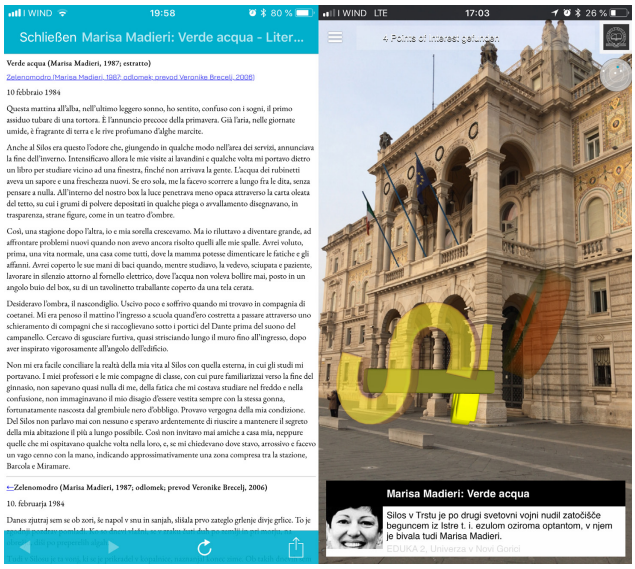
Slika 6. Druga in tretja od *Devinskih elegij*; objekt, ki nosi sklic na Adornovo interpretacijo pesmi (brskalnik Layar).

## 4.2. Digitalni zapis

V sodelovanju z informatiki sta Narvika Bovcon in Aleš Vaupotič zasnovala integracijo računalniškega zapisa in uporabniškega vmesnika (drugi in tretji segment projekta, glej zgoraj): realizacijo spomenikov v povečani resničnosti.[18] Prvi korak je izbira platforme in standardov.

### 4.2.1. Povečana resničnost na platformi Layar

Projekt se je vključeval v obstoječo serijo medijsko-kritičnih raziskav povečane resničnosti na platformi Layar (Bovcon et al., 2013), ki je v letu 2018 dosegla zelo visoko stopnjo dosegljivosti. Ob predstavitvah projekta (učiteljem osnovnih in srednjih šol) po Sloveniji in Italiji je bilo dovolj, da se je obiskovalkam povedalo ime aplikacije in ključno besedo za iskanje (ime projekta ali pa kaj podobnega), in že je tako rekoč vsaka gledala in rokovala s povečano resničnostjo na svoji napravi. To stanje je seveda drugačno, kot je bilo opaženo v Ljubljani na festivalu industrijskega oblikovanja leta 2012: profesionalne oblikovalke so seveda imele najnovejše in najzmogljivejše naprave in jim je bila zato povečana resničnost načeloma dostopna. Nikakor pa to v resnici ni veljalo za veliko večino tistih, ki so bile zgolj občinstvo leta 2012 – treba je bilo zagotoviti napravo in pomoč pri uporabi. Kot rečeno, danes teh težav v tej regiji v izobraževalni skupnosti ni več.




Slika 7. Roman *Zelenomodro*; izvirnik in prevod na spletni strani, ki je povezana s 3D objektom iz črk besede »silos«, ki so postavljene v zgradbo v Trstu, kjer so živeli optanti po 2. svetovni vojni – tudi pisateljica –, in v času trajanja projekta, zadnjih nekaj let, migranti nelegalno.

Za projekt je ključnega pomena to, da so tako vsa večpredstavnostna gradiva (3D objekti, videi, spletne strani z besedili idr.) kot tudi informacije o postavitvi 3D skulptur v realni prostor Tržaške in Goriške na strežnikih (in v repozitoriju) Univerze v Novi Gorici (izbrana je bila licenca CC-BY), kar omogoča večkratno uporabo gradiv v novih projektih.[19]

Poleg devinske sta nastali tudi pešpoti v Trstu in Gorici. V Trstu so jo določale štiri skulpture v povečani resničnosti: poleg del Pahorja in Marise Madieri (sliki 5, 7), tudi objekti za *Materado* (1960) Fulvia Tomizze in roman *Kreda in hijacinte* (2006) Irene Žerjal (slika 8). Vsaka povečava je vodila na spletno stran z odlomkom (slika 7), ki se je prikazala kar v brskalniku za povečano resničnost Layar, hkrati pa je imela aplikacija tudi pogled zemljevida, kar je omogočilo uporabnici, da se sprehodi do lokacij

[18] 3D modeli in literarne pešpoti: Narvika Bovcon, Aleš Vaupotič, programiranje: Robert Valentak (strežnik Layar), Borja Bovcon (aplikacija za Android, igralno okolje Unity).

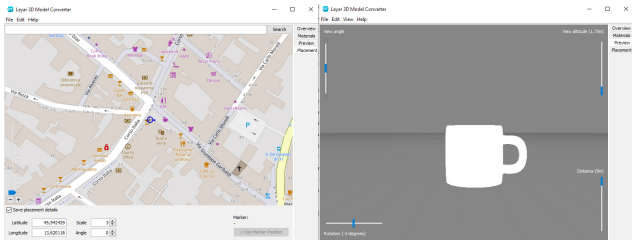
[19] http://sabotin.ung.si/~avaupotic/eduka2

iz literarne imaginacije Trsta (Toroš, 2011). Pot po Trstu, kot tudi ostale, smo večkrat prehodili po možnih poteh, ki so implicirane v razporeditvi in pogledih na objekte v povečani resničnosti (velikost, s tem povezana vidnost), da bi ocenili, kako se bo uporabnica vedla ter da bi pri tem dosegli najboljši učinek povečave mestne resničnosti. Mogoče se je sprehoditi med črkami besede »asimilacija« po pomolu Audace/San Carlo, v prikazanem odlomku iz romana Irene Žerjal je govora tudi o gradu Miramar, ki se ga vidi prek zaliva (slika 8). Postavitev objekta je najbolj enoznačna v primeru Pahorjevega *Kresa v pristanu*, ki opisuje požig slovenskega Narodnega doma med svetovnima vojnama.

Slika 8. Roman Irene Žerjal in del z rušilca Audace z modrim grafitom – trk več zapisov kulturnega spomina skozi prizmo-ekran pešpoti, tj. povečane resničnosti.

V primeru romana Marise Madieri je dostop do nekdanjega silosa pravzaprav prepovedan, gre za okolico železniške postaje, kjer je gibanje omejeno. Črkovni 3D objekt je primerno velik, da spodbuja domišljijo na ozadju pregrajenega pogleda na staro pristanišče. Objekt »odhod«, povezan z *Materado*, stoji nad morjem in je peš nedostopen.

Pešpot po Gorici sestavlja sedem lokacij: dve furlanski besedili, Piera Paola Pasolinija in Celsa Macorja, eno v italijanščini, dve v nemščini in dve v slovenščini.[20] V Goricah, kjer je sedež Univerze v Novi Gorici, smo v delo na povečani resničnosti vključili tudi študente (slika 9), načrtovali pa smo tudi delavnice za srednješolce.



Slika 9. Vmesnik za prevajanje standardnih 3D računalniških zapisov oblik v format za prikazovalnik Layar in vmesnik za postavitev objekta v prostor; model Aneja Žagarja za besedilo *Marietta. Ein Roman aus Görz*.

Pri vseh teh povečanih resničnostih je bil v prvi vrsti cilj, da obiskovalka začuti kompleksno povezanost kultur in jezikov, kot se kaže v literaturi, ki je nastala v regiji. Namen ni bil zabavati z igrivo očarljivostjo tehnike – niti ni šlo samó za likovno učinkovitost videopodobe –, ampak smo hoteli prepričati bodoče bralke, da posežejo po dejanskih literarnih delih, romanih, pesniških zbirkah. Enako je bilo pri načrtovani delavnici za srednješolce: v ospredju je razmislek o literarnih podobah ljudi, njihovih sobivanj v prostoru prek izdelovanja tudi lastnih zapisov, ki ostanejo del okolice … kar so navsezadnje že ves čas bili, saj literatura in naše ideje ne nastajajo v praznini, ampak so dialoški odziv na življenje v prostoru.

### 4.2.2. Menjava platforme za povečano resničnost – izzivi za trajnostno zasnovo pri izvedbi povečane resničnosti

Konec leta 2018 je storitev Layar – zaradi stečaja krovnega podjetja Blippar – prenehala delovati (Palmer, 2019). Trajnostna zasnova povečane resničnosti je velik izziv. Kot smo videli pri omembah večanja dostopnosti do platforme Layar – zaradi razširjenosti zmogljivejših naprav, ne nazadnje tudi zaradi vedno pogostejše uporabe večfunkcionalnih prenosnih naprav za delo in razvedrilo in zato blažitve strahu pred zapleteno tehniko – je fizično-tehnična rešitev najtesneje povezana z načini rabe strojne in programske opreme v konkretnih okoljih. Zagotoviti delovanje povečane resničnosti pomeni, da moraš priskrbeti napravo, to pa tako, da jo npr. dejansko izročiš uporabnicam, vendar v projektih, kot so naše pešpoti, to ne bi prišlo v poštev, saj ni informacijskih pisarn ali kontaktnih točk v fizičnem prostoru. Layar je zares omogočal, da prek te platforme naredimo povečano resničnost dostopno širokim množicam na območju med Slovenijo in Italijo brez vzpostavljanja nove hardverske infrastrukture.

Tok podatkov je bil takšen: uporabnica je prek App Store ali Google Play naložila majhno aplikacijo, nato je

[20] Alojz Gradnik *Jesenski večer v Medani*; Celso Macor *Da ponta di chista mont*; Carlo Michelstaedter *All'Isonzo*; Pier Paolo Pasolini *Dedica*; Janez Povše *Meja*; Ivo Volkmar (psevdonim) *Marietta: Ein Roman aus Görz*; Franz Xaver Zimmermann *Mittagskonzert im Giardino Pubblico*.

ta aplikacija sporočila centralnemu strežniku platforme Layar geolokacijo (in npr. nastavitev radija, kjer se povečave prikažejo), ta je potem poslal poizvedbo drugemu strežniku na Univerzi v Novi Gorici, ki je na podlagi posredovanih koordinat posredoval informacije o zanimivih točkah (*point of interest*, seznam *hotspots*) in njihove opise ter vse večpredstavnostne vsebine (3D objekte za prikaz prek posnetka kamere v usklajeni perspektivi, 2D slike v prostoru in kot del vmesnika, videe, besedila itn.; v obliki povezav). Torej, vsa vsebina, lokacijske koordinate povečav, modeli, besedila itn. je bila na strežnikih Univerze (kjer smo postavili ustrezno strežniško storitev) – to je bil eden izmed odločilnih kriterijev, da izberemo Layar, saj smo predvidevali, da, če platforma izgine, lahko izgubimo samo prikazovalnik, ne pa celotne povečane resničnosti. In to se je v resnici tudi zgodilo, izpadel je vmesni strežnik, ki je posredoval med podatki na našem strežniku in brskalniki na telefonih uporabnikov.

Migracija med prikazovalniki ni bila avtomatična, šlo je pravzaprav za rekonstrukcijo podobne platforme. Nova aplikacija je narejena z igralnim pogonom Unity in knjižnico Vuforia. Pri prikazovanju pogleda na virtualni objekt uporablja podatke o lokaciji iz GPS enote v telefonu uporabnice, tako izračunava oddaljenost do virtualnega objekta, ter podatke žiroskopa za zorni kót. V bazi (tokrat na uporabničinem telefonu) so shranjeni 3D virtualni objekti in pripadajoči opisi. Rezultat je torej aplikacija za operacijski sistem Android, ki prikazuje vse prej določene vsebine. Koordinate, opisna in vmesniška besedila ter povezave na slike so bile za namene prenosa zapisane v CSV datoteki[21] in potem vnesene v nov prikazovalnik povečane resničnosti. 3D modeli so morali biti glede formata datoteke seveda prilagojeni novemu programu.

Slika 10. Druga verzija povečane resničnosti na platformi Android-Unity-Vuforia; Janez Povše *Meja* (2014).

#### 4.2.3. Primerjava dveh verzij iste povečane resničnosti

Razlika je, da je Layar deloval prek internetne povezave, medtem ko rešitev v Unity deluje kot samostojna aplikacija (slika 10). To je hkrati prednost in slabost: prednost, ker deluje, če internet ni dosegljiv, slabost, ker se pojavijo omejitve pri količini objektov (oz. obratno, velikosti same aplikacije), saj so vse informacije (razen spletnih strani z literarnimi odlomki, ki so iste v obeh verzijah) v sami aplikaciji na telefonu ali tablici. Seveda je dejstvo, da uporabnica ni povezana z internetom posebna uporabniška pozicija, nepovezanost onemogoča npr. branje celotnih del, iskanje dodatnih informacij po svetovnem spletu …

Druga bistvena razlika je, da je Layar deloval na obeh glavnih mobilnih platformah, Unity pa smo v danih pogojih usposobili le v Androidu. Seveda bi kdo rekel, da je to zgolj tehnično vprašanje, aplikacijo se lahko prevede in prilagodi za iOS, vendar pa bi s tem zgrešili poanto, govorimo namreč o trajnosti novomedijskega objekta, kjer pomanjkanje finančnih sredstev in institucionalnih podpor za njegovo vzdrževanje lahko pomeni tudi kratko malo izginotje (tudi če ne pride do velikih strateških napak pri izbirah standardov).

Tretja razlika je v politiki pristopov na področju uporabniških vmesnikov: Layar je bil storitev, izvedba v Unity pa vsaj delno stopa po poti odprtokodnosti. Tržno usmerjene rešitve, kot je bil Layar, so precej bolj robustne in seveda vključujejo tehnično podporo, ki je v *ad hoc* rešitvi ne more biti – maksimum fleksibilnosti se praviloma povezuje z ranljivostjo in krhkostjo, netrajnostjo, še posebej na področju 3D svetov in videa. Npr. iPhone 6S Plus (2016) je prikazoval 3D objekte prek geometrijskih koordinat fizičnega prostora v brskalniku Layar precej bolj gladko kot naša rešitev na novejšem telefonu, prav tako najvišjega razreda, Samsung Galaxy Note 10 Plus (2019).

Naš namen je bil, da je povečana resničnost na internetu povezana z geografskimi koordinatami. Človek mora biti tam, kar je seveda velika zahteva, zato pa mora delovati na napravi, ki jo ima pri sebi ves čas. Če je bila namestitev Layarja izrazito preprosta, je bila situacija z na Unity temelječo aplikacijo podobna kot po letu 2010 – uporabnikom je bilo treba pomagati, da naložijo aplikacijo, medtem pa so npr. ostali v občinstvu zgubljali pozornost. Lahko sklenemo, da je povečana resničnost med še posebej ranljivimi digitalnimi vsebinami, razlog za to pa je, da bi morala delovati zunaj kontroliranega okolja, tako rekoč brez posebnega – posebej opaznega – vmesnika, samo s tistim, kar ima vsak že tako ali tako s seboj. Potemtakem se zdi, da je pri povečani resničnosti izbira uveljavljene storitve, naprav, ki in kot so že v rabi, prava.[22]

## 5. Literarna veda in umetniško raziskovanje

### 5.1. Video *TrsT_La Sufia la bora* Vanje Merviča[23]

Naše pešpoti so krepitev doživljanja prostora, kot se je že izoblikovalo v literaturi. Vstopanje v povečano resničnost pa povzroči doživetje – npr. Trsta – danes, ki je zopet lahko kristalizirano v umetniškem delu. V okvirih projekta EDUKA 2 smo v letu 2019 izvedli tudi naročilo umetniku in nastal je video trak na podlagi tržaške pešpoti – s tem smo v projekt pritegnili tudi posebno obliko raziskav, ki se odvija v umetnosti in ki je zelo dragocena tudi za humanistiko. Vsak od nas ima svojo interpretacijo sveta in osebno vizualizacijo literature. Likovni umetnik zna svojo vizijo posredovati drugim in jo nadgraditi v okvirih svojega umetniškega raziskovanja (slika 11).

[21] Ko smo arhivirali koordinate in druge gradnike povečane resničnosti (http://sabotin.ung.si/~avaupotic/eduka2 ), smo projekt kartirali tudi v *Google maps*, https://www.google.com/maps/d/viewer?mid=1pNDeF8nwrthzbbsGjLqfSK6dGWpFYHl8&hl=en&usp=sharing.

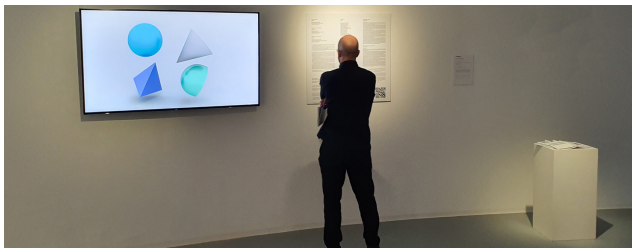
[22] Pri povečani resničnosti je pomemben vidik uporabniške izkušnje ritem usklajevanja slike realnočasovne kamere in računanja izrisa 3D koordinat prek slike – lahko bi ga imenovali ritem povečane resničnosti, ki je ali bolj tekoč, skoraj brez zamikov obeh 3D koordinatnih sistemov, ali pa zatikajoč. Gre za artefakt, ki ga najbrž ne bo mogoče v celoti odstraniti, to pa tudi najbrž ni potrebno, saj gre za element tega videojezika.

[23] https://vimeo.com/333978654, 2019, video trak, 20'45".

Slika 11. Video Vanje Merviča, štiri ekranske slike.

### 5.2. Literarne pešpoti kot novomedijska instalacija na festivalu Pixxelpoint 2019

Na 20. Mednarodnem festivalu sodobnih umetnosti Pixxelpoint v Novi Gorici in Gorici so bile pešpoti v povečani resničnosti predstavljene kot videoinstalacija v Mestni galeriji Nova Gorica, kot povečana resničnost v javnem prostoru in kot projekt hibridne umetnosti, ki povezuje umetniško in znanstveno raziskovanje z aktivnim udejstvovanjem v družbi (slika 12).



Slika 12. Instalacija v Mestni galeriji Nova Gorica; 4K video Narvike Bovcon in Aleša Vaupotiča, 3D objekt za Pasolinijevo furlansko pesem iz *Poesie a Casarsa* (1942).

***

Digitalna humanistika rahlja meje med disciplinami ved in znanosti ter tehnike. Gradi stičišča proti izobraževanju, medijem, oblikovanju in umetnosti. Algoritmi umetne inteligence so mehanski, človeško razumevanje je ustvarjalno, oboje se trenutno meša (Manovich, 2019). Pot naprej je negotova, kar je pravzaprav nekaj dobrega, pomeni možnosti, ki jih digitalni humanisti pregledujemo prek širokih interdisciplinarnih povezav. Povečana resničnost je medijski konglomerat, bodisi dodatek površinam ali pa človeškemu orientacijskemu prostoru zgoraj-spodaj-levo-desno-spredaj-zadaj-blizu-daleč. Postavitev na enoznačne geolokacije ponovno aktivira naravno okolje, ki je zaradi neustavljive želje po napredovanju v globalni družbi in kulturi ogroženo, vrnitev v resnično okolje, naravno in kulturno krajino – nam bližnjo –, pa nam omogoča, da smo bolj občutljivi za spremembe v njej.

## 6. O projektu

Projekt Univerze v Novi Gorici, pripravljen v sklopu EDUKA 2 *Čezmejno upravljanje izobraževanja*, Program sodelovanja Interreg V-A Italija-Slovenija s sredstvi ESRR (2017–19). Delo na raziskovalnem programu *Zgodovinske interpretacije 20. stoletja*, ki ga financira ARRS.

## 7. Literatura

Theodor W. Adorno. 1972. *Žargon pravšnjosti*. CZ.
Erich Auerbach. 1998. *Mimesis*. LUD Literatura.
Ronald T. Azuma. 1997. A Survey of Augmented Reality. *Presence*, 6(4):355–85.
Mihail Bahtin, 1999. *Estetika in humanistične vede*. SH.
Hendrik Birus. 2004. Goethes Idee der Weltliteratur. V: *Goethezeitportal*. http://www.goethezeitportal.de/db/wiss/goethe/birus_weltliteratur.pdf.
Pascal Bonitzer. 1985. *Slepo polje*. SH, Ljubljana.
Narvika Bovcon, Aleš Vaupotič. 2018. Artistic Visualizations and Beyond. V: *JTDH 2018*, str. 18–23, Ljubljana.
Narvika Bovcon, Aleš Vaupotič, Bojan Klemenc, Franc Solina. 2013. »Atlas 2012« Augmented Reality. V: *SouthCHI 2013*, str. 477–96, Maribor.
Marcel Cornis-Pope in John Neubauer. 2002. *Towards a History of the LiteraryCultures in East-Central Europe*. ACLS Occasional Paper, 52.
Vilém Flusser. 2002. *Digitalni videz*. Študentska založba.
Vilém Flusser. 2009. *Kommunikologie weiter denken*. Fischer.
Nick Hedley. 2017. Augmented reality and GIS. *Comprehensive Geographic Information Systems*, Elsevier, str. 355–68.
Olivier Hugues, Jean-Marc Cieutat, Pascal Guitton. 2011. GIS and Augmented Reality. V: *Handbook of Augmented Reality*, str. 850, Springer.
Lev Manovich. 2019. *Estetika umetne inteligence*. Mestni muzej, Zavod Basic, Ljubljana.
Lev Manovich. 2018. *Instagram and Contemporary Image*. http://manovich.net/index.php/projects/instagram-and-contemporary-image.
Lev Manovich. 2001. *The Language of New Media*. MIT.
Paul Milgram in Fumio Kishino. 1994. A Taxonomy of Mixed Reality Visual Displays. *IEICE Transactions on Information and Systems*, E77-D(12):1321–29.
Sebastijan Pregelj in Gašper Troha. 2010. *Literarne poti Ljubljane*. Študentska založba.
Maija Palmer. 2019. Goodbye Blippar, welcome back Layar? *Sifted*. https://sifted.eu/articles/goodbye-blippar-welcome-back-layar.
2006. *The Panorama Phenomenon*. Panorama Mesdag, Haag.
John W. Pavlik. 2016. Augmented Realty. V: *The International Encyclopedia of Communication Theory and Philosophy*, str. 144–9, John Wiley & Sons.
Janez Strehovec. 2007. *Besedilo in novi mediji*. LUD Literatura.
Ana Toroš. 2011. *Podoba Trsta in Tržaškega v slovenski in italijanski poeziji prve polovice 20. stoletja*. Založba Univerze v Novi Gorici.
Ana Toroš. 2019. Teaching minority literature: the case of Trieste. *Razprave in gradivo*, 83:83–93.
Aleš Vaupotič. 2002. On the Problem of Historical Research in Humanities. *Logos*, 2(3). http://kud-logos.si.
Aleš Vaupotič. 2014. Smisel videa. *Video et gaudeo*, str. 2–4, ZDSLU, Ljubljana.
2014. OGC, Wikitude, Layar, and Metaio invite Mobile World Congress attendees to AR Interoperability Demo. *Wikitude Blog*. https://www.wikitude.com/ogc-wikitude-layar-metaio-invite-mobile-world-congress-attendees-ar-interoperability-demo.

# The ssj500k Training Corpus for Slovene Language Processing

**Simon Krek[1,2], Tomaž Erjavec[1], Kaja Dobrovoljc[1,2], Polona Gantar[2], Špela Arhar Holdt[2], Jaka Čibej[1,2], Janez Brank[1]**

[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2]Faculty of Arts, University of Ljubljana, Slovenia
{janez.brank, kaja.dobrovoljc, simon.krek, tomaz.erjavec}@ijs.si
{spela.arharholdt, jaka.cibej, apolonija.gantar}@ff.uni-lj.si

**Abstract**

This paper presents recent developments and the content of the ssj500k training corpus, the largest and most widely used open-source collection of training data for Slovene language processing, which has been manually annotated with respect to segmentation, tokenisation, lemmatisation, JOS morphosyntax and dependency syntax, Universal Dependencies, semantic role labelling, named entities and verbal multi-word expressions. After a short history of the development of the corpus, we give an overview of the dataset as a whole, and the details of each annotation layer, including a survey of existing natural language processing tools that used it for training. Most ssj500k annotations were carried out using the dedicated Q-CAT querying-supported corpus annotation tool, which is also presented, and the directions for future development of the corpus are discussed.

## 1. Introduction

Training corpora, i.e. manually annotated language datasets, are essential to the development of natural language processing tools based on supervised machine learning. For Slovene, a South Slavic language with approximately two million speakers, the ssj500k training corpus (Krek et al., 2019) represents the largest collection of such manually annotated training data to date, a result of more than two decades of continuous development.

This first began within the multilingual MULTEXT-East project and its spin-offs (Dimitrova et al., 1998; Erjavec, 2004), providing the pivotal language resources for Slovene morphosyntactic annotation, including the specifications, an inflectional lexicon, and the "1984" morphosyntactically annotated corpus, which was later also used as the starting point for the first Slovene dependency treebank (Džeroski et al., 2006). However, given the limitations in size and diversity of the "1984" corpus, the "Linguistic annotation of Slovene" (Jezikoslovno označevanje slovenščine: JOS)[1] project (2007-2009) aimed to fill this gap by producing a balanced and representative set of text samples, the jos100k corpus (Erjavec and Krek, 2008), manually annotated with lemmas, morphosyntax and dependency syntax (Erjavec et al., 2010), all based on a revised and updated set of MULTEXT-East/JOS annotation guidelines. In the consecutive "Communication in Slovene" (SSJ)[2] project (2008-2013), additional 400,000 words were added to the jos100k corpus to form the ssj500k corpus with JOS/MULTEXT-East morphosyntactic annotation, as well as partial annotation for JOS dependency syntax and named entities (Holozan et al., 2008).

In subsequent years, the annotation of the ssj500k corpus has been incrementally developed within various other projects, such as the "Semantic Role Labelling in Slovene and Croatian" bi-lateral project (Gantar et al., 2018b), the PARSEME verbal multi-word expression annotation (Gantar et al., 2018a), Universal Dependencies for Slovene (Dobrovoljc et al., 2017) and Janes named entity annotation (Fišer et al., 2018), producing several new layers of grammatical annotation. The ssj500k corpus thus represents the largest, richest and most widely used training and testing dataset for Slovene language engineering research and development.

However, despite this pivotal role, the dataset as a whole lacks systematic and exhaustive documentation, not only in terms of its content and recent developments, but also in terms of the infrastructure that surrounds it, such as its formats and the querying and annotation tools. The aim of this paper is to remedy this and provide an overview of the latest ssj500k release, version 2.2 (Krek et al., 2019), by presenting its annotation layers and the state-of-the-art tools that employ them (Sections 2. and 3.), describing the surrounding infrastructure (Sections 4. and 5.) and discussing immediate and long-term future work (Section 6.).

## 2. Overview of the corpus

The ssj500k corpus (Krek et al., 2019) is a balanced collection of text samples from the FidaPLUS corpus of written modern Slovene (Arhar and Gorjanc, 2007), which includes fiction, non-fiction and periodical documents dating from 1990–2000. In total, the corpus contains 586,248 tokens belonging to 27,829 sentences inside 8,137 paragraphs sampled form 1,655 documents. However, due to temporal and financial constraints, not all annotation layers have been applied to the corpus in its entirety.

As summarised in Table 1, the entire corpus has been manually segmented into sentences and tokens and annotated for lemmatisation and JOS/UD morphosyntax. In contrast, other annotation layers cover only parts of the corpus, ranging from approximately half of the corpus (e.g. multi-word expressions) to approximately one fifth of the corpus (e.g. semantic roles). With the exception of the intermittent UD syntax (Section 3.3.), all layers cover a con-

[1]http://nl.ijs.si/jos/
[2]http://www.slovenscina.eu/

| Level | Tokens | Senten. | Docs |
|---|---|---|---|
| Segmentation | 586,248 | 27,829 | 1,655 |
| lemmatisation | 586,248 | 27,829 | 1,655 |
| JOS morphosyntax | 586,248 | 27,829 | 1,655 |
| UD morphology | 586,248 | 27,829 | 1,655 |
| JOS syntax | 235,864 | 11,411 | 617 |
| UD syntax | 140,670 | 8,000 | 581 |
| Semantic roles | 112,048 | 5,501 | 228 |
| Named entities | 194,637 | 9,488 | 498 |
| Multi-Word Expressions | 280,522 | 13,511 | 754 |

Table 1: Size of ssj500k annotation layers.

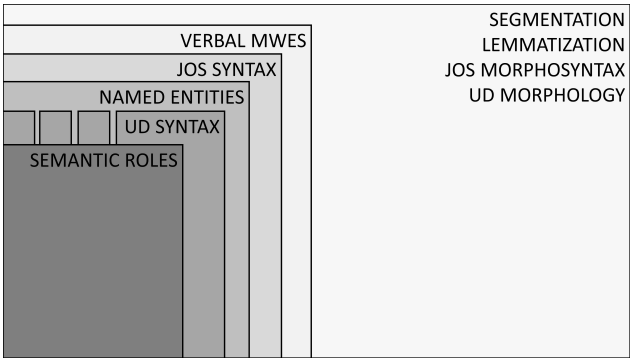



Figure 1: Illustration of nested annotation layers in ssj500k.

tinuous sequence of sentences starting at the beginning of the corpus, and can thus be considered as nested subsets, as illustrated in Figure 1.

# 3. Annotation Layers

## 3.1. Segmentation and tokenisation

Manual segmentation and tokenisation in ssj500k corpus follows rules implemented in the Obeliks4J tokeniser for standard Slovene.[3] The rule-based tokenisation and segmentation system was devised with the view that anyone can unambiguously predict how sequences of characters would be split into sentences and tokens in processed texts.

**Tokenisation**: Whitespace is the principal separator for tokens. Sequences of words that can be written both with or without space (e.g. *kdorkoli*, *kdor koli* "anybody, any body") follow the same principle. During tokenisation, the text is divided into sequences belonging to two categories: words (W) are alphanumeric strings, while the other category includes punctuation and symbol characters (C). C tokens are recognised on the basis of a predefined list of punctuation and symbol characters included in the tokeniser. C tokens may include only one punctuation or symbol character. Sequences of two or more C tokens (e.g. *?!*) are treated as sequences of separate tokens. If a string of alphanumeric characters between two spaces includes C characters, it is split into separate tokens (e.g. *AC/DC* will be split into three tokens *AC, /, DC*), with some exceptions, such as:

- Apostrophe becomes part of a W token if it is used without space on both sides (e.g. *O'Brian* or *mor'va* "we have to").
- Comma and colon become part of a W token if used without space on both sides and if the string contains only digits (e.g. *30:00, 200,000,000*).

Similar rules are applied for other C tokens, including those used in URLs and e-mail addresses where all C characters become part of a single W token (using a regular expression).

**Segmentation**: sentence segmentation is performed after tokenisation. The basic rule of "sentence terminal character + space + capitalised word" is complemented with several combinations of different punctuation characters.

| Markup Type | No. of occurrences |
|---|---|
| Texts | 1,655 |
| Paragraphs | 8,137 |
| Sentences | 27,829 |
| Punctuation tokens | 85,953 |
| Word tokens | 500,295 |

Table 2: Overview of segmentation and tokenisation annotations in ssj500k.

## 3.2. JOS Annotation

### 3.2.1. JOS Morphosyntactic Annotation

The morphosyntactic annotation in ssj500k follows the so called JOS morphosyntactic specifications, which are identical to the MULTEXT-East (Erjavec, 2012) morphosyntactic specifications for Slovene. In particular, Version 2.2 of the ssj500k corpus follows MULTEXT-East Version 6.[4] which, for Slovene, lists 1,900 morphosyntactic descriptions (MSDs) and their decomposition into features (e.g. part-of-speech, gender, number, case etc.). Furthermore, the specifications give the MSDs, the attributes, and their values both in English and in Slovene, so it is possible to switch between the morphosyntactic description language. As Table 3 shows, the ssj500k corpus contains approximately 80% of the 1900 theoretically possible MSDs.

| Annotation | No. of types |
|---|---|
| JOS lemmas | 38,818 |
| JOS morphosyntactic tags | 1,304 |

Table 3: Overview of JOS morphosyntactic annotations in ssj500k.

As part of the "Communication in Slovene" project, the lemmatisation guidelines for Slovene were also prepared. The guidelines rely on the MULTEXT-East part-of-speech categorisation and define capitalisation rules for the lemmatisation of (potentially) proper-noun-related words, and

[3] https://github.com/clarinsi/Obeliks4J

[4] MULTEXT-East morphosyntactic specifications Version 6 are maintained on https://github.com/clarinsi/mte-msd, and also available at http://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

other problematic issues in Slovene capitalisation (Holozan et al., 2008).

The lemmatisation and morphosyntactic annotations have been used in the development of several locally developed tagging tools, such as the Obeliks (Grčar et al., 2012) and ReLDI (Ljubešić and Erjavec, 2016) taggers, and the ensemble neural network tagger for Slovene (Belej, 2018).

Given that JOS/MULTEXT-East morphosyntactic annotations are also contained (in the `xpos` column) in Universal Dependencies SSJ Slovene treebank (cf. Section 3.3.), which is a subset of the ssj500k corpus, many other lemmatisation and tagging tools competing in the CONLL 2017 and 2018 shared tasks (Zeman and others, 2017; Zeman et al., 2018) have also been trained on this data. The StanfordNLP tool (Qi et al., 2018), in particular, has been trained on full ssj500k morphosyntactic annotations (Ljubešić and Dobrovoljc, 2019)[5] and gives currently the best tagging results for Slovene.

#### 3.2.2. JOS Syntactic Annotation

In both JOS and SSJ projects, the JOS dependency annotation scheme was also developed. The system was designed specifically for Slovene and is based on syntactic dependencies.

The main idea is that the information attributed on lower levels of annotation (MSD, lemmas) need not to be repeated at the syntactic level. The result is an extremely robust system comprising of only 10 labels (Table 4): 5 labels link elements at the word-phrase level, 4 labels pertain to sentence elements (predicate arguments), and the last label links the heads of sentences to the annotation root. The main features of the system are described in (Erjavec et al., 2010), and in more detail in the annotation guidelines (Holozan et al., 2008).

In ssj500k, 11,411 sentences have been annotated using the JOS dependency annotation scheme. This data was primarily used for training a MST-based dependency parser for Slovene (Dobrovoljc et al., 2012), but it has recently also been applied to a neural-network parsing architecture (Qi et al., 2018) to be used in JOS dependency annotation of the Gigafida 2.0 reference corpus of Slovene (Arhar Holdt et al., forthcoming).

| Level | Label | Relation | Tokens |
|---|---|---|---|
| Phrases | Atr | Attribute | 79,556 |
| | PPart | Predicate part | 16,138 |
| | Coord | Coordination | 6,530 |
| | Conj | Conjunction | 19,278 |
| | MWU | Multi-word unit | 828 |
| Arguments | Sb | Subject | 11,690 |
| | Obj | Object | 15,637 |
| | AdvM | Adverbial of manner | 5,762 |
| | AdvO | Other adverbials | 14,276 |
| Root | Root | All other relations | 66,169 |

Table 4: JOS dependency relation annotations in ssj500k.

[5] `https://github.com/clarinsi/classla-stanfordnlp`

### 3.3. Universal Dependencies

Universal Dependencies (UD) is a cross-linguistically consistent grammatical annotation scheme, providing a universal inventory of morphological and syntactic categories and guidelines for their application (Nivre et al., 2016). In its latest release (Nivre et al., 2019), it has been applied to more than 100 treebanks in over 70 world languages, including the written SSJ (Dobrovoljc et al., 2017) and spoken SST (Dobrovoljc and Nivre, 2016) treebanks for Slovene. The SSJ UD treebank is the result of an automatic rule-based conversion of the ssj500k subset annotated for JOS syntax (Section 3.2.), both on the level of morphology and dependency syntax.[6] The UD annotations are thus the only ssj500k layers that have been produced automatically, however, given the carefully designed conversion rules its quality resembles that of manual annotation.

#### 3.3.1. UD morphology

The conversion of JOS morphosyntactic tags (Section 3.2.) to UD part-of-speech categories and morphological features has been performed by a script which uses two semi-ordered tables (one for mapping the POS categories and the other for features). The large majority of rules are direct mappings of specific categories (e.g. conversion of JOS numerals (`M`) with `Form=letter` and `Type=ordinal` to UD adjectives (`ADJ`) with feature `NumType=Ord`) with the exception of the mappings to the UD determiner (DET) and auxiliary (AUX) POS categories, which also require a predefined lexicon and context information, respectively. This mapping has recently been applied to the entire ssj500k corpus (Dobrovoljc et al., 2019), providing much larger training data for UD morphology than the officially released UD SSJ treebank, which also requires full syntactic annotations. The benefits of such larger training set have already been attested in several UD tagging and lemmatisation experiments for Slovene and other South Slavic languages (Dobrovoljc et al., 2019; Ljubešić and Dobrovoljc, 2019).

| Annotation | No. of types |
|---|---|
| UPOS | 16 |
| FEATS | 1,008 |
| UPOS+FEATS | 1,141 |

Table 5: The number of UD POS categories, UD feature combinations and full UD morphology annotations in ssj500k.

#### 3.3.2. UD syntax

Due to several significant differences between the JOS and UD dependency annotation schemes (Dobrovoljc et al., 2017), most notably a much more fine-grained set of dependency labels in UD (i.e. 32 dependency relations), which also cover phenomena outside the predicate-argument structure (cf. Section 3.2.), approximately 250 rules have been designed for dependency relation and head

[6] The JOS to UD conversion scripts are available at `https://github.com/clarinsi/jos2ud`.

identification, taking into account various lexical, grammatical and contextual features of the tokens. Given the limited scope of such detailed conversion rules, only 8,000 out of 11,411 JOS dependency trees in ssj500k have been successfully converted to UD and released as the official UD SSJ treebank. Nevertheless, this dataset has had important implications for Slovene language processing infrastructure, as it has been featured in the development of many state-of-the-art NLP tools for parsing raw text to Universal Dependencies (Zeman et al., 2018).

| UD Dependency relation | Label | Tokens |
|---|---|---|
| Adjectival clause | acl | 2394 |
| Adverbial clause modifier | advcl | 1302 |
| Adverbial modifier | advmod | 11702 |
| Adjectival modifier | amod | 11800 |
| Appositional modifier | appos | 393 |
| Auxiliary | aux | 7161 |
| Case marking | case | 13495 |
| Coordinating conjunction | cc | 4668 |
| Clausal complement | cc:preconj | 62 |
| Preconjunct | ccomp | 1216 |
| Conjunct | conj | 5297 |
| Copula | cop | 2826 |
| Clausal subject | csubj | 558 |
| Unspecified dependency | dep | 8 |
| Determiner | det | 3311 |
| Discourse element | discourse | 68 |
| Expletive | expl | 2298 |
| Fixed MWE | fixed | 455 |
| Flat MWE | flat | 43 |
| Flat MWE: foreign | flat:foreign | 55 |
| Flat name | flat:name | 688 |
| Indirect object | iobj | 606 |
| Marker | mark | 4701 |
| Nominal modifier | nmod | 10352 |
| Nominal subject | nsubj | 7242 |
| Numeric modifier | nummod | 1610 |
| Object | obj | 7036 |
| Oblique nominal | obl | 9900 |
| Parataxis | parataxis | 1625 |
| Punctuation | punct | 18596 |
| Root | root | 8000 |
| Open clausal complement | xcomp | 1202 |

Table 6: Overview of UD dependency relations in ssj500k.

### 3.4. Named Entities

The first version of the corpus was already annotated with named entities (Štajner et al., 2013), but only about a quarter of the text had these annotations. Furthermore, there had been no annotation guidelines for the named entities, with the result that many annotations were inconsistently applied. In the scope of the Janes project (Fišer et al., 2018) we developed guidelines for named entity annotation in Slovene[7] and we used these guidelines to first correct the already existing NE annotations and then to extend the portion of the text that is annotated, so that around half of the ssj500k v2.0 corpus is annotated for named entities.

As summarised in Table 7, these include five categories: persons (i.e. names of people and pets, pseudonyms, fictional characters and named groups of people), person derivatives (i.e. possessive adjectives derived from personal names), locations (i.e. geographical and celestial entities, public places, shopping centres, etc.), organisations (i.e. organisation, company and institution names, media, art-related groups, etc.) and the miscellaneous category that includes various types of entities ranging from names of works of arts to registered names of products or stock market indices. With the exception of possessive adjectives (essential for tools dealing with data anonymisation), most named entity annotations were thus attributed to nouns or noun phrases, usually in a capitalised or acronym form. This subset of the ssj500k corpus has already been used in the development of the Janes-NER (Fišer et al., 2018),[8] a named entity recognition system for South Slavic languages.

| NE Type | Tag | No. of tokens |
|---|---|---|
| person | PER | 2,928 |
| person derivative | DERIV-PER | 163 |
| location | LOC | 1,967 |
| organisation | ORG | 1,356 |
| miscellaneous | MISC | 602 |
| Total | | 7,016 |

Table 7: Overview of named entity annotations in ssj500k.

### 3.5. Semantic Roles

Semantic Role Labelling (SRL) refers to the process of detecting and assigning semantic roles to semantic arguments determined by the predicate or a verb in a sentence. In comparison with syntactic trees, semantic role labelling requires a higher level of abstraction: namely, the same semantic content can be expressed with different syntactic relations. A semantic role labelling annotation scheme for Slovene was developed within the project Semantic Role Labelling in Slovene and Croatian (Gantar et al., 2018b) and follows previous SRL efforts for other languages, especially for Czech (the Prague Dependency Treebank tagset; Mikulová et al. (2006)), but also for English (FrameNet; Baker et al. (1998)); PropBank; Palmer et al. (2005)), Croatian (Crovallex; Preradović et al. (2009)) and other languages (AnCora, Taulé et al. (2011); SoNaR, Schuurman et al. (2010)). It was essential to take into account the specifics of the Slovene language, such as rich morphology and free word order. The final version of the tagset consists of 25 semantic labels: 5 for arguments, 17 for adjuncts, and

[7] The NER Guidelines are available from http://nl.ijs.si/janes/wp-content/uploads/2017/09/SloveneNER-eng-v1.1.pdf.

[8] https://github.com/clarinsi/janes-ner

3 labels for multi-word predicates and other multi-word expressions, as seen in Table 8. Each semantic label was defined and described in detail, which served as the guidelines for the annotation task.

In ssj500k, 5,491 sentences were annotated with semantic roles, with the first 500 sentences used for test annotation. The second phase included automatic annotation of the remaining 4,991 sentences and their manual revision. The resulting annotations in the corpus therefore represent the final role labelling system, tested both on Slovene and Croatian, which was already used for preliminary quantitative analyses (Gantar et al., 2018b). The ssj500k SRL data has also been used for modelling the mate-tools semantic role labeller (Björkelund et al., 2009; Gantar et al., 2018b), which has been applied to the Gigafida 2.0 reference corpus of Slovene.

| Semantic role | Tag | No. of tokens |
|---|---|---|
| actor | ACT | 5,269 |
| patient | PAT | 6,204 |
| recipient | REC | 1,064 |
| origin | ORIG | 290 |
| result | RESLT | 2,756 |
| time | TIME | 1,631 |
| duration | DUR | 414 |
| frequency | FREQ | 278 |
| location | LOC | 1,155 |
| source | SOURCE | 136 |
| goal | GOAL | 571 |
| event | EVENT | 104 |
| aim | AIM | 218 |
| cause | CAUSE | 355 |
| contradiction | CONTR | 128 |
| condition | COND | 315 |
| regard | REG | 452 |
| accompaniment | ACMP | 111 |
| restriction | RESTR | 6 |
| manner | MANN | 1500 |
| means | MEANS | 250 |
| quantity | QUANT | 250 |
| multiword predicate | MWPRED | 365 |
| modal | MODAL | 611 |
| phraseme | PHRAS | 211 |

Table 8: Overview of semantic role annotations in ssj500k.

### 3.6. Verbal Multi-Word Expressions

ssj500k was also annotated with verbal multi-word expressions (VMWEs) as part of the PARSEME COST Action Shared Task 1.1 (Ramisch et al., 2018), which focused on the automatic identification of VMWEs in running text and developed universal annotation guidelines (Candito et al., 2016) that were tested on 13 languages.

The corpus was annotated with four VMWE categories: a) *inherently reflexive verbs* (IRV), which occur with the independent morpheme *se/si* (*zdeti se* 'to seem'); b) *light-verb constructions* (LVC), which consist of a verb and nominal/prepositional phrase (e.g. *imeti mnenje* 'to have an opinion', *biti v dvomih* 'to be in doubt') and are divided into two subtypes: *full light-verb constructions* (LVC.full), where the verb contributes to the meaning on a predominantly categorical level, and *causal light-verb constructions* (LVC.cause), where the subject can be interpreted as the cause or source of the denoted action (e.g. *spraviti v smeh* 'to make someone laugh'); c) *inherently adpositional verbs* (IAV), which only occur with a prepositional morpheme (*simpatizirati z* 'to sympathise with') or change meaning when occurring with a prepositional morpheme (*biti za* 'be for, to support' vs. *biti* 'to be'); and d) *verbal idioms* (VID), which, even with certain syntactic conversions, must retain a meaning that is independent of the meanings of their elements (e.g. *plačati ceno* 'to pay the price').

A total of 13,511 sentences (approx. 48% of the corpus) were annotated using the FoLiA Linguistic Annotation Tool (FLAT). 2,290 of sentences contain at least one VMWE. On average, one VMWE is present in every fourth sentence. The total number of annotated VMWEs in the training corpus is 3,364, with 1,100 different VMWEs. The most frequent category (see Table 9 is IRV (48%) and the least frequent category is LVC.cause (2%).

| MWE Type | Tag | Tokens |
|---|---|---|
| verbal idiom | VID | 724 |
| inherently reflexive verb | IRV | 1627 |
| light verb construction (full) | LVC.full | 239 |
| light verb construction (cause) | LVC.cause | 64 |
| inherently adpositional verb | IAV | 710 |
| Total | | 3,364 |

Table 9: Overview of verbal MWE annotations in ssj500k.

## 4. Format and Distribution

The ssj500k corpus is encoded according to the Text Encoding Initiative Guidelines (TEI Consortium, 2018), in particular, to the TEI parametrisation that is maintained by the CLARIN.SI research infrastructure.[9] The TEI document contains an extensive header, where the metadata of the corpus is given. This includes the usual bibliographic information but also the description of the sources, of the editorial policies, a listing of the tags used in the corpus together with their quantities and explanations, and a list of revisions of the corpus.

Two further elements of the TEI header deserve a mention. First, the header includes taxonomies for the text types included in the corpus and for the various labels of linguistic annotations included in the corpus: the named entity types, the JOS and UD syntactic labels, the types of verbal multiword expressions and of the semantic roles. These taxonomies enable us to give an identifier to each category, which is then referred to in the corpus annotations, while also encoding the gloss of the label both in English and Slovene. The second element is the TEI list of prefix definitions, which contains the definition of prefixes used in

[9] This TEI parametrisation is available from `https://github.com/clarinsi/TEI-schema`.

corpus annotation, e.g. `mte:` for MULTEXT-East morphosyntactic annotations. Each prefix definition contains the prefix, a description of the prefix use, and two regular expressions: a match pattern and a replacement pattern. The two implement the conversion of the string following the prefix into a URI, e.g. that the annotation `mte:Npmsl` corresponds to the URL `http://nl.ijs.si/ME/V6/msd/tables/msd-fslib-sl.xml#Npmsl`.

The body of the corpus is composed of divisions, each containing the samples of one source text of the FidaPLUS corpus, with each division giving also the bibliographic and taxonomic data on the text. The divisions are then composed of paragraphs, and these of sentences. In Figure 2 we give an (abbreviated) example of a sentence together with its annotations.

As can be seen, each sentence contains words, punctuation symbols and white-space elements, and possibly segments marking-up the named entities. Each token is annotated for its MULTEXT-East morphosyntactic description, its UD morphological features, its lemma and is also given an identifier, so that linking-type annotations can refer to it. The linking annotations for the two types of syntactic dependencies (UD and JOS), verbal multiword units (not present in the example sentence) and semantic roles are then given at the end of the sentence as link groups. These contain links, which give the label of the relation and the links to the head and argument of the relation.

The ssj500k v2.2 corpus is available for download under the CC BY-NC-SA licence from the CLARIN.SI repository (Krek et al., 2019), where it is archived in several formats:

- in TEI, with all annotations in English (e.g. `Ncmsn` for Noun, Type=common, Gender=masculine, Number=singular, Case=nominative);
- in TEI, with all annotations in Slovene (e.g. `Somet` for Samostalnik, vrsta=občа, spol=moški, število=ednina, sklon=imenovalnik);
- in CONLL-U format, automatically derived from the TEI version;
- in the so-called vertical format used by well-known concordancers such as CQP and Sketch Engine, automatically derived from the TEI version; this pack also includes the registry file needed to mount the corpus under one of the mentioned concordancers.

The corpus is also available in the noSketch Engine and KonText[10] concordancers installed at CLARIN.SI, and its repository entry is linked to the concordancers. It is thus possible to search and display the corpus, including most of the provided annotations.

## 5. Q-CAT Annotation and Querying Tool

Most annotations presented in sections above were carried out using the Q-CAT Querying-Supported Corpus Annotation Tool (TBA), which has been specifically designed for the annotation of the ssj500k corpus and has evolved accordingly with the new annotation layers being added through the years. Recently, however, the tool has been substantially improved so as to allow user modifications of annotation settings, including the addition of new annotation layers of various types, and complex queries employing one or more annotation layers. Several functions improving the overall user experience have also been added, including dynamic user-defined visualisations.

In short, the Q-CAT GUI, a .NET application, which runs on the Windows operating system, allows the user to annotate sentences with three types of annotation: tags, links and chunks. A tag is the annotation attributed to individual tokens, which typically include information on its surface form, base form (lemma) and morphosyntactic tag. A link is a directed connection from one token to another (including an optional root element), while a chunk is annotation attributed to a set of one or more tokens, which are not necessarily adjacent. The sentence is displayed graphically with tags being visualised as blue tiles attaching to the surface form, links as arrows and chunks as coloured rectangles (cf. Figure 3).

The user may define and customise one or more levels of annotation (for example: named entities, multi-word expressions, etc.), with each level consisting of one or more annotation types (for example: named entities might be persons, organisations, or locations). Each annotation type can be associated with a shortcut key for easier editing and with a colour for the purposes of visualisation. A sentence may be annotated with annotations belonging to any number of levels, but only two levels of the type link can be shown simultaneously (one above the sentence and one below it).

Q-CAT also supports queries that search over all the sentences in the input TEI file, looking for groups of words that satisfy the search constraints. These can be specified on the level of tags, chunks or links, as shown in Figure 4, which illustrates the query or actants (SRL link of type `ACT`) of main verbs (JOS morphosyntactic tag `Vm`), which are not annotated as subjects (JOS dependency label `Subj`). All sentences and tokens matching the constraints are displayed, counted and can be saved either as a subcorpus (`.xml`) or a tab-separated list (`.txt`).

## 6. Conclusions and Future Work

We have presented the content and the recent developments of the ssj500k training corpus for Slovene, which has been manually annotated with respect to segmentation, tokenisation, lemmatisation, JOS morphosyntax and dependency syntax, Universal Dependencies, semantic role labelling, named entities and verbal multi-word expressions. As such, the ssj500k represents the largest open source collection of grammatically annotated training data for Slovene language processing to date, which has been used in the development of many supervised machine learning systems for Slovene. In addition to that, the ssj500k corpus represents one of the fundamental language resources for Slovene language research, with a well-maintained infrastructure enabling its downloading, browsing, querying and annotation.

[10] KonText (Machálek, 2020) is an alternative front-end to Manatee (Rychlý, 2007), the back-end of the Sketch Engine concordancer. It is available from `https://github.com/ufal/lindat-kontext`.

```
<s xml:id="ssj1.1.2">
  <w ana="mte:Ncmsn" msd="UposTag=NOUN|Case=Nom|Gender=Masc|Number=Sing"
    lemma="dogodek" xml:id="ssj1.1.2.t1">Dogodek</w><c> </c>
  <w ana="mte:Sl" msd="UposTag=ADP|Case=Loc"
    lemma="v" xml:id="ssj1.1.2.t2">v</w><c> </c>
  <seg type="name" subtype="loc">
    <w ana="mte:Npmsl" msd="UposTag=PROPN|Case=Loc|Gender=Masc|Number=Sing"
       lemma="Ankaran" xml:id="ssj1.1.2.t3">Ankaranu</w>
  </seg><c> </c>
  ...
  <w ana="mte:Ncfsn" msd="UposTag=NOUN|Case=Nom|Gender=Fem|Number=Sing"
     lemma="nesreča" xml:id="ssj1.1.2.t7">nesreča</w>
  <pc ana="mte:Z" msd="UposTag=PUNCT" xml:id="ssj1.1.2.t8">.</pc>
  <linkGrp corresp="#ssj1.1.2" targFunc="head argument" type="UD-SYN">
    <link ana="ud-syn:root" target="#ssj1.1.2 #ssj1.1.2.t1"/>
    <link ana="ud-syn:case" target="#ssj1.1.2.t3 #ssj1.1.2.t2"/>
    ...
  </linkGrp>
  <linkGrp corresp="#ssj1.1.2" targFunc="head argument" type="JOS-SYN">
    <link ana="jos-syn:Atr" target="#ssj1.1.2.t5 #ssj1.1.2.t1"/>
    <link ana="jos-syn:Atr" target="#ssj1.1.2.t3 #ssj1.1.2.t2"/>
    ...
  </linkGrp>
  <linkGrp corresp="#ssj1.1.2" targFunc="head argument" type="SRL">
    <link ana="srl:ACT" target="#ssj1.1.2.t5 #ssj1.1.2.t1"/>
    <link ana="srl:PAT" target="#ssj1.1.2.t5 #ssj1.1.2.t7"/>
  </linkGrp>
</s>
```

Figure 2: An example of an TEI annotated sentence *The event in Ankaran was a dramatic accident.* The content is abbreviated for display.

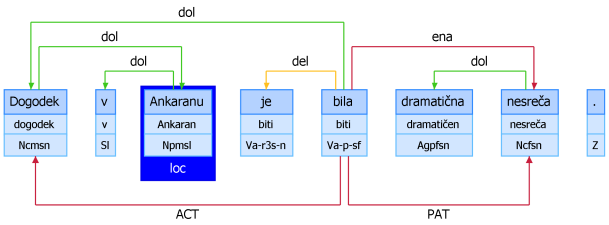



Figure 3: An example of an annotated sentence *The event in Ankaran was a dramatic accident.* in the Q-CAT tool.

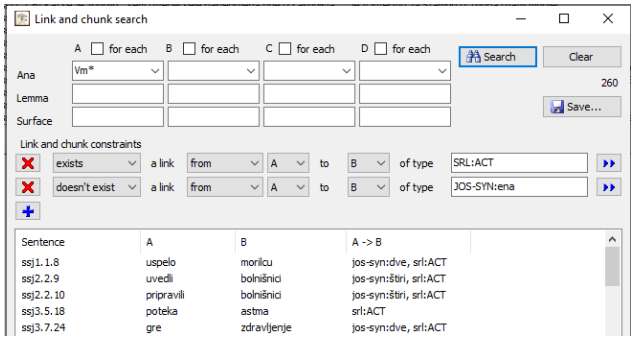



Figure 4: An example of a querying constraint in Q-CAT with results.

To support future research in both natural language engineering and linguistics, we will strive to ensure a continuous development of this dataset in the future as well. In addition to enlarging the coverage of existing annotation layers, especially those covering only part of the corpus, our immediate future work concerns adding new annotation layers, emerging from ongoing annotation campaigns, such as the annotation of non-verbal multi-word expressions (Gantar et al., 2019) and coreference resolution (Žitnik and Bajec, 2018), with possible modifications of the TEI and vertical formats as well.

Since the beginnings of ssj500k, several manually annotated training datasets for non-standard Slovene have also emerged, such as the Janes-Tag corpus of computer-mediated communication (Fišer et al., 2018), the goo300k corpus of historical texts (Erjavec, 2015) and the SST treebank of spoken Slovene (Dobrovoljc and Nivre, 2016), which have been shown to significantly improve non-standard language processing. In line with this, there is an ongoing project investigating the potential removal of the (relatively scarce) non-standard phenomena from the existing version of the ssj500k corpus, with the aim of providing a more reliable dataset for mainstream natural language processing tools, primarily developed for processing of standard Slovene.

## 7. Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. The work

described in this paper was funded by the Ministry of Education, Science and Sport within the "Communication in Slovene" project (3311-08-986003), and the Slovene Research Agency within the national research programmes "Language Resources and Technologies for Slovene" (P6-0411) and "Knowledge Technologies" (P2-0103), the national basic research projects "Linguistic annotation of Slovene language: methods and resources" (J2-9180) and "New grammar of contemporary standard Slovene: sources and methods" (J6-8256), and the Young Researcher Programme (MR-37487).

## 8. References

Špela Arhar and Vojko Gorjanc. 2007. Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa (The FidaPLUS Corpus: A New Generation of the Slovene Reference Corpus). *Jezik in slovstvo*, 52(2):95–110.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Primož Belej. 2018. Oblikoskladenjsko označevanje slovenskega jezika z globokimi nevronskimi mrežami. Master's thesis, Faculty of Computer Science and Informatics, University of Ljubljana.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, Colorado, June. Association for Computational Linguistics.

Marie Candito, Fabienne Cap Cap, Silvio Cordeiro, Vassiliki Foufi, Polona Gantar, Voula Giouli, Carlos Herrero, and et al. 2016. Parseme shared task 1.0 annotation guidelines - version 1.6b (last updated on november 26, 2016). Technical report, COST Action IC1207 – Parsing and multi-word expressions. Towards linguistic precision and computational efficiency in natural language processing (PARSEME).

Ludmila Dimitrova, Tomaz Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevic, and Dan Tufis. 1998. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern European languages. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 315–319, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Kaja Dobrovoljc, Simon Krek, and Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino (dependeny parser for slovene). In *Proceedings of the 8th Language Technologies Conference*, volume C, pages 42–47, Ljubljana, Slovenia, October. IJS.

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

Kaja Dobrovoljc, Tomaž Erjavec, and Nikola Ljubešić. 2019. Improving UD processing via satellite resources for morphology. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 24–34, Paris, France, 26 August. Association for Computational Linguistics.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtsky, and Andreja Žele. 2006. Towards a Slovene dependency treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Tomaž Erjavec and Simon Krek. 2008. The JOS morphosyntactically tagged corpus of Slovene. In *LREC 2008*.

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

Tomaž Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

Tomaž Erjavec. 2015. The IMP historical Slovene language resources. *Language Resources and Evaluation*, 49(3):753–775. `https://doi.org/10.1007/s10579-015-9294-7`.

Darja Fišer, Nikola Ljubešić, and Tomaž Erjavec. 2018. The Janes project: language resources and tools for Slovene user generated content. *Language Resources and Evaluation*. `https://rdcu.be/7RX4`.

Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Taja Kuzman, and Teja Kavčič. 2018a. Glagolske večbesedne enote v učnem korpusu ssj500k 2.1. In *Proceedings of the conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

Polona Gantar, Kristina Štrkalj Despot, Simon Krek, and Nikola Ljubešić. 2018b. Towards semantic role labeling in slovene and croatian. In *Proceedings of the conference on Language Technologies and Digital Humanities*, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

Polona Gantar, Jaka Cibej, and Mija Bon. 2019. Slovene multi-word units: Identification, categorization, and

representation. In *Computational and Corpus-Based Phraseology - Third International Conference, Europhras 2019, Malaga, Spain, September 25-27, 2019, Proceedings*, pages 99–112.

Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik (Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene). In *Proceedings of the 8th Language Technologies Conference*, volume C, pages 89–94, Ljubljana, Slovenia, October. IJS.

Peter Holozan, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman, and Aleš Velušček. 2008. Specifikacije za učni korpus. Projekt "Sporazumevanje v slovenskem jeziku" (Specifications for the Training Corpus. The "Communication in Slovene" project). Technical report, Project "Communication in Slovene".

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, and Anja Zajc. 2019. *Training corpus ssj500k 2.2*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1210`.

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, BSNLP@ACL 2019*.

Nikola Ljubešić and Tomaž Erjavec. 2016. Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Tomáš Machálek. 2020. KonText: Advanced and flexible corpus query interface. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France, May. European Language Resources Association.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2006. Annotation on the tectogrammatical level in the prague dependency treebank. annotation manual. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Joakim Nivre et al. 2019. *Universal Dependencies 2.4*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106, March.

Nives Mikelić Preradović, Damir Boras, and Sanja Kišiček. 2009. Crovallex: Croatian verb valence lexicon. In *Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces*, pages 533–538, June.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency Parsing from Scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium, October. Association for Computational Linguistics.

Carlos Ramisch, Silvo Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, and et al. 2018. Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions. In *LAW-MWE-CxG 2018, The 12th Linguistic Annotation Workshop (LAW XII) and the 14th Workshop on Multiword Expressions (MWE 2018)*, pages 222–240, Santa Fe. Ljubljana University Press, Faculty of Arts.

Pavel Rychlý. 2007. Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno. Masarykova univerzita.

Ineke Schuurman, Véronique Hoste, and Paola Monachesi. 2010. Interacting semantic layers of annotation in SoNaR, a reference corpus of contemporary written Dutch. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Mariona Taulé, M. Antònia Martí, and O. Borrega. 2011. Argument Structure Guidelines for Catalan and Spanish. Working paper 4: TEXT-MESS 2.0 (Text-Knowledge 2.0). Technical report, University of Barcelona.

TEI Consortium, editor. 2018. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Daniel Zeman et al. 2017. CoNLL 2017 Shared Task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.

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

Tadej Štajner, Tomaž Erjavec, and Simon Krek. 2013.

Named entity recognition in slovene text. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(2):58–81, Dec.

Slavko Žitnik and Marko Bajec. 2018. Odkrivanje koreferenčnosti v slovenskem jeziku na označenih besedilih iz coref149. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 6:37–67, 06.

# Epistemic modal adverbs in Slovenian academic discourse

**Jakob Lenardič*** **and Darja Fišer*†**

*Department of Translation
Faculty of Arts
University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana
jakob.lenardic@ff.uni-lj.si

†Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
darja.fiser@ff.uni-lj.si

**Abstract**

In this paper, we analyse Slovenian epistemic modal adverbs in the 100-million-token *KAS corpus of Slovenian PhD theses*. The focus of the analysis is a comparison of their usage in the humanities and social sciences on the one hand and natural and technical sciences on the other. Since modals are in principle ambiguous between epistemic and non-epistemic readings, we conduct a randomized concordance analysis with which we show that only those modals that are exclusively used in their epistemic sense are more frequent in the humanities and social sciences theses. We also show that the non-epistemic dispositional meaning of possibility, which is most commonly used in natural and technical sciences theses, does not constitute hedging.

## 1. Introduction

In this paper, we analyse Slovenian epistemic modal adverbs in the 100-million-token *KAS corpus of Slovenian PhD theses* (Erjavec et al., 2019c), comparing their use in the theses from the humanities and social sciences on the one hand and natural and technical sciences on the other.

Modals offer an interesting insight into academic discourse because they can pragmatically function as *hedges* (Lakoff, 1972; Hyland, 1996; Hyland, 1998), which are used by writers to present their claims with varying degrees of tentativeness. Hedging is a particularly important pragmatic strategy employed in academic writing, as it "enables writers to express a perspective on their statements, to present unproven claims with caution, and to enter into a dialogue with their audiences" and is therefore an "important means by which professional scientists confirm their membership in research communities" (Hyland, 1996, 251–252).

In the related work, which has been done primarily on the basis of English academic discourse, it is often shown that hedging is more characteristic of humanities and social sciences rather than natural sciences (Hyland, 1998; Takimoto, 2015), which reflects the general idea that humanities are more interpretative and less rooted in empirical research than natural sciences (Takimoto, 2015).

In this paper, we try to confirm whether this is also the case for Slovenian academic discourse. To our knowledge, this paper presents the first such comparative analysis of Slovenian academic disciplines from the perspective of hedging; in related work by e.g. Pisanski Peterlin (2010), the notion is exclusively discussed in the framework of translation theory in relation to how English hedges are translated into Slovenian and vice versa.

We present a descriptive quantitative analysis of 9 modal adverbs and then conduct a randomized concordance analysis with which we show that only those modals that are exclusively used in their epistemic sense and thus pragmatically correspond to hedging are also used more frequently in the humanities and social sciences subcorpora of the *KAS* corpus, which corresponds to similar findings for corpus-based analyses of English academic discourse reported by Hyland (1998), Takimoto (2015) and Rizomilioti (2006).

The paper is structured as follows. In Section 2., we lay out the relevant linguistic theory on modality and present the pragmatic notion of hedging, as well as discuss related work on corpus-based treatment of hedging in academic discourse. In Section 3., we present the corpus we used for our analysis from the perspective of the extra-linguistic metadata relevant for our purposes as well as discuss the selection criteria of the modal adverbs that we have analysed. In Section 4., we present the analysis. In Section 5., we discuss our findings by comparing them to those in the related literature and conclude the paper.

## 2. Theoretical framework

### 2.1. Epistemic and non-epistemic modalities

Modality has been defined in many different ways in the literature, but it is perhaps von Fintel (2006, 21) who most succinctly summarizes the notion:

> Modality is a category of linguistic meaning having to do with the expression of possibility and necessity. A modalized sentence locates an underlying or prejacent proposition in the space of possibilities [...] *Sandy might be home* says that

there is a possibility that Sandy is home. *Sandy must be home* says that in all possibilities, Sandy is home.

Modality thus evaluates a proposition from the perspective of the gradient from possibility to necessity.[1] Aside from this, modality comes in different semantic flavours that are contextually determined, and the usual linguistic distinction is made between epistemic modality on the one hand and non-epistemic modality on the other (Palmer, 2014), the latter of which is usually referred to as root modality (Coates, 1983) or circumstantial modality (Kratzer, 2012).[2]

Epistemic modality encompasses the speaker's judgement about the truth of the proposition (Palmer, 2014, 50); i.e., a modal like *mogoče* in the sentence *Ana je mogoče doma* "Ana might be home" constitutes an epistemic modal because its use asserts that the speaker is not completely certain that the prejacent i.e. unmodalised proposition *Ana je doma* "Ana is home" is true.

On the other hand, root (or circumstantial) modality also evaluates the proposition in the domain of possibility, but unlike epistemic modality does not tie the evaluation to the speaker's knowledge. An example of a non-epistemic modal is *lahko* in the sentence *Ta program se lahko namesti na Windows* "This program can be installed on Windows", where *lahko* is not used to indicate the speaker's knowledge about the truth of the expressed proposition but rather to attribute possible qualities to the subject NP *Ta program* "this program".

Modals are often ambiguous between epistemic and non-epistemic readings. For instance, *lahko* in the sentence *Ana je lahko doma, lahko pa je v šoli* is ambiguous between an epistemic reading that can be paraphrased as "'It is possible that Ana is at home or at school" and a root meaning that denotes permission that Ana is granted by someone else ("Ana is allowed to stay at home or in school").[3] Such often unpredictable ambiguity motivates the manual concordance analysis of the Slovenian modal adverbs that will be presented in Section 4.2.

Lastly, what is especially important for our purposes regarding academic discourse is that many root modal expressions display prominent meta-discursive usage, as in the case of reader-oriented meta-commentary clauses like *Kot lahko vidimo iz rezultatov* "As can be seen from the results". Such use along with the purely epistemic meaning often corresponds to the pragmatic notion of hedging (Hyland, 1996; Hyland, 1998; Grabe and Kaplan, 1997), which we introduce in the following section.

### 2.2. Hedging

In linguistics, Lakoff (1972, 471) was the first to use the term *hedges* to refer to those "words whose meaning implicitly involves fuzziness – words whose job is to make things fuzzier or less fuzzy". Lakoff (1972)'s basic concept is further explicated by Hyland (1996, 251), who claims that hedges are "any linguistic means used to indicate either (a) a lack of complete commitment to the truth of a proposition, or (b) a desire not to express that commitment categorically". Additionally, hedging not only involves markers of tentativeness but is typically extended to include rhetoric communicative strategies, i.e., politeness strategies by means of which the author implicitly includes the addressee in the discourse her or she is presenting (Grabe and Kaplan, 1997, 154).

Hyland (1996)'s definition of hedging overlaps quite significantly with that of epistemic modality defined in the previous section, but there is an important difference: a hedge is not a lexical property that holds of a specific category like modality, but rather corresponds to a pragmatic device that can in principle hold for any kind of lexical category given the right communicative context.

In terms of grammatical categories, hedging corresponds not only to modal verbs or adverbs, but to other lexical categories such as the use of certain reporting verbs that indicate the author's tentativeness (e.g., *We believe that*) as well as syntactic strategies such as the use of the passive rather than the active voice to syntactically omit the otherwise entailed agent of the verbal event (Rizomilioti, 2006, 56) or the use of inclusive plural pronouns to help establish rapport between the reader and the writer (Hyland, 1996).

### 2.3. Related work

In related work on hedging in academic discourse, researchers (Hyland, 1998; Rizomilioti, 2006; Pisanski Peterlin, 2010; Takimoto, 2015, a.o.) have generally taken into account all of the major categories that can in principle be used to hedge discourse, such as modal auxiliaries, modal and non-modal (e.g., approximators) adverbs and adjectives, and lexical verbs.

For instance, Takimoto (2015) analyses how hedges corresponding to 5 syntactic categories (adverbs, adjectives, auxiliaries, nouns, and verbs) are used across 4 different natural sciences disciplines and 4 humanities/social sciences disciplines, showing that "70% of all hedges and boosters were found in humanities and social sciences" (Takimoto, 2015, 103) and that philosophy contains "almost 5.3 times as many hedges and boosters as electrical engineering" (Takimoto, 2015, 103).[4] Similarly, Rizomilioti (2006, 64) compares the use of hedging between a

[1] Notions such as *possibility*, *likelihood*, and *necessity*, which are logically related by entailment, are also referred to as the *modal force* (Kratzer, 2012).

[2] To our knowledge, no one has yet attempted a comprehensive study of the syntactic and semantic properties of the Slovenian modal system in the context of descriptive Slovenian linguistics on par with e.g. Palmer (2014)'s work for English modal auxiliaries. Usually, researchers have studied Slovenian modals in relation to highly specialised topics in formal grammar – a good example of this is Marušič and Žaucer (2016)'s work, which proposes a generative syntactic explanation why the modal adverb *lahko* is a positive-polarity item – or have taken modals as a springboard for the study of non-linguistic topics in e.g. translation studies, rather than describing them in detail in their own right.

[3] This modal meaning involving obligation/permission is referred to as *deontic modality* by Palmer (2014).

[4] Some authors use the term *boosters* to describe those hedges that convey the author's certainty rather than tentativeness; since our analysis, presented in Section 4., does not show prominent differences between hedges and boosters, we use *hedges* as a catch-all term for expressing both tentativeness and certainty.

200,000 token corpus of journal papers in literary criticism and a comparable corpus of papers in biology, showing that there are more adverbs of uncertainty in the literary criticism corpus than in the biology corpus.

Given the high degree of lexical polysemy and the consequent likelihood that not all of the observed lexemes in the studied corpus function as hedges, a prominent strategy to filter out irrelevant data relies on the researcher having to read all the concordances that potentially correspond to hedges and then singling out only the relevant occurrences. For this to be possible, the corpora used in the related literature are usually quite small, generally consisting of 100,000–500,000 tokens and around 50–60 research articles (Thompson, 2000; Pisanski Peterlin, 2010; Hyland, 1998; Rizomilioti, 2006; Takimoto, 2015).

Nevertheless, despite such a strategy, the epistemic and non-epistemic notions of possibility seem conflated in some of the related work (Pisanski Peterlin, 2015). We therefore attempt to make our quantitative analysis of the modals more precise by making such a distinction between the modality types introduced in Section 2.1., arguing that only those instances of possibility expressed by the modals that correspond either to epistemic modality or the meta-discursive use constitute hedges, whereas non-epistemic meanings of possibility that correspond to dispositional ascriptions do not.

Our corpus, which we introduce in Section 3.1., is also significantly larger than those in the related literature, consisting of approximately 100 million tokens. Because a manual reading of such a large corpus was not a feasible option for us and because we wanted to reduce the amount of irrelevant data that in part arises from often unpredictable lexical polysemy, we conduct our analysis only on the basis of one word class, i.e., modal adverbs, which arguably constitute the most prominent category for expressing modality in Slovenian.

## 3. Methodology

### 3.1. The KAS Corpus of Academic Slovenian

The study presented in this paper has been carried out on the *KAS corpus of Slovenian PhD theses* (Erjavec et al., 2019c) (henceforth abbreviated as *KAS-dr*), which is a subset of the *KAS corpus of Slovenian academic writing* (Erjavec et al., 2019a). We have chosen the *KAS-dr* corpus because PhD theses represent the most uniform kind of Slovenian academic writing at the post-graduate level in comparison to e.g. the master's theses in the *KAS-mag 1.0* (Erjavec et al., 2019b) corpus, which, because of the Bologna reform, constitutes two types of writing referred to as "master's thesis", where the pre-Bologna variant is longer, more detailed and generally closer to PhD theses in terms of academic maturity than the Bologna variant.

The *KAS-dr* corpus consists of 1569 PhD theses, which together amount to approximately 100 million tokens – the PhD corpus thus represents roughly 6% of the entire 1.7-billion-token KAS corpus.The theses were written between 2000 and 2018 at Slovenian universities and other academic institutions.[5] The corpus is linguistically annotated but is also marked up for several extra-linguistic metadata categories that are tailored to the genre of academic theses, the most relevant for our purposes being the publisher and CERIF (Common European Research Information Format).

The Publisher information corresponds to the institution or faculty where the thesis was defended. There are a total of 44 different publisher abbreviations, 36 of which are faculties of the Universities of Ljubljana, Maribor, Nova Gorica, and Primorska. The remaining 8 are research institutes with their own study programmes or private and semi-private colleges.

The corpus represents a very diverse breadth of scientific (sub)disciplines so each thesis has been assigned to (at least) one of the five top-level CERIF[6] categories: BIO(MEDICAL SCIENCES), HUM(ANITIES), PHYS(ICAL SCIENCES), SOC(IAL SCIENCES), TECH(NOLOGICAL SCIENCES). Since the CERIF categories represent a generalised division of academic disciplines, they are particularly well-suited for comparative corpus analyses of academic genres, especially given the diverse disciplinary scope of the individual publishers included in the corpus.

The CERIF division of the theses in the *KAS-dr* corpus is given in Table 1.

| CERIF | Size (in tokens and %) | |
|---|---|---|
| BIO | 9,075,823 | 10% |
| HUM | 12,911,252 | 14% |
| PHYS | 9,785,950 | 11% |
| SOC | 46,758,605 | 52% |
| TECH | 11,993,724 | 13% |
| Σ | 90,525,354 | 100% |

Table 1: The five CERIF subcorpora of *KAS-dr*

As shown in Table 1, the five CERIF subsets of *KAS-dr* are unequal in size, with the SOC(IAL SCIENCES) subset accounting for half of the corpus. Consequently, we will provide frequency counts for our modal adverbs that are relativised to a million tokens in addition to their absolute frequencies (see Section 4.1.). Furthermore, the total token size (90,525,354) listed in Table 1 is slightly smaller than that of the entire *KAS-dr* corpus (101,473,395); this is because approximately 9% of the PhD theses are assigned to multiple CERIF categories, while the texts that we take into account include the majority of the theses with only one CERIF label.

### 3.2. The Modal Adverbs

The modal adverbs analysed in this paper are given in Table 2. There are 5 adverbs that denote possibility (*lahko, mogoče, možno, morda, morebiti*), 2 adverbs that denote likelihood (*najbrž, verjetno*), and 2 adverbs that denote certainty (*zagotovo, gotovo*).

The modals were selected in the following way. We first extracted all the lemmas in the *KAS-dr* corpus that are morphosyntactically tagged as either adverbs or as particles. Note that the Slovenian descriptive grammar *Slovenska*

[5] For a description of the KAS corpus, see Erjavec et al. (2020).

[6] https://www.eurocris.org/cerif/main-features-cerif.

*slovnica* (Toporišič, 2004), which is the basis for the MULTEXT tagset[7] used by the *KAS* corpora (Erjavec, 2012), postulates that particle is a separate word class. Toporišič (2004, 445–449), rather unusually, defines the particle class solely in terms of its semantic rather than syntactic properties, claiming that the category is distinct from e.g. adverbs in that it consists of semantically abstract clausal modifiers i.e. propositional operators rather than event modifiers such as manner or time adverbials. While most of the lexemes in Table 2 are tagged as adverbs in the corpus, *morda*, *najbrž*, and *morebiti* are tagged as particles, even though their syntactic distribution is prototypically adverbial.[8] From this extracted list of adverb and "particle" lexemes in the corpus, we selected all that semantically correspond to epistemic modals and have a minimum absolute frequency of 500 tokens in the entire *KAS-dr* corpus, which allowed us to omit very infrequent and stylistically marked modals like *nemara* "perhaps".

| MODAL | AF | RF |
|---|---|---|
| *lahko* "possibly" | 328,481 | 3628.6 |
| *verjetno* "likely" | 14,357 | 158.6 |
| *morda* "possibly" | 10,538 | 116.4 |
| *zagotovo* "certainly" | 3,630 | 40.1 |
| *gotovo* "certainly" | 3,458 | 38.2 |
| *mogoče* "possibly" | 2,194 | 24.2 |
| *možno* "possibly" | 1,518 | 16.8 |
| *najbrž* "likely" | 1,151 | 12.7 |
| *morebiti* "possibly" | 1,048 | 11.6 |

Table 2: The most frequent epistemic modal adverbs in the *KAS-dr* corpus, sorted by absolute frequency (AF) and relative frequency (RF)

The nine lexemes in Table 2 largely correspond to the epistemic modal adverbs identified for Slovenian by Pisanski Peterlin (2015, 31). However, in contrast to her approach, our selection criteria were stricter in that we excluded those adverbs that are frequently ambiguous between a modal and non-modal (e.g., manner) interpretation.[9] Such an ambiguous modal is *očitno* "apparently", as shown by the two possible paraphrases of example (1), where the first corresponds to a modal interpretation denoting the speaker's attitude towards the proposition while the other to a non-modal interpretation in which the adverb specifies the manner of the verbal event.

(1) Voda je očitno narasla.
"It appears that the water has risen."
"The water has risen in an obvious manner."

Discounting such ambiguous adverbs reduced the amount of irrelevant data, i.e., it helped us ensure that our comparative analysis is not hindered by the noise due to such polysemy.

# 4. The Analysis

## 4.1. Quantitative Analysis of Modal Adverbs Across Disciplines

Table 3 provides the distribution of the 9 modal adverbs in focus across the 5 CERIF subcorpora, listing both the absolute frequencies for the occurrence of each modal within a subcorpus as well as the relative frequency normalized to 1,000,000 tokens, which is needed for comparison because of the unequal sizes of the subcorpora.

We now compare the relative frequencies of the modals between the subcorpora. For easier comparison, we highlight in Table 3 the two highest relative frequencies for each modal by marking them in bold.

The modals can be divided into two groups.[10] The first group consists of 4 modals (*lahko, verjetno, mogoče, možno*), whose highest or second-highest frequency is in the natural science (BIO, PHYS) or technical (TECH) subcorpora. The modals in this group are on average 1.4 times more frequent in the BIO, PHYS and TECH subcorpora than in the HUM and SOC subcorpora; in other words, they are more frequently used in PhD theses from the natural/technical sciences than in theses within the humanities or social sciences.

The second group consists of 5 modals (*morda, zagotovo, gotovo, najbrž, morebiti*), whose highest relative frequencies are consistently observed in the HUM and SOC subcorpora. The five modals in this group are on average 2.2 times more frequent in the HUM and SOC subcorpora, meaning that in contrast to the first group they are more characteristic of PhD theses in the humanities and social sciences than theses from the natural/technical sciences.

The distribution of the first group is slightly uneven. Two modals – namely, *lahko* "possibly" and *možno* "possibly" – display both the highest and second-highest relative frequencies in either the BIO, PHYS or TECH subcorpora. For instance, the two highest relative frequencies of *možno* are 29.8 instances per million in the TECH subcorpus and 27.6 instances per million in the PHYS subcorpus, which is more than two times higher than the relative frequency of the adverb in the entire *KAS-dr* corpus (cf. Table 2) and around four times higher than the relative frequency of this adverb in the HUM corpus, which is 7.7 tokens per million.

In the case of the likelihood adverb *verjetno*, only its highest relative frequency (310.8 per million tokens) is observed in a natural sciences subcorpus i.e. BIO, while its second-highest frequency is in the HUM corpus (187.4 per million tokens). The distribution of the possibility adverb

[7] https://www.sketchengine.eu/slovene-tagset-multext-east-v5/

[8] In other words, there are no categorical differences between e.g. *verjetno*, which is tagged as an adverb, and *najbrž*, which is tagged as a particle. For simplicity's sake, we thus refer to all the 9 lexemes in Table 2 as adverbs.

[9] Admittedly, *lahko* also has a manner interpretation i.e. "easily". However, this use is incredibly rare – in our analysis of a randomized set of 250 concordance examples (cf. Section 4.2.) for this adverb, there was only 1 example, given in (i), where *lahko* is used in its comparative form *lažje* and corresponds to the non-modal manner usage.

(i) [...] zaradi česar *lažje* in pogosteje prihaja do sprememb v vrednostih indikatorjev.
"... because of which changes in the values of the indicators occur more frequently and more easily."

[10] In Table 3, this division is shown by the dashed line.

| | BIO | | HUM | | PHYS | | SOC | | TECH | |
|---|---|---|---|---|---|---|---|---|---|---|
| MODAL | AF | RF | AF | RF | AF | RF | AF | RF | AF | RF |
| *lahko* | 30,405 | 3350.1 | 40,223 | 3115.3 | 46,481 | **4749.8** | 159,948 | 3420.7 | 51,424 | **4287.6** |
| *verjetno* | 2,821 | **310.8** | 2,420 | **187.4** | 1,406 | 143.7 | 6,563 | 140.4 | 1,147 | 95.6 |
| *mogoče* | 201 | 22.1 | 269 | 20.8 | 196 | 20.0 | 1,224 | **26.2** | 304 | **25.3** |
| *možno* | 121 | 13.3 | 99 | 7.7 | 270 | **27.6** | 671 | 14.4 | 357 | **29.8** |
| *morda* | 893 | 98.4 | 2,365 | **183.2** | 861 | 88.0 | 5,967 | **127.6** | 452 | 37.7 |
| *zagotovo* | 260 | 28.6 | 584 | **45.2** | 342 | 34.9 | 2,149 | **46.0** | 295 | 24.6 |
| *gotovo* | 153 | 16.9 | 917 | **71.0** | 297 | 30.5 | 1942 | **41.5** | 148 | 12.3 |
| *najbrž* | 63 | 6.9 | 324 | **25.1** | 108 | 11.0 | 596 | **12.7** | 60 | 5.0 |
| *morebiti* | 61 | 6.7 | 165 | **12.8** | 66 | 6.7 | 680 | **14.5** | 76 | 6.3 |

Table 3: Modal adverbs in *KAS-dr*; the relative frequency RF is normalized to a million tokens

*mogoče* is similarly uneven in that its highest frequency (26.2 per million) is in the SOC subcorpus, while its second-highest frequency (25.3 per million) is in the TECH subcorpus; we will account for this distributional property in Section 4.2.

In the second group, the highest and second highest frequencies are observed consistently in the HUM and SOC subcorpora. Furthermore, the differences between the highest and lowest relative frequencies are quite large. For instance, the highest relative frequency of *morda* is 183.2 per million tokens in the HUM subcorpus, which is around twice as many as in the BIO (98.4 per million) and PHYS (88.0 per million) subcorpora, and almost five times as many as in the TECH (37.7 per million) corpus.

Similarly, the second-highest frequency of *morebiti* is 12.8 per million tokens in the HUM corpus, which is almost twice as high as in the natural sciences/technical subcorpora, where it is 6.7 per million in the BIO and PHYS subcorpora and 6.3 in TECH subcorpus , all of which are lower than the 11.6 token-per-million frequency of *morebiti* in the entire *KAS-dr* corpus in Table 2.

### 4.2. Epistemic and non-epistemic usage

In order to account for the pattern presented in the previous section, which is the fact that 5 out of the 9 analysed modal adverbs occur most frequently in the humanities (HUM) and social sciences (SOC) subcorpora of *KAS-dr* while the remaining adverbs are equally or even more prominent in the three natural sciences/technical subcorpora, we have examined a randomized set of 250 concordance examples for each of the nine adverbs. We manually annotated each concordance line for the type of modality that the example expresses, given that modals are in principle ambiguous between epistemic and non-epistemic readings, as was discussed in Section 2.1.

The results of the concordance analysis are presented in Table 4.[11] The analysis has first shown that the use of the modal adverbs in the *KAS-dr* corpus can be grouped according to three major types of modality – epistemic use on the one hand and two types of circumstantial/root modality that we label as the dispositional use and the discursive use on the other.

Second, and most importantly, Table 4 shows that the distribution of epistemic and non-epistemic meanings of the adverbs generally follows the distribution of the modals in the CERIF corpora (Table 3) in the following sense: all the five modals that are used most frequently in the social sciences and humanities subcorpora (*morda*, *najbrž*, *zagotovo*, *gotovo* and *morebiti*) are also almost exclusively used to denote epistemic modality, whereas the other modals – with the exception of *verjetno*, which is also used exclusively as an epistemic modal – are to varying degrees ambiguous between the epistemic and non-epistemic readings.

We now take a closer look at the three types of meaning identified in the concordance analysis, and relate the use of modality to the notion of hedging that was introduced in Section 2.2. For instance, let's first take *morda*, which is used as an epistemic modal in 240 (96%) of the randomized concordances and only in 7 (4%) as a non-epistemic modal in the discursive sense, as being representative of the group that is almost exclusively epistemic. Sentence (2), which is taken from a thesis defended at the Faculty of Social Sciences, exemplifies this epistemic usage, while sentence (3), taken from a thesis at the Faculty of Pedagogy, exemplifies one of the few cases of the non-epistemic meta-discursive use of this modal.

(2) *Morda* je to eden od razlogov, da znanstvena skupnost ni bila uspešna pri svojem "programu" izboljšanja javnega razumevanja znanosti in znanstvene pismenosti.
"Perhaps this is one of the reasons that the scientific community wasn't successful in implementing their proposed program for improving the public understanding of science and scientific literacy."

(3) Zato lahko *morda* na tem mestu poudarim strinjanje z Banduro (1997), da je samoučinkovitost precej povezana s samouravnavanjem [. . .]
"This is why I can (perhaps) emphasise my agreement with Bandura (1997) that self-effectiveness is related to self-regulation."

Pragmatically, *morda* in its epistemic sense in example

[11] Note that, in Table 4, the number of included concordances for each modal is not always exactly 250, like 248 in the case of *možno*. The lower number in these cases is due to a few instances of incorrect part-of-speech tagging in the corpus (e.g., some syncretic premodifying adjectives, like *možno* in the accusative/instrumental NP *možno analizo* "possible analysis", are incorrectly tagged as adverbs); we have discarded such irrelevant occurrences from our analysis.

| MODAL | EPISTEMIC | | DISPOSITION | | DISCURSIVE | |
|---|---|---|---|---|---|---|
| *lahko* | 25 | 11% | 117 | 47% | 105 | 42% |
| *mogoče* | 150 | 60% | 97 | 39% | 3 | 1% |
| *verjetno* | 250 | 100% | 0 | 0% | 0 | 0% |
| *možno* | 6 | 2% | 233 | 94% | 9 | 4% |
| *morda* | 240 | 96% | 0 | 0% | 7 | 4% |
| *najbrž* | 250 | 100% | 0 | 0% | 0 | 0% |
| *zagotovo* | 243 | 98% | 0 | 0% | 4 | 2% |
| *gotovo* | 245 | 98% | 0 | 0% | 5 | 2% |
| *morebiti* | 250 | 100% | 0 | 0% | 0 | 0% |

Table 4: The epistemic/root distribution of the modal adverbs in *KAS-dr*

(2) corresponds to Hyland (1996, 256–257)'s notion of an *accuracy-based hedge*, as it is used by the writer to denote his or her uncertainty about the validity of the proposition in the example; i.e., that whatever is denoted by the demonstrative *to* "this" in the main clause is indeed one of the reasons for the lack of success on part of the scientific community. All the epistemic examples with the remaining modals (which we do not exemplify here so as not to exceed the scope of the paper) also function as similar accuracy-based hedges, where the sole semantic and pragmatic difference is in the modal force of the lexeme in question; i.e., a modal like *najbrž* "likely" denotes a greater degree of the speaker's commitment to the truth of the proposition than *morda* or *morebiti* "possibly".

By contrast, *morda* in sentence (3) clearly does not denote the writer's uncertainty, and could be freely omitted from the sentence without a change in the propositional truth-commitment. It is rather used as part of a metadiscursive strategy with which the writer "acknowledge[s] the reader's role in ratifying knowledge" (Hyland, 1996, 258), in the sense that the lexical meaning of possibility which is inherently entailed by the modal "subtly hedges the universality of a writer's claim by implying that a position is an individual interpretation" (ibid.).

Such metadiscursive use is most prominent with the modal *lahko*, having been observed in 105 (42%) out of a total 250 of the randomized set of concordances. The sentence in (4), which is taken from a thesis at the Biotechnical Faculty, exemplifies this usage.

(4) Zaključimo *lahko*, da alkidni premazi na osnovi organskih topil izkazujejo nižje kontaktne kote na obeh substratih kot vodni akrilni premazi [. . .]
"We can conclude that alkyd coatings on the basis of organic solvents show smaller contact angles on both substrates than aqueous acrylic coatings. . ."

In all the 105 examples with the meta-discursive use of *lahko*, the modal adverb is used with directive verbs that are inflected for the so-called inclusive plural, like *zaključimo* "we conclude" in example (4). According to Takimoto (2015, 99), the use of "inclusive pronouns (e.g., *we*) [. . .] enables the writers to produce more interpresonal signals to the readers, which may allow the writers to share contexts with the readers and draw on their assumed belief specific to a particular field of study". In other words, the inclusive inflection emphasises the meta-discursive use of *lahko* as a hedge that is reader-oriented rather than accuracy-oriented (Hyland, 1996). Note that the remaining modals which are also used in this meta-discursive role (*mogoče, možno, morda, zagotovo, morebiti*) do not as consistently pattern with the inclusive plural inflection (cf. example (3), where the first person is used), which may possibly correlate with the fact that their use in this role is much less frequent in comparison to *lahko*, this being the de-facto modal for expressing meta-discursive commentary.

Finally, we turn to the fact that the modals *lahko*, *mogoče*, and *možno* convey, in addition to the epistemic and meta-discursive meanings, the root modal meaning that we refer to as the dispositional use. Sentence (5), which is taken from a thesis at the Faculty of Medicine, exemplifies this meaning with the modal *možno*, which is by far the most frequently used in this sense (233 or 94% examples), while sentence (6), which is from a thesis at the erstwhile Faculty of Electrical Engineering, Computer Science and Information Sciences, contains the modal *mogoče*, which is used in the dispositional sense in 97 (39%) of the concordance examples.[12]

(5) Upliniti je *možno* najrazličnejšo biomaso (les, oglje, kokosove olupke, riževe lupine).
"It is possible to gasify many kinds of biomass (wood, charcoal, coconut peels, rice husks)."

(6) Celoten grafični vmesnik je zasnovan tako, da ga je *mogoče* hitro prilagoditi potrebam metode [. . .]
"The entire GUI is designed in such a way that it can be easily tailored to the needs of the method."

In such cases, the modals are used to denote possibility in its root non-epistemic sense. This kind of modality is not concerned with the knowledge or attitude of the writer (as in the case of epistemic modals and those used in the meta-discursive sense), but is rather used to convey the characteristic properties (i.e., the disposition) on the basis of which

[12] In standard descriptive Slovenian linguistics, the lexemes *možno* and *mogoče* in sentences like (5) and (6) are usually referred to as adverbs; see e.g. the *Dictionary of Standard Slovenian* entry for *možno* (`https://hdl.handle.net/11346/IA5E`). Note, however, that in both examples *možno* and *mogoče* require that the VP be infinitival. It would therefore be more precise to analyse the two lexemes as predicative adjectives, on par with those heading extrapositional *it*-constructions in English like *It is possible to*+$VP_{inf}$ (linden and Davidse, 2009). Conversely, adverbs in clausal adjunct positions are unable to govern the syntactic properties of other sentential constituents in such a way.

the underlying subject NP can be used in some way; for instance, example (6) says that the GUI is such that it is possible to tailor it to the needs of whatever is the method in question.

Palmer (2014, 38) claims that such subject-oriented modality is actually "not strictly a kind of modality at all, modality being essentially subjective", and that such modals are used "to make purely objective statements about the subject of the sentence" (ibid.). From the perspective of pragmatics, it does not seem that such dispositional modals actually constitute hedging of any kind given that they are used to convey objective properties of what the authors are describing in a given example. It should be noted that Hyland (1998, 5) claims that "hedges are the means by which writers can present a proposition as an opinion rather than a fact: items are only hedges in their epistemic sense, and only when they mark uncertainty". Example (5) does not involve the speaker's opinion one way or the other, hence it is not a hedge.

In relation to our observation that the non-epistemic use of modal adverbs is correlated with natural sciences and technical theses rather than those in the humanities or social sciences, it is then not surprising that *možno*, which is used in such dispositional contexts in the overwhelming majority of the randomized concordance examples, is by far the most frequent in the PHYS and TECH subcorpora (27.6 and 29.8 tokens per million, respectively), while its frequency in e.g. the HUM subcorpus, i.e., 7.7 tokens per million, is well below the 16.8 tokens-per-million frequency for the entire *KAS-dr* corpus (cf. Table 2). That is, *možno*, which is used almost exclusively as a non-attitudinal dispositional modal, is well suited for the natural sciences, which are generally objective in that they deal "with numerical data, which is more likely to generate a more precise picture of their findings" (Takimoto, 2015, 95) than e.g. the presumably more subjective and less empirical humanities.

By contrast, *mogoče*, which is lexically perfectly synonymous with *možno* in that it can in principle convey all the three observed types of modality, is used as a dispositional modal only in 39% cases, while 60% of the examples constitute the epistemic use in parallel with *morda* in example (2), so its use is more evenly distributed between natural/technical sciences on the one hand and social sciences/humanities on the other, as shown by the fact that the two highest relative frequencies are in the technical subcorpus (i.e., 25.3 per million in TECH) and in the social sciences subcorpus (i.e., 26.2 per million in SOC).

## 5. Discussion and conclusion

In this paper, we have analysed modal adverbs in the 100-million-token *KAS corpus of Slovenian PhD theses*, comparing their frequency and use between humanities and social sciences subcorpora on the one hand and natural sciences and technical subcorpora on the other. As one of our main contributions to the research on hedging, we have taken into account the fact that modals are in actual usage often unpredictably ambiguous between epistemic and non-epistemic readings, and argued that only those modals that either convey epistemic judgements or metadiscursive commentary also function as hedges, whereas those that express dispositional possibilities do not.

On the basis of this distinction, we have shown that 5 out of the 6 modals that are almost exclusively used in the epistemic sense (i.e., *morda, najbrž, zagotovo, gotovo, morebiti*[13]), and that thereby constitute accuracy-based hedges displaying varying degrees of the authors' tentativeness about the truth of the proposition, are used 2.2 times more frequently in Slovenian PhD theses in the humanities and social sciences rather than the natural and technical sciences.

This result is generally consistent with findings in similar studies that compare the use of adverbial hedging between humanities disciplines on the one hand and natural sciences on the other. For instance, Takimoto (2015) shows that in his corpus, the highest relative frequency of English adverbs of possibility, namely 1200 per million, is in the humanities disciplines, while the second highest is in the social sciences disciplines (800 per million) and the lowest in the natural sciences disciplines (600 per million) (Takimoto, 2015, 102). This is consistent with our findings, since we have for instance shown that the highest frequencies of the epistemic possibility adverb *morda* are also in humanities and social sciences (183.2 and 127.6 per million respectively), while the third highest, i.e. 98.4 per million is in the biology subcorpus. Similarly, Rizomilioti (2006, 64) shows that there are more adverbs of uncertainty in the literary criticism corpus (a total of 212 examples) than in her comparable biology corpus (a total of 181 examples), whereas we have shown an even greater difference – on average, the purely epistemic modals (that is, the accuracy-based hedges) in our corpus are 2.2 times more frequent in the humanities and social sciences.[14]

However, Rizomilioti (2006) shows that on the whole, i.e., when taking into account other categories such as modal auxiliaries, epistemic adjectives, and epistemic lexical verbs (e.g., *believe*, *indicate*), hedging is actually less

[13]The modal *verjetno* thus remains unaccounted for, so explaining the fact that it is the most frequent in natural sciences discourse, i.e., BIO in Table 3, despite its purely epistemic meaning is left for future work.

Nevertheless, we speculate that the difference arises because *verjetno* does not seem to be completely synonymous with *najbrž* even though both entail likelihood. *Verjetno* seems to have a stronger evidential meaning (i.e., the speaker has some empirical evidence for judging the given proposition as likely), whereas *najbrž* seems more rooted in non-evidential inference. A similar claim has been made for the distinction between the certainty modal auxiliaries in English, where the "difference between *will* and *must* is that *will* indicates what is a reasonable conclusion, while *must* indicates the only possible conclusion on the basis of the evidence available" (Palmer, 2014, 57).

If *verjetno* thus truly has a stronger evidential meaning than *najbrž*, it would then come as no surprise that it is more frequent in biomedical sciences, where empirical evidence abounds.

[14]We do note, however, that the empirical vs. non-empirical divide partially transcends the distinction between humanities/social sciences on the one hand and natural/technical sciences on the other, but is rather influenced by the methodological framework adopted by the researcher. Thus, a thesis in a humanities discipline may be more concerned with empirical data than other theses in the same discipline.

frequent in her literary corpus than it is in the biology corpus, which contradicts the findings reported by Takimoto (2015), who shows that hedging is the most frequent in humanities disciplines across all lexical categories.

The findings in the related work are thus to a degree inconsistent. Consequently, for future work we would like to extend our analysis of modals in the *KAS-dr* corpus to other word classes as well, such as adjectives (which are in many cases cognates of the adverbs, like ADJ *možen* "possible" vs. ADV *možno* "possibly") as well as lexical verbs reporting epistemic judgements. We will thereby be able to further ascertain whether expressions of epistemic modality are really more characteristic of humanities and/or social sciences disciplines across the board, as claimed by Takimoto (2015) and Hyland (1998), or whether they are a quirk of a specific word class, such as adverbs, as shown by Rizomilioti (2006).

## 6. Acknowledgments

We would like to thanks the anonymous reviewers for their comments and suggestions. The work described in this paper was funded by the Slovenian Research Agency within the national research programme *Slovene Language – Basic, Contrastive, and Applied Studies* (P6-0215) and within the national basic research project *Slovene Scientific Texts: Resources and Description* (J6-7094, 2014-2017).

## 7. References

Jennifer Coates. 1983. *The Semantics of the Modal Auxiliaries*. Croom Helm, London and Canberra.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2019a. *Corpus of Academic Slovene KAS 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1244.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2019b. *Corpus of Academic Slovene (MSc/MA theses) KAS-mag 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1266.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2019c. *Corpus of Academic Slovene (PhD theses) KAS-dr 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1265.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2020. The kas corpus of slovenian academic writing. *Language Resource and Evaluation*. Submitted/under review.

Tomaž Erjavec. 2012. Multext-east: morphosyntactic resources for central and eastern european languages. *Language Resources and Evaluation*, 46:131–142. https://doi.org/10.1007/s10579-011-9174-8.

William Grabe and Robert B. Kaplan. 1997. On the writing of science and the science of writing: Hedging in science text and elsewhere. In János S. Petöfi, editor, *Hedging and Discourse*, pages 151–167. de Gruyter, Berlin and New York.

Ken Hyland. 1996. Talking to the academy: Forms of hedging in science research articles. *Written Communication*, 13(2):251–281. https://www.doi.org/10.1177/0741088396013002004.

Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins, Amsterdam.

Angelika Kratzer. 2012. The notional category of modality. In *Modals and Conditionals: New and Revised Perspectives*, pages 27–69. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199234684.003.0002.

George Lakoff. 1972. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, 2(4):458–508. https://www.jstor.org/stable/30226076.

An Van linden and Kristin Davidse. 2009. The clausal complementation of deontic-evaluative adjectives in extraposition constructions: a synchronic-diachronic approach. *Folia Linguistica*, 43(1):171–211. https://doi.org/10.1515/FLIN.2009.005.

Franc Marušič and Rok Žaucer. 2016. The modal cycle vs. negation in slovenian. In Franc Marušič and Rok Žaucer, editors, *Formal Studies in Slovenian Syntax*, pages 167–192. John Benjamins, Amsterdam. https://www.doi.org/10.1075/la.236.08mar.

F. R. Palmer. 2014. *Modality and the English modals*. Routledge, Abingdon-on-Thames.

Agnes Pisanski Peterlin. 2010. Hedging devices in slovene-english translation: A corpus-based study. *Nordic Journal of English Studies*, 9(2):171–193. http://doi.org/10.35360/njes.222.

Agnes Pisanski Peterlin. 2015. So prevedena poljudnoznanstvena besedila v slovenščini drugačna od izvirnih? korpusna študija na primeru izražanja epistemske naklonskosti. *Slavistična revija*, 63:29–44. https://srl.si/ojs/srl/article/view/COBISS_ID-57701986.

Vassiliki Rizomilioti. 2006. Exploring epistemic modality in academic discourse using corpora. In *Information Technology in Languages for Specific Purposes*, Educational Linguistics 7, pages 53–71. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-28624-2_4.

Masahiro Takimoto. 2015. A corpus-based analysis of hedges and boosters in english academic articles. *Indonesian Journal of Applied Linguistics*, 5(1):95–105. https://doi.org/10.17509/ijal.v5i1.836.

Paul Thompson. 2000. Modal verbs in academic writing. In Bernhard Kettemann and Georg Marko, editors, *Teaching and Learning by Doing Corpus Analysis – Proceedings of the Fourth International Conference on Teaching and Language Corpora*, pages 305–328.

Jože Toporišič. 2004. *Slovenska Slovnica*. Založba "Obzorja", Maribor.

Kai von Fintel. 2006. Modality and language. In Donald M. Borchert, editor, *Encyclopedia of Philosophy – Second Edition*, pages 20–27. MacMillan Reference USA, Detroit.

# Nadgradnja Zgodovinarskega indeksa citiranosti

**Katja Meden***, **Ana Cvek**†

*Odsek za tehnologije znanja, Institut ≫Jožef Stefan≪
Jamova cesta 39, 1000 Ljubljana,
katja.meden@ijs.si

†Inštitut za novejšo zgodovino, zunanja sodelavka
Privoz 11, 1000 Ljubljana,
ana.cvek3@gmail.com

**Povzetek**

Začetki Zgodovinarskega indeksa citiranja segajo v leto 2003, ko so raziskovalci Inštituta za novejšo zgodovino začeli spremljati in sistematično popisovati citate za prijave projektov in programov na ARRS. Citatni indeks je v letih doživel nekaj nadgradenj, poskusov harmonizacije podatkov in prečiščevanja relacijskih baz, vendar je bilo v zadnjih letih ugotovljeno, da sistem ne zadostuje potrebam indeksatorjev in uporabnikov. Pred nadgradnjo smo izvedli analizo podatkov, kjer so se identificirale največje težave. Nadgradnja je potekala v dveh delih, v prvem delu smo nadgradili administrativni del, v drugem delu pa spletno aplikacijo. Zgodovinarski indeks citiranja je bil med nadgradnjo tehnično posodobljen in s tem oblikovan tako, da je intuitiven za indeksatorje in uporabnike.

**The Historiography Citation Index Upgrade**

The beginnings of the Historiography Citation Index reach into the year 2003 when the researchers of the Institute of Contemporary History felt the urge to begin systematically indexing citations for the intent of applying for funding of research projects and programs. Over the years HCI was a subject of several upgrades and data harmonization attempts. Despite this fact, the system was proven to have several shortcomings, and thus, another system update was needed. Before the update, extensive analysis was performed to identify the most problematic aspects of the system. The upgrade was performed in two parts. The first part consisted of administration system remodeling, which was followed by a frontend redesign. With the update, HCI was technically improved, which resulted in a more intuitive system to its indexators and users.

## 1. Uvod

Ocenjevanje uspešnosti raziskovalcev na področju humanistike je v primerjavi z drugimi raziskovalnimi področji, predvsem naravoslovnimi, že od samih začetkov precej prikrajšano. Med drugim ocenjevanje temelji na frekvenci citiranosti, te podatke pa pridobimo iz različnih citatnih indeksov, kot sta na primer Web of Science (v nadaljevanju WOS) in Scopus. Monografije so primarni produkt raziskovalnega dela na področju humanistike in družboslovja (Glänzel in Schoepflin, 1999; Huang in Chang, 2008; Nederhof, 2006; Hicks, 2004). V nasprotju z vrednotenjem raziskovalne uspešnosti v naravoslovju se ta področja teže vrednotijo, predvsem zaradi dejstva, da so monografije po večini bolj obsežne kot znanstveni članki (Kousha et al., 2011), in visokih kriterijev vključevanja publikacij v obstoječe indekse citiranja, na primer WOS in Scopus. Med pomembnejše kriterije spadajo redno izhajanje serijske publikacije, jezik publikacije, recenziranost, spoštovanje mednarodnih standardov (kot so informativni naslov, povzetek, popolna bibliografska informacija za vse citirane reference), poleg pogojev pa težavo predstavlja tudi indeksiranje monografij. Obstoječi citatni indeksi se namreč bolj osredotočajo na serijske publikacije. Web of Science zajema okoli 12.000 znanstvenih revij in samo okoli 50.000 monografij, medtem ko Scopus zajema več 21.500 znanstvenih revij in 113.000 znanstvenih monografij. Število monografij v indeksu Scopus odraža večji obseg monografij v primerjavi z WOS, pa vendar monografije v primerjavi s številom znanstvenih člankov v revijah predstavljajo zgolj zanemarljiv del citatnega indeksa (Južnič, 2017). ZIC trenutno vsebuje 4.837 vseh vnosov, od tega 2.901 vnos serijskih publikacij in 1.936 vnosov monografij in poglavij iz monografij, kar predstavlja razmerje 59,9 % serijskih publikacij ter 39,1 % monografij in poglavij iz monografij. Neenakosti pri vključevanju publikacij v citatne indekse so na Inštitutu za novejšo zgodovino skušali zamejiti že v letu 2003. Raziskovalci so začutili potrebo po spremljanju in sistematičnem popisovanju citatov za prijave projektov in programov, kar predstavlja zametek Zgodovinarskega indeksa citiranja (v nadaljevanju ZIC). Osnovni namen je bil ustvariti bazo citatov iz slovenskih zgodovinskih monografij, osrednjih znanstvenih časopisov in revij (Lazarević in Zemljič, 2003). Začetna shema baze, ki je bila precej enostavna, je ob svojem nastanku dobro služila potrebam raziskovalcev, vendar so se sčasoma pokazale pomanjkljivosti (Pančur et al., 2014), ki so vodile v nadaljnje nadgradnje, poskuse harmonizacije podatkov in prečiščevanja relacijskih baz. Zadnja nadgradnja je potekala leta 2012 in predstavlja osnovo in temelj nadgradnje, ki je predstavljena v nadaljnjem besedilu članka.

## 2. Cilji in potek nadgradnje

Pri postopku nadgradnje smo z uporabo sodobnih tehnologij in estetsko privlačne grafične podobe želeli preoblikovati administratorski spletni vmesnik in indeksatorju omogočiti prijazno in pregledno izkušnjo pri urejanju podatkov. Najpomembnejši cilj nadgradnje je bila postavitev ZIC kot ločene aplikacije. Ker je baza MySQL trenutno integralni del portala SIstory in se upravlja s pomočjo skupne administracije, je treba podatkovno bazo ZIC postaviti kot ločeno aplikacijo na poddomeni portala SIstory. Razlog za to je načrtovana postavitev nove digitalne knjižnice por-

tala SIstory kot samostojnega repozitorija z ločeno administracijo. Poleg ločene baze in administracije smo pri nadgradnji upoštevali naslednje sklope problemov. V prejšnji nadgradnji uvoz in izvoz podatkov nista bila mogoča, zato smo želeli to omogočiti. Prav tako smo želeli, da je spletna aplikacija narejena modularno, kar bo omogočalo dodajanje novih funkcionalnih rešitev. Pri uporabniškem vmesniku smo želeli, da je stran prijazna za mobilne obiskovalce, pri iskalniku pa smo želeli doseči hitro in pregledno iskanje po podatkih. Nadgrajeni administracijski modul naj bi omogočal enostavnejši dostop in upravljanje vseh podatkov ter z geslom zaščiten dostop do administracije. Izbrani osnovni podatki morajo biti z ustreznim vmesnikom prosto dostopni strojnemu zajemu podatkov(Pančur, 2019b).

Pri postavljanju ciljev in procesu nadgradnje smo izhajali iz temeljnih načel Raziskovalne infrastrukture slovenskega zgodovinopisja (v nadaljevanju RI INZ), ki vključujejo uporabo uveljavljenih in razširjenih tehnologij, ki jih člani infrastrukture dobro poznajo in obvladajo (načeli enostavnosti in poznavanja), modularno nadgrajevanje obstoječih tehnologij (načelo fleksibilnosti) in uporabo odprtih ali lastniških standardov (načelo odprtosti) (Pančur in Šorn, 2019). V procesu nadgradnje smo tako uporabljali tehnologije, ki jih priporoča RI INZ (Pančur, 2019a), in upoštevajo načeli enostavnosti in poznavanja HTML5 in CSS3, najnovejše verzije PHP, MySQL, ElasticSearch engine, JavaScript in JavaScript knjižnice. Pomemben vidik nadgradnje je tudi vidik interoperabilnosti, ki se v svojem pomenu prepleta z načelom fleksibilnosti. Fleksibilnost in interoperabilnost sistema želimo doseči z implementacijo aplikacijskega profila MODS za uvoz in izvoz metapodatkov v različnih formatih, ki podpirajo nadaljnjo diseminacijo in izmenjavo podatkov z drugimi informacijskimi sistemi. Nadgradnja je potekala v posameznih sklopih, ki so opisani v nadaljevanju besedila.

## 3. Rezultati nadgradnje

Nadgradnja je potekala v dveh delih: prvi del se nanaša na administrativni sistem SIstory. Nadgradnja v tem delu zajema preoblikovanje mask in njihovih polj, postavitev nove XML sheme po standardu MODS za uvoz in izvoz podatkov, iskalnik, ki temelji na tehnologiji ElasticSearch, ter migracije vrednosti ločenih polj Avtor(ji). Drugi del se osredotoča na nadgradnjo spletne aplikacije in uporabniškega vmesnika. Pri programski nadgradnji smo sodelovali z zunanjimi sodelavci Infrastrukture.

### 3.1. Administrativni sistem Sistory

#### 3.1.1. Maske za vnos podatkov

Glavna sprememba v administracijskem sistemu (admin) je prehod s prej enotne maske na dve ločeni. Enotna maska je vsebovala tri razdelke: *Splošni podatki, Podatki o viru* in *Vsebinska obdelava*. Vnos podatkov v maske poteka ročno, podatkovna polja v enotni maski pa so bila nejasna (npr. ponavljanje polja za vnos id številke COBISS, imena avtorja idr.), nekatera tudi brez pomena za potrebe citatnega indeksa. Tako je bil na primer razdelek Vsebinska obdelava za citatni indeks povsem neuporaben, saj vsak zapis vsebuje identifikatorje s povezavami na zapise publikacij (COBISS, SIstory) s polnim metapodatkovnim opisom.

Iz enotne maske sta nastali dve neodvisni maski za vnos podatkov v ZIC V2. Iz maske za vnos publikacije sta nastali dve - *maska za vnos monografij* in *maska za vnos serijskih publikacij*, ki dovoljujeta natančnejši opis glede na publikacijo, ki jo indeksiramo. Vsaka izmed mask, tako kot v prejšnji verziji, vsebuje tudi masko za vnos citatov. Maske so bile oblikovane na podlagi zaznanih težav v prejšnjem administracijskem sistemu, o katerih so poročali indeksatorji, ter na podlagi potreb za opis določene publikacije in citatnega indeksa. Spodnja tabela (Tabela 1) prikazuje polja oziroma metapodatke za opis posameznih del in citatov.

Večina elementov, potrebnih za opis publikacij, je ostala nespremenjena. Po opravljeni analizi elementov mask smo izpostavili ključna polja za potrebe opisa publikacij in njihovih citatov. Večina polj je splošne narave (npr. avtor, naslov, leto, kraj itd.), publikacije, ki jih vnašamo (monografije in serijske publikacije), pa se med seboj razlikujejo v določenih vidikih. Ločeni maski s prilagojenimi polji omogočata (z indeksatorskega vidika) kakovostnejšo indeksacijo publikacije. Elementi so bili spremenjeni ali prilagojeni, saj določeni niso bili ažurirani (na primer element Tipologija) ali niso omogočali dovolj natančnega opisa (element Avtor). Pri poljih Avtor in Urednik smo metapodatkovno polje ločili na dve polji: Ime in Priimek. S tem smo zagotovili natančnejši, bolj strukturiran opis in posledično boljše prikazovanje podatkov. Zaradi nove strukture polja je bilo za povezovanje vrednosti polj treba opraviti migracijo vrednosti iz starih, neločenih polj v nova, strukturno ločena polja v obliki Priimek, Ime (za namen prikaza). Nekaterih elementov iz stare maske v novih maskah nismo vključili, npr. Ključne besede ali Država, saj so bili za opis publikacij v citatnem indeksu nepotrebni. Dodani so bili tudi novi elementi, ki jih starejša maska za vnos podatkov ni vsebovala, ker ti podatki še niso bili potrebni. Tu govorimo predvsem o maski za vnos serijskih publikacij in citatov, kjer smo dodali polji DOI in URL, ki omogočata enoznačno, trajno identifikacijo, prav tako pa poleg polja Sistory ID uporabniku omogočata hiter dostop do publikacije. Pri analizi obstoječih zapisov se je izkazalo, da so zapisi pomanjkljivi in neenotni. Do takšnih napak je prihajalo predvsem zato, ker indeksatorji niso imeli nobenih konkretnih navodil in so publikacije v maski (glavni vnos in citat) vpisovali po lastni presoji. Zato smo se pri nadgradnji odločili, da indeksatorjem ponudimo pomoč, ki jim bo olajšala vnos podatkov, še bolj pomembno pa je, da bi s temi navodili oz. pomočjo radi zagotovili čim bolj enotno indeksacijo ter pravilnejše in natančnejše zapise v indeksu. Ob vsakem polju se pri vseh treh maskah nahaja opis polja z navodili za vnos in primeri, ki naj bi bili indeksatorju v pomoč oz. oporo pri vpisovanju podatkov. Tu velja poudariti, da se zavedamo, da bo do napak kljub pomoči še vedno prihajalo, saj se podatki vpisujejo ročno. S tem, da nudimo navodila za vnos, poskušamo zmanjšati število pogostih napak.

#### 3.1.2. ElasticSearch iskalnik in filtriranje

Iskalnik ElasticSearch je distribucijsko, odprtokodno in analitično orodje za vse vrste podatkov, skupaj z besedilnimi, številčnimi, geoprostorskimi, strukturiranimi in nestrukturiranimi podatki, in temelji na knjižnici Lucene Apa-

Tabela 1: Metapodatki mask za vnos podatkov.

| Metapodatek | min/max. št | Podatkovni tip | Maska (Mono., Serijska, Citat) | Primer |
|---|---|---|---|---|
| Cobiss ID | 0,1 | ID | M,S,C | 3278924 |
| Sistory ID | 0,1 | ID | M,S,C | handle.net/11686/4320 |
| ISBN | 0,1 | ID | M | 987-961-3421-43 |
| ISSN | 0,1 | ID | S | 0353-0329 |
| Jezik | 1,1 | ISO639-2b | M,S | slv - slovenski |
| Tipologija | 1,1 | COBISS tipologija | M,S | 1.16 - Samostojni znan. sestavek |
| Tip | 0,1 | interni seznam | M | Poglavje v monografiji |
| Avtorji | 1,neomejeno | niz | M,S,C | Marko Zajc |
| Naslov | 1,1 | niz | M,S,C | Slovenski intelektualci in.. |
| Vzporedni naslov | 0,1 | niz | M,S | Slovenian Intellectuals... |
| Naslov zbornika | 0,1 | niz | M | Slovenija v Jugoslaviji |
| Naslov vira | 0,1 | niz | S | Prispevki za novejšo zgodovino |
| Uredniki | 0,neomejeno | niz | M | Zdenko Čepič (ur.) |
| Kraj | 0,1 | niz | M,S,C | Ljubljana |
| Založba | 0,1 | niz | M,S,C | Založba INZ |
| Leto | 0,1 | številčna vrednost | M,S,C | 2015 |
| Letnik | 0,1 | številčna vrednost | S,C | 57 |
| Številka | 0,1 | številčna vrednost | S,C | 1 |
| Zbirka | 0,1 | niz;št. vrednost | M | Vpogledi; 10 |
| Stran | 0,1 | št. vrednost | M,S,C | 241 - 256 |
| DOI | 0,1 | ID | S,C | 10.1090/019339135 |
| Baza citatov INZ | 0,1 | gumb | M,S | DA |
| Citat na strani | 1,1 | št.vrednost | C | 34 |
| Vir | 0,1 | niz | C | Prispevki za novejšo zgodovino |

che. Postopek zajema vrednosti se začne s tako ime novanim *data ingestion*, v katerem so surovi podatki zajeti v iskalnik iz različnih virov. V naslednjem delu se oblikuje ElasticSearch index, ki je zbirka med seboj povezanih dokumentov. Vsak izmed dokumentov je povezan s ključi (imena, podatkovna polja ali lastnosti) in njihovimi vrednostmi (niz, številke, Boolovi operatorji, nabor vrednosti ...), indeks sam pa omogoča izvajanje kompleksnih iskalnih poizvedb.

Za implementacijo iskalnika ElasticSearch za ZIC v administrativnem sistemu podatke zajamemo iz relacijske baze, ki temelji na tehnologiji MySQL (Elastic, nd). Indeksirani ključi so v tem primeru podatkovna polja, ki bodo služila iskalnim poizvedbam, in njihove vrednosti (ki so večinoma besedilni nizi ali številčne vrednosti). Iskalnik ponuja izvajanje kompleksnih iskalnih poizvedb, ZIC uporablja funkcijo *simple string query*.

```
GET /_search
{
  "query": {
    "simple_query_string" : {
        ""query": "Mojca + Šorn +
        \"Življenje Ljubljančanov
        med drugo svetovno vojno\""
        "fields": ["title^5", "body"],
        "default_operator": "and"
    }
  }
}
```

Funkcija uporablja preprosto sintakso za besedilne iskalne poizvedbe, na podlagi katere vrača iskalne rezultate z uporabo parserja.

Za iskalnik v spletni aplikaciji indeksiramo zgolj polji *Avtor* in *Naslov*, filtri v spletni aplikaciji pa imajo indeksirana polja (in njihove vrednosti) *Identifikator*, *Avtor*, *Naslov*, *Tipologija*, *Leto*, *Kraj* in *Št. citatov*. V administrativnem sistemu je bil filter nadgrajen. Prej je omogočal filtriranje po naslednjih parametrih: *Avtor*, *Leto*, *Naslov*, *Vir* in *Kraj*. Ti po mnenju indeksatorjev niso omogočali učinkovitega in natančnega iskanja zapisov znotraj baze. Novi filtri vsebujejo večje število parametrov: *Tip (monografija/serijska publikacija)*, *ID*, *Avtor*, *Naslov*, *Leto* in *Vir*. Iskalnik ElasticSearch podpira tudi funkcijo samodokončanja iskalne poizvedbe, poznano tudi pod imenom *Autocomplete* ali *Completion suggester*. Funkcija je optimizirana za hitrost tipkanja, saj se prilagaja hitrosti tipkanja iskalne poizvedbe, ki jo uporabnik vnese. Podpira izključno *type as you go* funkcijo in ni mišljena za samodejno korekcijo iskalne poizvedbe ali funkcije *Ali ste mislili* (Elastic, nd). V našem primeru se na funkcijo samodokončanja, enako kot pri osnovnem iskalniku, vežeta zgolj polji Avtor in Naslov.

#### 3.1.3. Uvoz in izvoz metapodatkov - MODS aplikacijski profil

XML ali eXtensible Markup Format prihaja iz družine označevalnih jezikov, kot sta SGML in HTML. Vendar pa se od omenjenih formatov razlikuje predvsem po svoji

fleksibilnosti v primerjavi s HTML omogoča oblikovanje lastnih označevalcev (angl. tag) in s tem predstavlja enega izmed najpogosteje uporabljanih standardov za izmenjavo podatkov v digitalni humanistiki (W3C, 2008). Že v prejšnjih verzijah baze je izvoz podatkov bil mogoč v formatu XML. Shema je predpostavljala lastne označevalce in ni upoštevala kateregakoli metapodatkovnega standarda. Kot je bilo že omenjeno, to pomeni zmanjšano stopnjo interoperabilnosti podatkov. Iz tega razloga smo shemo in polja prenesli v metapodatkovno shemo MODS. Postopek je vključeval tri faze:

- Pregled elementov stare sheme, ki je za svoje označevalce upoštevala imena, kot so *OpTipBiblEnote* ali *OpSistoryUrnId* del označevalca 'Op' se nanaša na publikacijo, ki jo opisujemo (Op = original publi cation), 'Pv' pa označuje podatke za vir publikacije, sledi interno poimenovanje polja (ki ustreza imenu polja, iz katerega vzamemo podatke).

- Preslikava internih polj (poimenovanje po meri) v metapodatkovni standard MODS in komentiranje kode (navodila za programerja, iz katerih polj v stari metapodatkovni shemi se vežeje vrednosti v nove elemente). Iz ene sheme sta nastali dve novi, upoštevali smo novo strukturo mask za vnos podatkov, tako kot smo predhodno enotno masko razdelili na masko za monografije in serijske publikacije. V aplikacijskem profilu v skupnem metapodatkovnem zapisu v formatu XML sta ločena zapisa mask definirana z elementom mods in identifikatorjem ID=pub za oznako zapisa za monografijo ali serijsko publikacijo (na primer *mods ID="pub.224"*) ali elementom relatedItem in identifikatorjem za oznako navedenih del, na primer *relatedItem type="references"ID="ref.1"*.

- Prenos vrednosti iz starih internih polj v polja MODS ima svoje prednosti poleg dejstva, da tako povečamo interoperabilnost svojih podatkov z drugimi sistemi, s tem pridobimo večjo strukturiranost in pogosto tudi dodatne podatke, ki jih v stari shemi ne bi mogli implementirati. Element OpJezik ima za svojo vrednost na primer le številčno vrednost "21", kar se navezuje na interni nekontroliran seznam jezikovnih vrednosti, novi element pa v svoji strukturi dovoljuje navedbo avtoritete in tipa poimenovanja. Tako poleg jezikovne kode pridobimo tudi podatek o standardu oziroma kontroliranem seznamu, ki je bil uporabljen, s tem pa tudi standardiziramo vrednost zapisa.

Slika 1 prikazuje strukturo in del elementov stare, interne metapodatkovne sheme.

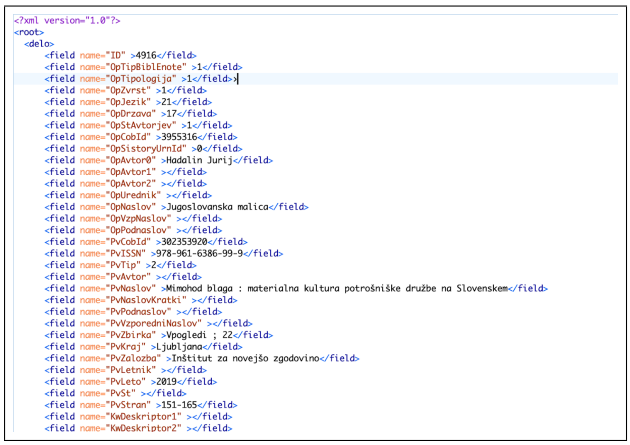
Spodaj sta prikazana star in nov način poimenovanja ter primerjava strukture posameznega zapisa:

Interna shema ZIC (element Avtor):

```
<field name="OpAvtor0">Hadalin Jurij</field>
```

aplikacijski profil v XML:

```
<name type="personal">
     <namePart>Priimek Ime avtorja</namePart>
  <role>
     <roleTerm type="code">cre</roleTerm>
     <roleTerm>Avtor</roleTerm>
  </role>
     <namePart type="family">Priimek</namePart>
     <namePart type="given">Ime</namePart>
</name>
```



```
<?xml version="1.0"?>
<root>
  <delo>
      <field name="ID" >4916</field>
      <field name="OpTipBiblEnote" >1</field>
      <field name="OpTipologija" >1</field>
      <field name="OpZvrst" >1</field>
      <field name="OpJezik" >21</field>
      <field name="OpDrzava" >17</field>
      <field name="OpStAvtorjev" >1</field>
      <field name="OpCobId" >3955316</field>
      <field name="OpSistoryUrnId" >0</field>
      <field name="OpAvtor0" >Hadalin Jurij</field>
      <field name="OpAvtor1" ></field>
      <field name="OpAvtor2" ></field>
      <field name="OpUrednik" ></field>
      <field name="OpNaslov" >Jugoslovanska malica</field>
      <field name="OpVzpNaslov" ></field>
      <field name="OpPodnaslov" ></field>
      <field name="PvCobId" >302353920</field>
      <field name="PvISSN" >978-961-6386-99-9</field>
      <field name="PvTip" >2</field>
      <field name="PvAvtor" ></field>
      <field name="PvNaslov" >Mimohod blaga : materialna kultura potrošniške družbe na Slovenskem</field>
      <field name="PvNaslovKratki" ></field>
      <field name="PvPodnaslov" ></field>
      <field name="PvVzporedniNaslov" ></field>
      <field name="PvZbirka" >Vpogledi ; 22</field>
      <field name="PvKraj" >Ljubljana</field>
      <field name="PvZalozba" >Inštitut za novejšo zgodovino</field>
      <field name="PvLetnik" ></field>
      <field name="PvLeto" >2019</field>
      <field name="PvSt" ></field>
      <field name="PvStran" >151-165</field>
      <field name="KwDeskriptor1" ></field>
      <field name="KwDeskriptor2" ></field>
```

Slika 1: Metapodatkovna polja maske za vnos podatkov pred nadgradnjo.

Interna shema ZIC (element Jezik)

```
<field name="OpJezik" >21</field>
```

Aplikacijski profil v XML:

```
<language>
  <languageTerm type="code">slv</languageTerm>
  <scriptTerm type="code">Latin</scriptTerm>
</language>
```

Interna shema ZIC (element Tipologija):

```
<field name="OpTipologija" >1</field>
```

Aplikacijski profil:

```
<classification authorityURI
="https://www.izum.si/">101</classification>
```

Z novim aplikacijskim profilom, ki izhaja iz metapodatkovnega standarda MODS, smo namesto internih metapodatkovnih elementov v shemi uporabili obstoječi in razširjeni metapodatkovni standard MODS. S tem smo naslovili dve izmed temeljnih načel: poznavanje oziroma uporabo poznanih in razširjenih tehnologij ter načelo interoperabilnosti. Format XML nam namreč zagotavlja lažje izmenjevanje in diseminacijo podatkov z drugimi sistemi.

#### 3.1.4. Migracija vrednosti polj Avtorji

Enega izmed večjih problemov, ki nam ga je delno uspelo rešiti med nadgradnjo, predstavlja migracija vrednosti polja Avtor(ji) iz skupnega polja v dve ločeni. Problem je nastal zaradi neenotnega zapisa oziroma različnih oblik vrednosti Priimek in Ime (oblike: *Priimek, Ime; Ime in Priimek, Ime, Priimek ...*) ter naštevanja več avtorjev v enem polju (*Avtor1; Avtor2 ...*), ki so bili med seboj ločeni z različnimi ločili. Ta problem nam je uspelo rešiti zgolj

delno: migracija, ki je potekala strojno, je bila uspešna na poljih, ki so se med seboj ujemala, pri določenih zapisih pa to ni bilo mogoče (primer *Ime Ime, Priimek*), zato zahteva ročne popravke. Te napake bomo lahko odpravili po začetku procesa prečiščevanja baze, ki pa za zdaj še ni predviden.

### 3.2. Spletna aplikacija in uporabniški vmesnik

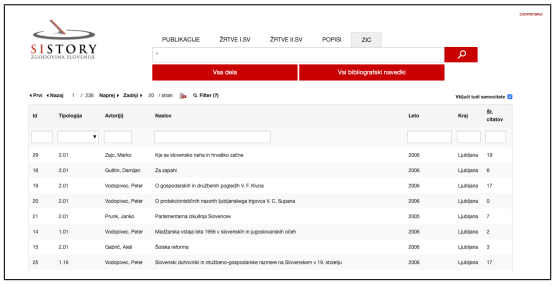
#### 3.2.1. Podatkovna baza vseh del in podatkovna baza vseh bibliografskih navedb

Spletna aplikacija vsebuje dve podatkovni bazi bazo *Vsa dela* in podatkovno bazo *Vse bibliografske navedbe*. Razlog za dve medsebojno ločeni bazi leži v prikazu rezultatov, še natančneje v prikazu števila prejetih citatov pri določenem zapisu. Pri izpisu rezultatov je na voljo število citatov, ki jih je določeno delo prejelo, vendar ti podatki morda niso pravilni, ker se število prejetih citatov določenega dela veže na ujemanje naslova pri glavnem vnosu (maska za vnos glavnega zapisa) in pri citatu (maska za vnos citata). Kot pa smo omenili že zgoraj, nemalokrat pride do napak. Zaradi tega je potrebna druga baza *Vse bibliografske navedbe*, po kateri je omogočeno brskanje z uporabo filtrov. Ta baza dovoljuje uporabniku dodaten in bolj natančen vpogled v citate, saj tu dejansko vidimo vse vnesene citate, indeksatorjem pa predstavlja dodatno orodje za lažje popravke že obstoječih zapisov (preglednejše iskanje zapisov slabše kakovosti).

#### 3.2.2. Prikaz iskalnih rezultatov

Iskalni rezultati so prikazani v obliki tabel, ki uporabnikom ponujajo tudi filtriranje rezultatov oziroma omogočajo oženje iskalne poizvedbe znotraj tabele. Rezultate je mogoče tudi razvrščati. Poleg filtriranja je uporabniku omogočen izvoz zadetkov na seznamu rezultatov in posameznega zadetka v formatu PDF. Za uporabnike sta prav tako pripravljeni tudi dve vrsti pomoči osnovna razlaga uporabe citatnega indeksa na prvi strani ZIC (iskanje/brskanje) in pa manjši namig pri uporabi filtrov s primeri uporabe ločil. Prikaz posameznega zapisa uporabniku dovoljuje vpogled v osnovne podatke (metapodatke dela), osnovne podatke vseh del, v katerih je bil citiran, in avtorjev seznam literature. Podatki so prikazani v dveh ločenih tabelah, *Citirano v* in *Seznam literature*, zapisi so med seboj povezani.

Med oblikovanjem vmesnika so v vmesnih fazah sodelovali raziskovalci/uporabniki, s katerim smo testirali odzive na novi vmesnik, novo podatkovno strukturo in nove funkcionalnosti. Največ težav je predstavljala terminologija, predvsem na podlagi dejstva, da se zgodovinarsko dojemanje terminov literature in virov precej razlikuje od pojmovanja na področju tehnologije. Nerodna poimenovanja iz prejšnje verzije vmesnika (*Avtor citira, Citiranost Avtorja*) je bilo treba nadomestiti s terminom, ki bo uporabnikom razumljiv. Kot že omenjeno, smo se na podlagi tega odločili za osnovno iskanje in dve ločeni bazi, ki sta po številnih preimenovanjih pridobili ime *Vsa dela* in *Vsi bibliografski navedki*. Čeprav sta imeni daljši, smo prednost namenili razlagi terminov, saj so uporabniki menili, da sta ti poimenovanji najbolj jasni in logični. Poleg terminologije je problem predstavljala tudi postavitev elementov na spletni strani (predvsem gumbi). Tu se je izkazalo, da je uporabnike precej medla postavitev gumbov za obe bazi, saj so mislili, da s klikom na npr. *Vsa dela* dobijo vsa dela iskanega avtorja. Težavo smo odpravili tako, da smo ustvarili različne statične verzije uporabniškega vmesnika in s pomočjo uporabnikov določili najbolj jasno in intuitivno.




Slika 2: Trenutni uporabniški vmesnik ZIC-a.

#### 3.2.3. Uporaba indeksa citiranosti

Primarni uporabniki citatnega indeksa so raziskovalci, ki lahko v sistemu enostavno preverijo št. prejetih citatov za posamezno avtorsko delo; če je to indeksirano v sistem. Poleg izpisa iz sistema SICRIS (Slovenian Current Research Information System), ki je osnova za vrednotenje znanstvene uspešnosti na posameznem raziskovalnem področju, lahko izpis iz ZIC predstavlja dodano vrednost pri prijavljanju projektov ali programov na področju humanistike. Poleg raziskovalcev si z ZIC lahko pomagajo tudi uredniki revij, ki želijo preveriti, kolikokrat so bili posamezni članki citirani, in s tem upravičijo obstoj revije. Poleg primarne naloge, ki je nudenje vpogleda v št. prejetih citatov, pa indeks ponuja tudi druge možnosti, ki jih stari ZIC ni ponujal. Te naj bi uporabniku omogočile prijetnejšo interakcijo s sistemom. Ena izmed takšnih funkcionalnosti je npr. možnost *prijaznega kopiranja*, ki uporabniku omogoča lažje navajanje virov v svojih delih, saj ZIC ponuja skoraj popolne bibliografske podatke, ali npr. izpis št. citatov v formatu PDF ipd.

## 4. Zaključek

Sistem je bil že v svoji začetni zasnovi izjemno ambiciozen in zaradi načina objavljanja na področju zgodovinopisja izjemno potreben. Vendar je Zgodovinarski indeks citiranja zadnja leta nekoliko stagniral. Po pregledu in analizi podatkov smo ugotovili, da je nadgradnja potrebna, saj sistem ne zadostuje potrebam indeksatorjev in uporabnikov. Začeli smo nadgradnjo administrativnega dela, kjer smo preoblikovali oz. nadgradili nove maske, nadgradili metapodatkovno shemo oziroma ustvarili nov aplikacijski profil na podlagi metapodatkovnega standarda MODS, filtre in dodali pomoč indeksatorjem, ki naj bi pripomogla k poenotenim zapisom. Poleg administrativnega dela smo nadgradili tudi uporabniški vmesnik z občasnim testiranjem baze in njenih komponent z raziskovalci. Z omenjeno nadgradnjo smo rešili večino zaznanih problemov, od nejasnih in nepotrebnih polj vnosa podatkov in razčlenitve mask, ki indeksatorju omogočajo lažje in natančnejše oblikovanje zapisov, oblikovanja aplikacijskega profila MODS, ki

omogoča lažji uvoz in izvoz podatkov, do uporabniku prijaznejšega vmesnika itd. Vseh težav pa zaradi omejitev, povezanih z ročnim vnosom podatkov, ni bilo mogoče v celoti rešiti. To velja predvsem za postopek migracije polja Avtorji, kjer bo problem v celoti rešen šele po prečiščenju cele baze podatkov. Postopek prečiščenja bo pripomogel tudi k poenotenju zapisov, kar bo omogočalo, da uporabniki v sistemu pridobijo zanesljive in kakovostne informacije. Pri nadgradnji Zgodovinarskega citatnega indeksa smo dosegli zastavljene cilje. Sistem smo tehnično posodobili in ZIC postavili kot ločeno spletno aplikacijo na poddomeni portala SIstory. Spletna aplikacija je narejena modularno, zato je mogoče dodajati nove funkcionalne rešitve, iskalnik s tehnologijo ElasticSearch pa omogoča natančnejše in preglednejše iskanje po podatkih. Z nadgradnjo smo oblikovali sistem, ki je intuitiven za indeksatorje in uporabnike, s tem pa zagotovili, da ZIC služi svojemu namenu.

## 5. Zahvala

Raziskavo je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru programa Raziskovalne infrastrukture slovenskega zgodovinopisja (I0-0013) in slovenske raziskovalne infrastrukture DARIAH SI.

## 6. Literatura

Elastic. n.d. What is elasticsearch. `https://www.elastic.co/what-is/elasticsearch`.

Wolfgang Glänzel in Urs Schoepflin. 1999. A bibliometric study of reference literature in the sciences and social sciences. *Information Processing & Management*, 35(1):31–44.

Diana Hicks, 2004. *The four literatures of social science*, pogl. Handbook of quantitative science and technology research, str. 476–496. Kluwer.

Mu-hsuan Huang in Yu-wei Chang. 2008. Characteristics of research output in social sciences and humanities: from a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, 59(11):1819–1828.

Primož Južnič. 2017. Bibliometrijski indikatorji. Dostopno na: `https://www.youtube.com/watch?v=l9W5glZl97I&feature=youtu.be`.

Kayvan Kousha, Mike Thelwall in Somayeh Rezaie. 2011. Assessing the citation impact of books: The role of google books, google scholar, and scopus. *Journal of the American Society for Information Science and Technology*, 62(11):2147–2164.

Žarko Lazarević in Igor Zemljič. 2003. Slovenski zgodovinarski indeks citiranosti - izhodišča in pomisleki. Neobjavljena dokumentacija.

Anton Nederhof. 2006. Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, 66(1):81–100.

Andrej Pančur, Mojca Šorn in Jurij Hadalin. 2014. Slovenski indeks citiranosti (sici): Načrt izgradnje in delovanja. Tehnično poročilo, Inštitut za novejšo zgodovino. Dostopno na: `https://www.sistory.si/11686/36153`.

Andrej Pančur in Mojca Šorn, 2019. *Zbornik ob 60-letnici Inštituta za novješo zgodovino*, pogl. Raziskovalna infrastruktura Inštituta za novejšo zgodovino. Inštitut za novejšo zgodovino.

Andrej Pančur. 2019a. Preprosta raziskovalna infrastruktura za kompleksne raziskovalne podatke v humanistiki - si4 (Simple research Infrastructure FOR complex research data in digital humanities). Neobjavljena dokumentacija.

Andrej Pančur. 2019b. Specifikacije za izvedbo naročila izdelave Zgodovinarskega indeksa citiranosti (ZIC). Neobjavljena dokumentacija.

W3C. 2008. Extensible markup language (xml) 1.0 (fifth edition). Tehnično poročilo, World Wide Web Consortium. Dostopno na: `https://www.w3.org/TR/xml/`.

# Grammatical Footprint of Socially Unacceptable Facebook Comments

**Kristina Pahor de Maiti*, Darja Fišer♦*, Nikola Ljubešić♦‡, Tomaž Erjavec♦**

* Faculty of Arts, University of Ljubljana, Slovenia
Aškerčeva cesta 2, 1000 Ljubljana
kristina.pahordemaiti@ff.uni-lj.si
darja.fiser@ff.uni-lj.si
‡ Faculty of Computer and Information Science, University of Ljubljana, Slovenia
Večna pot 113, 1000 Ljubljana
♦ Jožef Stefan Institute, Ljubljana, Slovenia
Jamova cesta 39, 1000 Ljubljana
nikola.ljubesic@ijs.si
tomaz.erjavec@ijs.si

**Abstract**

This study investigates the morphosyntactic characteristics of Facebook comments in order to identify those that prominently feature in socially unacceptable discourse (SUD) online with a specific focus on violent vs. offensive comments. By analyzing the frequency and keyness of the observed features, we learned that grammatical differences between socially acceptable and unacceptable comments do exist, but are subtle. Furthermore, we found that differences amplify with the severity of SUD, meaning that the language of violent comments differs more from socially acceptable comments than the offensive ones. Finally, our analysis shows that offensive comments include descriptive and dialogic language, while in the violent comments the focus is on actions and generalizations.

## 1. Introduction

Although socially unacceptable discourse (SUD) is not a new phenomenon, it has become much more prominent with the development of social media platforms. In this paper, we use SUD to refer to communication practices that are "abusive, insulting, intimidating, harassing, and/or incite to violence, hatred, or discrimination« (Erjavec and Kovačič, 2012). Online SUD is especially worrisome due to its wide reach, which is why the last decade has seen growing work in the field of natural language processing (NLP) focusing on the development of tools for automatic detection of SUD (for an overview see Fortuna and Nunes, 2018) but also in psychology, psychiatry, law, criminology and communication studies (Waqas et al., 2019) which shows the complexity of the phenomenon. Furthermore, study outcomes are usually difficult to generalize, since SUD is extensively influenced by many factors, such as the communicative context, the identity of the author, the target of the comment, etc. (Schmidt and Wiegand, 2017). In addition, since SUD is realized through language, its full description should be based on findings from multiple languages. However, until now the majority of the work has been performed on English data, with other languages coming into focus only recently (Corazza et al., 2020). This paper represents an attempt to contribute to the understanding of Slovene SUD by analyzing morphosyntactic properties of SUD and identifying grammatical features that are characteristic of SUD.

This paper is structured as follows: Section 2 reviews related work; Section 3 outlines the study design; Section 4 describes the dataset used; Sections 5 and 6 present and discuss the results; and Section 7 concludes the paper with final thoughts and directions for future work.

## 2. Related work

In this paper, we follow the reasoning of Hopper (1987) and Scheibman (2002) who state that grammatical structure to a certain extent depends on the discourse which is shaped by social interactions and the speaker's point of view. Lakey (2016), for example, showed that distancing behavior related to fear reflects on the morphosyntactic level through distancing markers such as the use of genitive instead of accusative for direct objects in simple fear constructions. Moreover, the author (ibid.: 132) claims that "it is likely that other emotions have bodily or behavioral reflections in the morphology and syntax of languages". Therefore, we can expect SUD to exhibit characteristic grammatical features in relation to socially acceptable discourse (henceforth referred to as non-SUD). We base this assumption on findings from our previous study (Pahor de Maiti et al., 2020) which showed that nonstandard linguistic features of SUD often manifest through nonstandard syntactical constructions (e.g., infinitives in place of finite verb forms). In this analysis, we focus on the morphosyntactic level as the first step towards a comprehensive linguistic description of SUD.

Several computer linguistics studies have already shown that grammatical features, such as tense and aspect markers, conjunctions, PoS distribution, etc. can help improve the performance of SUD-classifiers (cf., Clarke and Grieve, 2017; Vidgen and Yasseri, 2019; Zhang et al., 2018). However, grammatical feature selection has usually not been theoretically underpinned from the linguistic point of view. Nonetheless, certain researchers (e.g., Alorainy et al., 2019; Burnap and Williams, 2016) ground their feature selection in the conceptual framework of othering (Jensen, 2011) which defines actions that are focused on creating the in-group and the out-group (e.g., in language use: a pronounced use of the pronoun pair *us/them*).

Furthermore, linguistic studies of various text genres have shown a correlation between grammar and discourse. Werner (2019), for example, showed that hip hop discourse

manifests through specific grammatical structures (e.g., copula absence, multiple negatives). By analyzing key morphosyntactic features of presidents' speeches, Fidler and Cvrček (2019) found that differences in word class distribution is indicative of a speakers' representation of reality (e.g., noun-heavy discourse de-emphasizes the agent of the action), while morphosyntactic features point to differences in speakers' style (e.g., a finite verb form as opposed to the nominalized phrase suggests explicit reference and allows/encourages feedback; grammatical case indicates different participants roles). Grammatical specificities in different text genres have also been found in Slovene texts. Logar and Erjavec (2018) showed that academic Slovene is characterized by common nouns, adjectives and abbreviations. Zwitter Vitez and Fišer (2015) showed that spoken and online texts contain more pronouns, particles and interjections than traditional written texts, but that there are fewer verbs and conjunctions in online texts with respect to spoken text. In contrast to these studies which compared different genres, ours is more fine-grained in that we are comparing different types of comments within a single text genre, therefore expecting to identify subtler differences.

## 3. The dataset

The dataset used in this paper is extracted from the Slovene part of the FRENK corpus[1] which contains comments on Facebook posts published by the three most popular Slovene news media according to the Alexa service: 24ur.com, SiOL.net, and Nova24TV. The corpus contains comments posted from 2010 to 2017 on the topics of migrants and the LGBTQ+ community which were manually annotated for the type of SUD according to the project-specific annotation schema (e.g., classifying comments into those that express violent actions, contain offensive language, etc.) and its target (e.g., migrants, LGBTQ, commenter, etc.) (Ljubešić et al., 2019). The corpus has also been tagged with morphosyntactic descriptions and lemmatized with the CLASSLA-StanfordNLP tagger (Ljubešić and Dobrovoljc, 2019).

Based on the SUD type, the comments were compiled into three subcorpora in the noSketch Engine concordancer (Kilgarriff et al., 2014) available at CLARIN.SI[2]:

- **NON-SUD**: comments without any elements of SUD (e.g., *Jaz sem navedel dejstva... kaj pa vidva? /I stated the facts… what about you two?/*[3]);
- **OFFENSIVE**: comments that contain SUD in the form of obscene and abusive content, but without incitement to violent acts (e.g., *Ja res si primitiven,dokazano,po tvojem besednjaku /You really are primitive,it's a fact,proven by your vocabulary/*); and
- **VIOLENT**: comments that contain SUD in the form of threating or violent content inciting to physical violence (e.g., *Metk v celo vsakemu /Bullet to the forehead for each one of them/*).

Table 1 shows the structure of our dataset and the size of the subcorpora. First, we can observe that the share of SUD (i.e. the VIOLENT and OFFENSIVE subcorpora) is roughly the same (52.09%) as the share of the NON-SUD subcorpus. Second, the VIOLENT subcorpus is approximately 10 times smaller than the other two subcorpora. This is not surprising since the amount of violent comments on social media is generally low (below 7 %; estimations, however, vary across studies (Berglind et al., 2019; Vidgen and Yasseri, 2019), but to a certain extent this could also be influenced by the Facebook removal policy[4] of hateful content prior to our harvesting. Next, we can see an uneven distribution between the two topics (i.e., migrants and LGBTQ) in all subcorpora, but the largest difference is in the VIOLENT subcorpus. This might be due to the fact that at the time of data collection, news reports on migration issues were more divisive than on LGBTQ+. The final observation concerns the difference in comment length. The median for comment length for the studied subcorpora is 11 tokens for NON-SUD, 20 tokens for OFFENSIVE and 11 tokens for VIOLENT, which shows that offensive comments are longer than the violent and the acceptable ones. Performing the Kruskal-Wallis statistical non-parametric test over the length of the comments of the three subcorpora shows that the OFFENSIVE comments are statistically significantly longer than the comments in the remaining two subcorpora with a p-value < 0.001. This indicates that offensive comments could be more descriptive.

| | | VIOLENT | % | OFFENSIVE | % | NON-SUD | % | TOTAL | % |
|---|---|---|---|---|---|---|---|---|---|
| **MIGRANTS** | # of comments | 381 | 4.09 | 2,945 | 31.64 | 2,624 | 28.19 | 5,950 | 63.93 |
| | # of tokens | 6,848 | 2.77 | 96,807 | 39.17 | 48,862 | 19.77 | 15,2517 | 61.72 |
| **LGBTQ** | # of comments | 88 | 0.95 | 1,434 | 15.41 | 1,835 | 19.72 | 3,357 | 36.07 |
| | # of tokens | 1,941 | 0.79 | 48,612 | 19.67 | 44,060 | 17.83 | 94,613 | 38.28 |
| **TOTAL** | # of comments | 469 | 5.04 | 4,379 | 47.05 | 4,459 | 47.91 | 9,307 | 100.00 |
| | # of tokens | 8,789 | 3.56 | 145,419 | 58.84 | 92,922 | 37.60 | 247,130 | 100.00 |

Table 1: Dataset structure

## 4. Study design

In this study, we analyze the data for the three subcorpora separately, and then jointly compare the VIOLENT and OFFENSIVE subcorpora (SUD) against the NON-SUD subcorpus. To this end, we use well-established corpus analysis techniques that are built into the noSketch Engine concordancer, namely frequency and keyword lists (generated for MSD tags) for quantitative analysis, and concordances and collocation candidates (McEnery and Hardie, 2011) for qualitative analysis. For comparative purposes, we observe the data on the scale which delineates the increasing severity of SUD: non-SUD → offensive → violent.

[1] Due to the limitations of the Facebook Terms of Service, the FRENK corpus cannot be made publicly available.
[2] http://clarin.si/noske/
[3] All examples in this paper are taken from our dataset which is in Slovene. The translations were made by the authors.
[4] https://www.facebook.com/communitystandards/hate_speech

The quantitative part of the analysis includes obtaining data regarding frequency and keyness of grammatical tags which is extracted both for the part-of-speech (henceforth PoS) and fine-grained morphosyntactic (henceforth MSD) level. Due to the considerable difference in the size of the subcorpora (see Table 1), the calculations take into account relative frequencies. For the computation of key grammatical tags for the VIOLENT/OFFENSIVE subcorpus, the NON-SUD subcorpus is used as the reference. We analyze 5 top-ranking key PoS tags and 50 top-ranking key MSD tags, using the MULTEXT-East Slovene Morphosyntactic Specifications[5] (Erjavec, 2012). We then perform additional frequency analyses of prominent features determined on the basis of their frequency or keyness score (e.g., for demonstrative pronouns, imperative verb forms, prepositions). To avoid erroneously labeling features as SUD-specific, we check them against the GigaFida 2.0 representative corpus of standard Slovene (Krek et al., 2020). In addition, we compare our findings with the Janes corpus of Slovene netspeak (Fišer et al., 2020) in order to differentiate between characteristics of general netspeak and online SUD.[6] Even though both the GigaFida and the Janes corpus each contain a wide range of genres within their scope, the techniques we employ (frequency and keyness) extract the most characteristic features of standard written Slovene overall on the one hand and of Slovene netspeak on the other.

With the qualitative approach we further explore prominent features defined on the basis of keyness analysis by studying concordances and collocations of key grammatical forms (e.g., prepositions, numerals, non-finite verb forms), as well as by manually categorizing 100 random concordances of the selected word form into more fine-grained grammatical categories which are not provided by the MULTEXT-East morphosyntactic tagset. Based on the categorization by Toporišič (2004, 2007), we use the following categories:

- for proper nouns: *proper* (with a subcategory *personal*), *geographical*, *material*, *error;*
- for demonstrative pronouns: *classifying* (e.g., *ta* /*this*/), *qualifying* (e.g., *tak* /*such*/), *possessive* (e.g., *tega* /*from this one*/), *quantifying* (e.g., *toliko* /*so much*/);
- for participial adjectives: *demonstrative* (ending on -n/-t/-l; e.g., *zablokiran* /*blocked*/) and *qualifying* (ending in -č/-ši; e.g. *bodeč* /*barbed*/).

## 5. Results

This section presents the results of the grammatical analysis of SUD on the PoS and MSD level.

### 5.1. PoS level

The analysis of SUD on the PoS level takes into account frequencies of word classes and the 5 top-ranking key PoS tags which were computed for the OFFENSIVE and VIOLENT subcorpora with respect to NON-SUD subcorpus and vice versa.

#### 5.1.1. Offensive comments

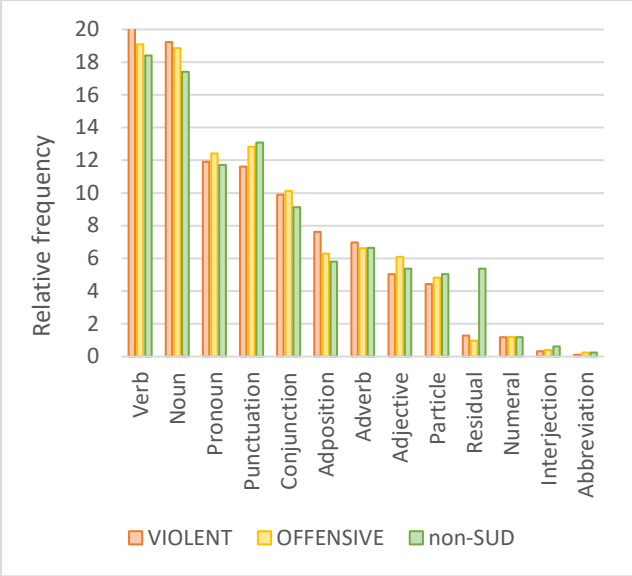



Figure 1: Relative frequency distribution of PoS for the three subcorpora. It is arranged in a descending order according to the values of the VIOLENT subcorpus.

| | OFFENSIVE | VIOLENT |
|---|---|---|
| the most key | Adjective | Adposition |
| | Conjunction | Verb |
| | Adposition | Noun |
| | Noun | Conjunction |
| the least key | Pronoun | Adverb |
| the least key | Adverb | Punctuation |
| | Punctuation | Particle |
| | Particle | Interjection |
| | Interjection | Abbreviation |
| the most key | Residual | Residual |
| | **NON-SUD** | |

Table 2: The 5 top-ranking key PoS tags (OFFENSIVE/VIOLENT vs. NON-SUD and vice versa). The blue cells indicate the function word classes.

Figure 1 shows that the PoS frequency distribution within the OFFENSIVE and the NON-SUD subcorpus is very similar. In fact, the two subcorpora only differ in the ranking of the low-frequency word class, namely residuals[7] which are 5.6 times more frequent in the NON-SUD subcorpus. Furthermore, in the OFFENSIVE subcorpus there are more nouns and conjunctions than in the NON-SUD subcorpus, and more punctuation marks and adjectives than in the VIOLENT subcorpus.

A more detailed frequency analysis shows that the main residual subclass in all subcorpora is the Xf-subclass which represents foreign words. There are a quarter more foreign

[5] http://nl.ijs.si/ME/V6/msd/html/msd-sl.html#msd.msds-sl

[6] It should be noted that different taggers have been used for FRENK, Janes and GigaFida 2.0. However, since the same training data was used for all the taggers, we do not expect any differences in the overall distribution of tags to be due to the tagger used, but rather due to the difference in texts.

[7] These are the elements that do not belong to any other category. In our case mainly strings of foreign words, emoticons, emojis, URLs, hashtags, etc.

elements in the NON-SUD subcorpus in contrast to the OFFENSIVE (and the VIOLENT) subcorpus. However, the opposite is observed for the second most frequent residual subclass, namely the Xe-subclass which denotes emoticons and emojis[8] where there are a quarter fewer emotive elements in the NON-SUD subcorpus in relation to the OFFENSIVE (and VIOLENT) subcorpus.

As expected, these findings are confirmed by the analysis of the 5 top-ranking key word classes for the OFFENSIVE and VIOLENT subcorpus. In addition to what we have seen from Chart 1, Table 2 shows that offensive comments compared to non-SUD comments are also characterized by adjectives, adpositions and pronouns.

#### 5.1.2. Violent comments

While in the OFFENSIVE subcorpus, the PoS distribution differences in relation to the NON-SUD subcorpus can be detected only in the lower part of the PoS frequency list, differences in the VIOLENT subcorpus appear over the entire PoS list. As Chart 1 shows, there are more pronouns than punctuation marks, more adpositions than adverbs and more particles than residual elements in the VIOLENT subcorpus when compared to the word class distribution in the NON-SUD subcorpus. Furthermore, Figure 1 shows that in the VIOLENT subcorpus, there are more verbs, nouns and adpositions than in the NON-SUD subcorpus, and more verbs and adpositions than in the OFFENSIVE subcorpus. It can also be observed that there are fewer adjectives, punctuation marks and certain functions words (esp. conjunctions, pronouns and particles) in the VIOLENT subcorpus in relation to the OFFENSIVE subcorpus. Keyness analysis of word classes for the VIOLENT subcorpus presented in Table 2 confirms these results and extends the set of typical word classes to conjunctions and adverbs.

### 5.2. MSD level

The analysis on the MSD level took into account 50 top-ranking key MSD tags which were then organized and analyzed according to parts-of-speech. The keyness of the MSD tags was extracted for the VIOLENT and OFFENSIVE subcorpora with respect to the NON-SUD subcorpus. An excerpt is shown in Table 3.

| **OFFENSIVE** | | | | | |
|---|---|---|---|---|---|
| **Tag** | **Description** | **Example /EN/** | **Freq** | **Freq/mill** | **Score** |
| Appnpn | Participial adjective, positive, neuter, plural, nominative | vzgojena /raised/ | 13 | 89.4 | 90.4 |
| Ps3mpap | Possessive pronoun, 3rd person, masculine, plural, accusative, plural possessor | njihove /their/[9] | 12 | 82.5 | 83.5 |
| Ps1nsap | Possessive pron., 1stp., neuter, sing., accus., plu.possessor | naše /our/ | 11 | 75.6 | 76.6 |
| Npfpa | Proper noun, feminine, plural, accusative | ZDA /the USA/ | 11 | 75.6 | 76.6 |
| Ps1msip | Possessive pron., 1st p., masc., sing., instr., plu. possessor | našim /our/ | 9 | 61.9 | 62.9 |
| **VIOLENT** | | | | | |
| Vmen | Main verb, perfective, infinitive | zapret /close/ | 233 | 26,510.4 | 5.9 |
| Appfsa | Participial adjective, positive, fem., singular, accusative | bodečo /barbed/ | 7 | 796.5 | 5.7 |
| Pd-fsa | Demonstrative pronoun, fem., singular, accusative | to /this/ | 20 | 2,275.6 | 4.6 |
| Pd-mpa | Demonstrative pronoun, masc., plural, accusative | te /these/ | 16 | 1,820.5 | 4.6 |
| Pd-mpi | Demonstrative pronoun, masc., plural, instrumental | takimi /these/ | 5 | 568.9 | 3.8 |

Table 3: The 5 top-ranking key MSD tags.

#### 5.2.1. Offensive comments

The 50 top-ranking key MSD tags of the OFFENSIVE subcorpus include six word classes in the order presented below.

**Adjective.** The most key MSD tag is the participial adjective in plural. By analyzing 100 random participial adjectives (e.g., *odprt* /*open*/), we found that the great majority of them are descriptive adjectives (e.g., *požgan* /*burnt*/), and only 2 are qualifying adjectives (e.g., *vladajoč* /*governing*/). Among all key adjectival MSD tags, the most frequent ones are the possessive adjectives which are followed by the already mentioned participial and general adjectives (e.g., *primeren* /*appropriate*/). Key adjectives appear in singular and plural form.

**Pronoun.** On the list of 50 most key MSD tags for the OFFENSIVE subcorpus, the pronoun appears in the second place. In addition, pronominal MSD tags are also the most numerous group of tags among the 50 top-ranking key tags. The key pronominal MSD tags include the following pronoun types[10] (indicated in the descending order by their keyness score): possessive pronouns, general, indefinite, reflexive, demonstrative and personal pronouns. The most key pronominal MSD tag is the possessive pronoun. Possessive pronouns mainly express first person plural possession, followed by third and second person plural and second person singular. By taking into account the ratio of actual and possible forms for a particular pronoun type, we see that the most prominent pronoun type are demonstrative pronouns, followed by general and indefinite pronouns. Key demonstrative pronouns show a balanced distribution for gender, and they appear in singular and plural.

We supplemented this results with data from the Janes corpus, which represents Slovene netspeak, and from

[8] Emoticon is a representation of a facial expression with a combination of standard keyboard characters (e.g., :-)), whereas an emoji is a character that represents facial expressions, emotions, other notions or objects in a form of a symbol or icon (e.g., ☺).

[9] Despite *their* being a possessive determiner in English and not a pronoun as written in *Description*, the translation reflects usage in Slovene.

[10] We use *tag* to refer to the complete MSD tag (e.g., Ps3mpap), *type* refers to the next lower level of division (e.g., possessive pronouns), and *form* refers to the third level of division (e.g., 1st person singular possessive pronoun).

GigaFida 2.0 corpus, which represents standard Slovene. A comparison of lemma frequencies of demonstrative pronouns[11] shows that the most frequent ones in all three datasets are classifying and qualifying demonstrative pronouns. However, it could be observed that the share of the qualifying demonstrative pronouns is the highest in the OFFENSIVE subcorpus (1.1/1.5/1.6 times higher than in the NON-SUD/Janes/GigaFida 2.0 corpus).

**Noun.** Nouns appear third on the list of 50 top-ranking key MSD tags after adjectives and pronouns. The most key nominal MSD tag represents proper nouns. There are 64% of proper nouns in the nominative case (vs. 73% in NON-SUD), 14% in the locative (vs. 12% in NON-SUD), 10% in the accusative (vs. 6% in NON-SUD), 9% in the genitive (vs. 7% in NON-SUD), and finally 2% in the dative (vs. 1% in NON-SUD) and 1% in the instrumental case (vs. 0,9% in NON-SUD). By manually analyzing 100 random concordances of proper nouns in the OFFENSIVE subcorpus, we found that the majority belongs to the class of proper names of which 71% represent personal names (and 60% of those represent direct address of other commenters), while there are 1.7 times fewer geographical names. The proper noun MSD tag is followed by the common noun MSD tag which has a much lower keyness score.

**Verb.** Verbal MSD tags represent main verbs and have a high keyness score (they appear in the top quarter of the list). The most key verb type is the participle which is used to form composed grammatical tenses (past and future verb tense) and modes. Further analysis of tense distribution in the OFFENSIVE subcorpus showed (1) that the present tense is more used than the past or future tense, (2) that there are 1.4 times more past tense verb forms than future tense forms, and (3) that 3rd person verb forms are 3–5 times more frequent in all three verb tenses with respect to 1st or 2nd person verb forms. Apart from participles, key verb MSD tags also include two finite verb forms: the negative "to have" in 2nd person singular present tense, and the imperative. Collocations for the 2nd person negative *imeti/to have* show that this verb is predominantly used in constructions that build up the argument of the statement on the basis of discreditation of the addressee (e.g., *nimaš pojma /you don't have a clue/*). The most frequent imperatives proved to be the same as in standard Slovene as represented in the GigaFida 2.0 reference corpus (e.g., *iti /go/, verjeti /believe/, pomagati /help/*). From the list of collocation candidates, it could be observed that the most frequent imperative *iti/to go* predominantly appears in offensive constructions (e.g., *pejdi u kurac; idi od koder si prišu /go fuck yourself, go back from where you came from/*).

**Numeral.** There are three key numeral MSD tags on the list of 50 top-ranking key MSD tags. The most key numeral MSD tag appears in the upper half of the list. All key numeral MSD tags indicate numerals that are spelled-out as opposed to written in digits. Concordances for all three key forms show that the lemma of the most frequent numeral is *prvi /first/* which is in majority of cases used to underline the statement (*Haha pa ti **prvi** svoje ne spostujes /Haha but you're **the first** one not respecting your lot/*). Considering its frequency, it is followed by the numeral *drug /other/* in all of the concordance examples (*Moje življensko vodilo pa je med **drugim** tudi to, da ne solim pameti /Among **others**, my motto is also that I don't try to be a smart-ass./*) and is used to indicate indefiniteness.

**Adverb.** On the list with the 50 top-ranked key MSD tags, we find only one type of adverb, namely the participial adverb, and it appears as the last item on the list. Concordances for this particular tag show that in all but one case the adverb is *sodeč/judging from/*[12] and it is mainly combined with 2nd person verb forms. If we check the use of this adverb in standard Slovene as represented in the GigaFida 2.0 reference corpus, we can see that this adverb predominantly appears in printed periodicals and is used to accompany an argument.

### 5.2.2. Violent comments

The 50 top-ranking key MSD tags of the VIOLENT subcorpus include six word classes in the order presented below.

**Verb.** The list of top 50 key MSD tags includes both auxiliary and main verbs. The verb category is represented by the most key MSD tag of the VIOLENT corpus, namely the infinitive of a main verb, and by the highest number of different MSD tags on the list. Almost half of the key verb MSD tags on the list represent non-finite verb forms, that is participles and infinitives. Collocation analysis for these two verb types show that the basic meaning of the participles is non-hateful (e.g., *priti /come/, narediti /do/, morati /must/, delati /work/*), it is rather the context that renders the comment socially unacceptable, while the infinitives directly express violent actions (e.g., *streljati /shoot/, poslati /send/, pobiti /kill/, zapreti /close, imprison/*). The grammatical tense distribution analysis for the finite verb forms shows that the present tense prevails heavily and that there are 1.5 times more future tense verbs forms than past tense verb forms. Furthermore, we can observe a considerably higher frequency of the 3rd person verb forms across the verbal tenses in comparison to 1st person verb forms (which are from 3 to 7 times less frequent) and 2nd person verb forms (which are from 7 to 30 times less frequent).

**Adjective.** The adjective is the second-highest ranked key word class for the VIOLENT subcorpus. The most key adjectival type is the participial adjective (e.g., *imenovan /named/*). By analyzing 100 random participial adjectives, we found that the great majority belongs to the category of descriptive adjectives (e.g., *ubit /killed/*), and only 7 to qualifying adjectives (e.g., *bodeč /barbed/*). Among all key tags that are denoting adjectives, the most frequent ones are the general adjectives in positive form which are followed by participial adjectives. Key adjectives appear in singular and plural form.

**Pronoun.** On the list of 50 top-ranking key MSD tags for the VIOLENT subcorpus, the first pronominal MSD tag appears in the third place. Among the key pronominal MSD tags, we can find the following pronoun types, arranged in descending order by their keyness score: demonstrative pronouns, personal, general and reflexive pronouns. The most key MSD tag represents the singular demonstrative pronoun in accusative case. Demonstrative pronouns also have the highest ratio of actual and possible forms for a

[11] The analysis took into account the 5 most frequent demonstrative pronouns as they appear in the GigaFida 2.0 corpus. We did not control for their nominal or adjectival use.

[12] We are ignoring the orthographical variations.

particular pronoun type. They mainly appear in plural and the accusative case. In the VIOLENT subcorpus (like in the NON-SUD subcorpus, Janes corpus of Slovene netspeak, and GigaFida 2.0 reference corpus of standard Slovene), the most frequent demonstrative pronouns are classifying and qualifying demonstrative pronouns, but the share of the latter is the highest in the VIOLENT subcorpus (1.004/1.3/1.4 times higher than in the NON-SUD/Janes/GigaFida 2.0). Personal pronouns in the VIOLENT subcorpus mainly represent 3rd person plural, the remaining forms indicate 1st person plural and 3rd person singular. Finally, general pronouns indicate singular and plural, with the plural form appearing in the top quarter of the list of top 50 key MSD tags.

**Numeral.** There are two key numeral MSD tags on the list of 50 top-ranking key MSD tags. The most key numeral MSD tag appears in the top quarter of the list. Both key forms indicate numerals that are spelled-out as opposed to written in digits. The analysis of concordances shows that the lemma of the most frequent numeral is *en* /*one*/ which is used to stress the statement (e.g., *te islamiste bi na križ nabijala* ***enga*** *po* ***enga*** /*I would crucify those Islamists* ***one*** *by* ***one***/), or express indefiniteness and generalization (e.g., *ja* ***en*** *dan jim bo trebo tistole raketo ubrnt na glavo* /*well* ***one*** *day that missile should be directed to them*/. The next key numeral, namely *drug* /*other*/ (e.g., *ne zaslužjo si* ***druzga*** /*they don't deserve* ***other***/), also indicates indefiniteness.

**Noun.** Nouns appear in the upper half on the list of the top 50 key MSD tags for the VIOLENT subcorpus. They are also among the three most prominent word classes on this list with regard to the number of different MSD tags (after verbs and pronouns). The big majority belongs to the group of common nouns and only one key MSD tag represents proper nouns. There are 59% of proper nouns in the nominative case, 16% in the locative, 12% in the accusative, 10% in the genitive, and finally 2% in the instrumental and 1% in the dative case. By manually analyzing 100 random concordances of proper nouns in the VIOLENT subcorpus, we found that geographical names are only slightly (1.2 times) less frequent than proper names, 76% of which represent personal names (and 53% of those represent direct address of other commenters).

**Preposition.** On the list of 50 top-ranking key MSD tags, we find one tag indicating prepositions. The MSD tag appears in the lower half of the list and represents prepositions in the accusative case. The most frequent forms are *v* /*in*, *to*/, *na* /*on*/ and *za* /for, behind/. A comparison with GigaFida 2.0 reference corpus of standard Slovene showed that these prepositions are in fact among the most frequently used prepositions in Slovene, but the normalized frequency preposition in the accusative case actually proved to be more frequent in the VIOLENT corpus (33,109.57 per million for VIOLENT subcorpus against 23,665.94 per million for the GigaFida 2.0 corpus). In addition, collocation analysis of prepositions in the VIOLENT subcorpus showed that prepositions are typically used in socially unacceptable constructions (e.g., *na mejo* /*on the border*/, *v glavo* /*to the head*/, *v morje* /*into the sea*/, *v rit* /*in the ass*/).

## 6. Discussion

This section discusses the results of the study and is based on the comparison of characteristics between offensive and violent comments with the non-SUD comments.

### 6.1. SUD against non-SUD on the PoS level

The results show that differences in grammatical structure of the comments increase relatively to the severity of SUD. This means that differences between OFFENSIVE and NON-SUD subcorpora are less pronounced than between VIOLENT and NON-SUD subcorpora. This graduality can be illustrated with the following example: the OFFENSIVE comments differ from NON-SUD comments in having fewer residuals, while VIOLENT comments differ from NON-SUD comments in having fewer residuals and punctuation marks.

We can also see that SUD comments are characterized by adpositions, conjunctions and nouns. In addition to these three word classes, adjectives and pronouns are typical of offensive comments, while verbs and adverbs are typical of violent comments. Furthermore, calculations showed that comments in the OFFENSIVE subcorpus are statistically significantly longer than comments in the other two subcorpora. This reveals greater descriptiveness of offensive comments on the one hand, and greater emphasis on the action in the violent comments on the other. Based on the results, we can also conclude that SUD comments are characterized by content words (verbs, nouns, adjectives, adverbs) and certain function words, i.e., the ones used to form complex discourse (conjunctions, adpositions, pronouns), while function words used as discourse markers (particles, interjection) are more typical of non-SUD comments. This indicates more spontaneous, dialogic nature of non-SUD comments, and more complex, argumentative nature of SUD comments.

### 6.2. SUD against NON-SUD on the MSD level

The 50 top-ranking key MSD tags of the two SUD subcorpora computed against the NON-SUD subcorpus both include six word classes of which the majority (5 out of 6) is shared: adjectives, nouns, numerals, pronouns and verbs. The 6th key word class represents adpositions for the VIOLENT subcorpus, and adverbs for the OFFENSIVE subcorpus. Despite the similar set of word classes among the top 50 key MSD tags, their frequency and distribution show that adjectives and pronouns prevail in the OFFENSIVE subcorpus, while nouns, pronouns and verbs appear to be the most key features of the VIOLENT subcorpus. This is another confirmation that violent comments lexicalize action, while offensive comments are more descriptive. In addition, the results suggest that violent comments typically contain features that can convey generalizations, while typical features of offensive corpus exhibit a tendency towards dialogism. These characteristics, as will be shown in the following paragraphs, are manifested on the level of several grammatical forms which were identified as key forms of SUD.

**Descriptiveness and dialogic style.** The descriptiveness of the offensive comments is first shown in pronouns where we can observe high keyness of possessive and demonstrative pronouns. Possessive pronouns only appear as a key pronoun type in the OFFENSIVE subcorpus and they mainly represent 1st person plural possession, followed by 3rd and 2nd person plural possession. Key demonstrative pronouns appear in both

SUD subcorpora and are of two types: classifying (e.g., *ta* /*this*/) and qualifying (e.g., *tak* /*such*/). However, we could observe that the share of qualifying demonstrative pronouns is higher in offensive comments in comparison to the violent ones, which once more indicates a greater descriptiveness of offensive comments.

Next, descriptiveness is also shown on the level of adjectives. The most key MSD form of the OFFENSIVE subcorpus is the participial adjective. This adjective type is among key MSD tags also in the VIOLENT subcorpus, but detailed analysis shows that the ratio between descriptive and qualifying participial adjectives is higher in the OFFENSIVE subcorpus. In addition, similar to pronouns, possessive adjectives are only typical of the OFFENSIVE subcorpus.

Next, we can observe descriptiveness on the level of nouns. Proper nouns appear high on the list of key MSD tags of the OFFENSIVE subcorpus, while in the VIOLENT subcorpus the higher keyness pertains to common nouns. Although proper nouns do not directly indicate a more descriptive style, they correlate to possessive pronouns and adjectives and thus contribute to higher descriptiveness. In summary, the keyness of demonstrative pronouns, especially qualifying ones, shows that referring to the quality of the referent is an important argumentation strategy in offensive comments and SUD in general. Furthermore, the keyness of proper nouns, which also have a higher proportion of direct address than proper nouns in violent comments, and high keyness of forms expressing possession, which only appear in offensive comments, reveal that offensive comments are more descriptive and dialogic.

A more pronounced dialogic nature of offensive comments with respect to violent comments can also be shown on the basis of grammatical case and person used in the two SUD subcorpora. Offensive comments exhibit more prominent use of nouns in the nominal and the dative case than violent comments. The nominative usually refers to the responsible entity, while the dative is likely to focus the discourse on purposes and opponents, rather than on actions (Fidler and Cvrček, 2019). In addition, 2nd person forms appear as key pronoun and verbal forms only in the offensive comments, and overall, there are more 2nd person verb forms in the OFFENSIVE subcorpus. In fact, there are considerably fewer 2nd person verb forms in the VIOLENT subcorpus with regard to the 1st/3rd person verb forms on one hand, and with regard to OFFENSIVE and NON-SUD subcorpus on the other. The first two grammatical persons indicate the speech participants, more specifically the first one refers to the speaker and the second person to the addressee, while the third person marks the non-active participants (Toporišič, 2004).

As mentioned above, the dialogic nature of offensive comments is shown on the level of verbs where we can observe more finite verb forms than in violent comments. Despite the fact that in both SUD subcorpora the most key verb tag is actually a non-finite verb, a closer look reveals that this non-finite form in the VIOLENT corpus represents infinitives, while in the OFFENSIVE subcorpus it is the participle which is used to form tenses and modes and is therefore an integral part of finite verb forms. Dialogic style and descriptiveness are also manifested through the use of the participial adverb which is a key MSD tag specific to OFFENSIVE subcorpus. The predominant form of the adverb (i.e., *sodeč* /*judging from*/) is mainly used in combination with 2nd person verb forms. In general standard Slovene (as represented in the GigaFida 2.0 corpus), this adverb is predominantly used in printed periodicals, which makes it interesting that such adverb should appear as a key feature of offensive online comments. This suggests that users try to objectivize and reinforce their argument by mirroring the style of printed publications (***Sodec*** *po tvojih stavkih si ti nepismena in potrebna stara kura* /***Judging*** *from your sentences you are the illiterate and horny old hen*/).

**Generalization and action-orientation.** Typical grammatical features of violent comments indicate generalizations and incitement to action. First, generalization can be observed on the level of the grammatical person. The analysis showed a high proportion of 3rd person personal pronouns, and 3rd person verbal forms (in comparison to 1st and 2nd person forms, as well as in comparison to offensive and non-SUD comments). The third grammatical person indicates the non-participant of the conversation and can have human or non-human referent (Toporišič, 2004). As such it can be used to express generalizations, a type of categorization, that can be ideologically consequential since generic *they* often refers to people viewed as outsiders (Scheibman, 2007). The keyness of the (1st and) 3rd person pronouns might thus be explained with the concept of othering which states that SUD is built on the opposition of the in-group and out-group. This especially pertains to violent comments where we can observe a stronger presence of 1st and 3rd person verb forms (as opposed to offensive comments where the 2nd person is as frequent as 1st person) which shows that the division between the two groups stand out more in the violent comments.

In addition, prevalence of the 3rd person verb forms coupled with the finding on predominance of qualifying demonstrative pronouns explained above further suggests that the non-active participant of the discourse and their characteristics are in the spotlight of the comment. Furthermore, strong presence of 1st person plural verb forms in the VIOLENT subcorpus with respect to OFFENSIVE and NON-SUD subcorpus »inherently reports opaque referentiality« (Fidler and Cvrček, 2019). The tendency for generalization is also shown through the use of general pronouns and numerals that were identified as key forms in both SUD subcorpora. However, key numeral forms are, apart from those that convey indefiniteness and generalization, also used to underline the statement and could as such be used to underline the action.

We have claimed that violent comments are more action-oriented. The analysis showed that almost half of all key verb forms are represented by non-finite verb forms, namely participles and infinitives. In fact, the infinitival verb form represents the most key MSD tag of the VIOLENT subcorpus. A closer look at these forms reveals that the participles describe neutral processes, while the infinitives evoke violent actions. Action-orientation as well as the tendency for generalizations can be observed through the use of non-finite verb forms since finite verbs provide explicit reference and option for feedback, while in noun-heavy, nominal phrases the action is taken for granted (Fidler and Cvrček, 2019). The latter could therefore also be applied to phrases with non-finite verbs forms. Action-orientation can further be supported with the results of frequency analysis of verbal tenses which shows that there is from 1.6 to almost 2 times more future tense in the

VIOLENT subcorpus with regard to the OFFENSIVE or NON-SUD subcorpus, respectively.

Action-orientation once more, despite less prominently, manifests through the use of grammatical person. We observed relatively high keyness of 1st person singular verb forms (esp. future tense forms) which points to explicit speaker agency and their intention to claim credit for their actions (Fidler and Cvrček, 2019). In addition, action-orientation can also be detected through the use of numeral forms that underline the meaning and through the use of prepositions which appear more frequently in violent comments than in general Slovene. The analysis showed that violent comments are characterized by prepositions in the accusative case. Closer examination revealed that they are more frequent in violent comments than in general standard Slovene and that they are predominantly used to form inappropriate constructions. This might be a promising finding with regard to automated detection of SUD.

**Graduality**. Based on our results, we were able to detect a gradual amplification of observed phenomena proportionately to the severity of SUD (i.e., non-SUD→offensive→violent comments). First, a comparative analysis of the grammatical case of proper nouns showed that the share of all of the cases individually gradually increases from the socially acceptable towards the least socially acceptable comments, except for the nominative case where the share increases in the opposite direction. The only deviation is the dative case which is most prominently used in the OFFENSIVE subcorpus. Next, comparative analysis of proper nouns showed that in all three subcorpora, proper nouns include more personal names than geographical names. The share of personal names (including direct individual addresses) decreases in the direction from socially acceptable towards the least socially acceptable comments, while the share of geographical names increases in this same direction. Both grammatical case and proper noun distribution support the claim that SUD comments in general are less dialogic than non-SUD.

In addition, graduality can be observed in the distribution of the grammatical person. Analysis of grammatical person distribution among verbal forms showed that despite the fact that 3rd person verb forms predominate in all three subcorpora, the share of 3rd person forms actually increases in the direction from non-SUD to violent comments. On the contrary, the share of 1st person verbal forms is smaller in the two SUD subcorpora than in the non-SUD subcorpus. The 1st person singular verb forms, for example, are most used in non-SUD comments. This is not unexpected since the »first person singular is the prototypical site for expression of speaker point of view« (Scheibman, 2007) with which the speaker overtly takes the responsibility for the statement (Fidler and Cvrček, 2019). It should be noted, however, that since we are analyzing comments to media posts of which a considerable part is dedicated to reporting events, such distribution of grammatical person could be expected to a certain extent. Nevertheless, this cannot fully explain the identified graduality phenomenon.

Next, graduality can be observed in the use of demonstrative pronouns. The analysis showed that the less formal the language is (i.e., from general standard language (GigaFida 2.0) towards netspeak (Janes corpus), non-SUD, offensive and violent comments), the higher the share of qualifying demonstrative pronouns is. We should note, however, that comparison of the results with the non-SUD subcorpus showed that the degree of social unacceptability pertaining to the comments under scrutiny might not be the sole influencing factor. The fact that the share of qualifying demonstrative pronouns is only slightly higher in SUD than in non-SUD comments indicates that the discourse topic might be impacting the results as well.

Throughout our analysis of offensive and violent comments we have noticed that graduality also applies to the use of future tense since the share of future tense in violent comments strongly exceeds the share in offensive as well as non-SUD comments. In addition, we observed that the forms in dual number are only typical of offensive comments. Comparative analysis once more confirmed the phenomenon of graduality: if there are 0.27% of verb forms in dual[13] in the NON-SUD subcorpus, there are 0.14% per million in the OFFENSIVE subcorpus and only 0.11% in the VIOLENT subcorpus. In addition, this again shows that offensive comments are more dialogic and personal than violent comments.

## 7. Conclusion

The aim of this paper was to produce a grammatical description of SUD on the level of word classes and morphosyntactic labels in order to answer the question whether SUD is realized through any specific grammatical features. To this end, we performed a comparative analysis of the FRENK corpus that contains Facebook comments and contrasted two SUD subcorpora (offensive and violent comments), and a NON-SUD subcorpus.

We found that the grammatical skeleton of SUD has its distinctive characteristics in comparison to non-SUD, but that the differences are subtle. Furthermore, we showed that these differences increase proportionately to the degree of social unacceptability of comments. The more we move away from socially acceptable discourse, the larger the set of specific grammatical features or the larger their share. For example, frequency-wise, internal word class distribution of offensive comments differs from that of non-SUD comments only in the representation of residuals, while violent comment-specific are also pronouns and prepositions. Likewise, the graduality can be observed, for example, in the use of qualifying demonstrative pronouns or future tense where their frequency increases proportionately to the intensity of SUD in the comments.

The analysis shows that offensive comments have greater similarities with non-SUD than with violent comments. We could observe that with respect to SUD comments, non-SUD comments are characterized by discursive function words. On the other hand, content words and function words used to form complex sentences prevail in SUD comments. Based on the analysis of key morphosyntactic tags, we can conclude that offensive comments tend to be more dialogic and descriptive than violent comments (e.g., by using more 2nd person and possessive forms), whereas violent comments exhibit a greater inclination toward generalization and action-

[13] Other word classes also agree in number and could therefore be used to check the frequency of grammatical forms in dual number. We opted for verbs because forms in dual are more frequently preserved with verbs that for example with nouns.

orientation (e.g., by using more infinitives and future tense).

This study does not offer a full description of grammatical structure of SUD, but rather indicates the tendencies of online SUD on the morphosyntactic level. The main limitation pertains to the small size of the VIOLENT subcorpus which prevented generalizations of certain infrequent phenomena but the issues of size were taken into account in study design and the selection of the analytical techniques used. It should also be noted that the dataset was automatically tagged which does not ensure error-free output, but based on manual inspection of random samples of the annotated concordances in the qualitative analysis, the quality of the output was sufficient for this study.

The main contribution of this study is that it shows that the grammatical footprint of SUD is different from non-SUD (e.g., prevalence of future tense and 3rd person verbal forms, higher frequency of prepositions, prominent use of qualifying demonstrative pronouns and infinitives). This could be a promising finding for the development of automated detection of SUD. In our future work, we envisage supplementing these results with lexico-syntactic analysis and observations of the phenomenon at higher levels of linguistic descriptions. Furthermore, based on the prominent features from this study, we plan to analyse certain stylistic aspects of SUD, such as the expression of agency.

## 8. Acknowledgement

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project »Resources, methods, and tools for the understanding, identification, and classification of various forms of socially unacceptable discourse in the information society« (J7-8280, 2017–2021), the Slovenian-Flemish bilateral basic research project »Linguistic landscape of hate speech on social media« (N6-0099, 2019–2023) , and ARRS P6-0215 research program (Slovene language – basic, contrastive, and applied studies).

## 9. References

Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L Williams. 2019. “The Enemy Among Us” Detecting Cyber Hate Speech with Threats-based Othering Language Embeddings. *ACM Trans. Web TWEB* 13(3), pp. 1–26.

Tor Berglind, Björn Pelzer, and Lisa Kaati. 2019. Levels of hate in online environments. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Vancouver, BC, Canada, pp. 842-847.

Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Sci.* 5(11).

Isobelle Clarke and Jack Grieve. 2017. Dimensions of abusive language on twitter. In: *Proceedings of the First Workshop on Abusive Language Online.* Vancouver, BC, Canada, pp. 1–10.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A Multilingual Evaluation for Online Hate Speech Detection. *ACM Trans. Internet Technol. TOIT* 20(2), pp. 1–22.

Karmen Erjavec and Melita Poler Kovačič. 2012. “You Don’t Understand, This is a New War!” Analysis of Hate Speech in News Web Sites’ Comments. *Mass Commun. Soc.* 15(6), pp. 899–920.

Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Lang. Resour. Eval.* 46(1), pp. 131–142.

Tomaž Erjavec, Nikola Ljubešić, and Darja Fišer. 2018. Korpus slovenskih spletnih uporabniških vsebin Janes. In: *Viri, orodja in metode za analizo spletne slovenščine*, Fišer, D. (ed.). Znanstvena založba Filozofske fakultete, Ljubljana, pp. 16–43.

Masako Fidler and Václav Cvrček. 2019. Keymorph analysis, or how morphosyntax informs discourse. *Corpus Linguistics and Linguistic Theory*, 15(1), pp. 39–70.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv. CSUR* 51(4), pp. 1–30.

Paul Hopper. 1987. Emergent grammar. Berkeley Linguistics Society 13(1), pp. 139–157.

Sune Qvotrup Jensen. 2011. Othering, identity formation and agency. *Qual. Stud.* 2(2), pp. 63–78.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1(1), pp. 7–36.

Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraz Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. 2020. Gigafida 2.0: The Reference Corpus of Written Standard Slovene. In: *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, France, pp. 3340–3345.

Holly Lakey. 2016. The Grammar of Fear: Morphosyntactic Metaphor in Fear Constructions. Ph.D. thesis, University of Oregon. http://hdl.handle.net/1794/20415

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*, Florence, Italy, pp. 29–34.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. In: Ekštein, K. (ed.) Text, Speech, and Dialogue. TSD 2019. Lecture Notes in Computer Science, vol. 11697. Springer, Cham. pp. 103–114.

Nataša Logar and Tomaž Erjavec. 2018. Strokovno-znanstvena slovenščina: besednovrstne in oblikoskladenjske značilnosti. In: *Proceedings of the conference on Language Technologies & Digital Humanities*, Ljubljana, Slovenia, pp. 175–180.

Tony McEnery and Andrew Hardie. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge: University Press.

Kristina Pahor de Maiti, Darja Fišer, and Nikola Ljubešić. 2020. Nonstandard linguistic features of Slovene socially unacceptable discourse on Facebook. In: Fišer, D. and Smith, P. (eds.): *The Dark Side of Digital*

*Platforms: Linguistic Investigations of Socially Unacceptable Online Discourse Practices*, Ljubljana: University Press, Faculty of Arts, pp. 12–35.

Joanne Scheibman. 2002. Point of view and grammar: structural patterns and subjectivity in American English conversation. Amsterdam: J. Benjamins.

Joanne Scheibman. 2007. Subjective and intersubjective uses of generalizations in English conversations. In: Englebretson, R. (ed.): *Stancetaking in Discourse*, J. Benjamins, pp. 111–138.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Valencia, Spain, pp. 1–10.

Jože Toporišič. 2004. *Slovenska slovnica*. Obzorja: Maribor.

Jože Toporišič. 2007. *Slovenski pravopis*. Slovenska akademija znanosti in umetnosti: Ljubljana.

Bertie Vidgen and Taha Yasseri. 2019. Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1), pp. 66-78.

Ahmed Waqas, Joni Salminen, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. 2019. Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PloS One* 14(9).

Valentin Werner. 2019. Assessing hip-hop discourse: Linguistic realness and styling. *Text Talk* 39(5), pp. 671–698.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-GRU based deep neural network. In: Gangemi et al. (eds): *The Semantic Web.* Springer: Cham, pp. 745–760.

Ana Zwitter Vitez and Darja Fišer. 2015. Novi mediji in govorjena slovenščina: zaznavanje, kodifikacija, analiza. In Smolej, M. (ed.).: *OBDOBJA 34: Slovnica in slovar – aktualni jezikovni opis*, Znanstvena založba Filozofske fakultete: Ljubljana, pp. 881–890.

# Načrtovanje jezikovne komponente generativnega nasprotniškega nevronskega omrežja

**Martin Pernuš, Simon Dobrišek**

Fakulteta za elektrotehniko
Univerza v Ljubljani
Tržaška cesta 25, 1000 Ljubljana
{martin.pernus, simon.dobrisek}@fe.uni-lj.si

**Povzetek**

Področje ustvarjanja slik je kot pomembno področje umetne inteligence v zadnjih letih dobilo velik zagon. Vizualno prepričljivo ustvarjanje slik omogočajo generativni modeli, ki lahko sliko pogojijo tudi z jezikovnim opisom. Model, ki trenutno vrača najboljše rezultate na tem področju je pogojeno generativno nasprotniško omrežje. V tem članku opišemo uspešnejše postopke modeliranja tekstovnega opisa za generativna nasprotniška omrežja, način učenja takih modelov, podatkovne zbirke in metrike za vrednotenje.

**Abstract**

The field of image generation is an important area of artificial intelligence research that gained a lot of traction in the recent years. The most visually promising results are achieved by generative models that can be conditioned on language description. The predominant model in the generative modelling area is conditional generative adversarial network. In this paper we describe the most successful attempts on the combination of language modelling and generative adversarial networks, their training procedures, datasets and evaluation metrics.

## 1. Uvod

Področje računalniškega vida in umetne inteligence je v zadnjih letih doseglo velike uspehe na področju ustvarjanja slik. V ozadju teh rezultatov so generativni modeli globokih nevronskih omrežij (v nadaljevanju generativni modeli). Generativni modeli z modeliranjem ciljne verjetnostne porazdelitve podatkov omogočajo ustvarjanje novih vzorcev. V osnovi so generativni modeli nepogojeni, če pa izhod generativnega modela pogojimo z zunanjim signalom, pa govorimo o pogojenih generativnih modelih.

Danes najbolj uporabljeni generativni modeli so avtoregresivni modeli, pretočni modeli, variacijski avtokodirniki in generativni nasprotniški modeli (angl. Generative Adversarial Network, GAN). Slednji še posebej izstopajo, saj so zmožni ustvarjanja fotorealističnih in vizualno prepričljivih slik različnih objektov, kompleksnih scen in visokoresolucijskih slik, ki jih praktično ni več mogoče razlikovati od resničnih slik. Prav tako so se modeli GAN izkazali kot modeli, ki jim zlahka vnesemo zunanji signal kot pogoj za ustvarjanje ciljne slike z želeno vizualno semantično informacijo. Zunanji signali so lahko različne oblike, lahko je preprost, kot npr. razred izhodne slike, lahko pa tudi bolj kompleksen, kot npr. podroben opis želene izhodne slike. Slednji način obravnavamo v našem članku.

Glavni namen članka je pregled najnovejših metod tistih komponent generativnih modelov, ki skrbijo za opis informacije v jezikovni obliki. V članku se bomo osredotočili na tiste komponente, ki se uporabljajo v modelih GAN, saj ti prevladujejo na področju pogojenih generativnih modelov. Na koncu predlagamo še nekaj možnih rešitev za izboljšanje teh modelov.

## 2. Sorodna literatura

Generativni nasprotniški modeli so bili prvič opisani v članku (Goodfellow et al., 2014), kjer so se avtorji osredotočali na modeliranje obraznih slik brez zunanjih signalov. Nadgradnje prvotne zasnove modela so se v literaturi osredotočale predvsem na spremembo arhitektur modelov in spreminjanje prvotne kriterijske funkcije. V članku (Radford et al., 2015) so predlagali konvolucijsko nevronsko omrežje DCGAN, kar je omogočalo zmanjšanje števila parametrov v celotni arhitekturi ob hkratni izboljšani kvaliteti ustvarjenih slik. S postopnim načinom učenja modela ProGAN (Karras et al., 2018) so prvič ustvarili slike s kar milijon slikovnimi elementi. Model je bil nadgrajen v (Karras et al., 2019a) in (Karras et al., 2019b), kjer je model StyleGAN z vnosom stohastične raznolikosti in napredne arhitekture ustvaril slike obrazov, ki jih praktično ni več mogoče razlikovati od slik resničnih ljudi. Ti modeli so nepogojeni, saj gre za neposredno modeliranje ciljne verjetnostne porazdelitve brez kakršnekoli dodatne informacije in lastnostih na končni sliki. Pri spremembi kriterijske funkcije je bil napredek narejen na področju nenasičenih kriterijskih funkcij, kar v praksi omogoča stabilnejše učenje (Arjovsky et al., 2017), (Mao et al., 2017), (Nowozin et al., 2016), (Gulrajani et al., 2017).

Poleg nepogojenih GAN modelov so se razvili tudi pogojeni modeli GAN, ki izhodno sliko pogojujejo glede na vhodni signal. Model Pix2pix (Isola et al., 2017) je ciljni slog slike oblikoval glede na vhodno sliko. Model GauGAN (Park et al., 2019) je sliko pogojil z visokoresolucijsko sliko, ki lokacijsko opiše želeno semantično informacijo. BigGAN (Brock et al., 2019) je trenutno najboljši pogojeni generativni model, ki ciljno sliko ustvari le glede na informacijo o želenem razredu. Model GeNeVA-GAN (El-Nouby et al., 2019) je tekstovne opise uporabil za postopno dodajanje preprostih geometrijskih objektov na sliko, kot so kocka, krogle in piramide.

Pogojeni modeli, ki so v našem članku najbolj relevantni, pogojujejo zunanjo sliko glede na tekstovni opis. Prvi

večji preboj na tem področju so naredili v članku (Reed et al., 2016b), kjer so ustvarili slike velikosti $64 \times 64$. Nadaljnje delo se je osredotočalo na izboljšanje vizualne prepričljivosti slik in višanje resolucije, kot npr. StackGAN (Zhang et al., 2017), AttnGAN (Xu et al., 2018) in MirrorGAN (Qiao et al., 2019), ki bodo podrobneje opisani v 4. poglavju.

# 3. Ozadje

## 3.1. Generativna nasprotniška omrežja (GAN)

Omrežje GAN sestoji iz dveh modelov: generator $G$ in diskriminator $D$. Celotno omrežje implicitno optimiziramo na tak način, da bo generator $G$ modeliral ciljno verjetnostno porazdelitev (v našem primeru verjetnostno porazdelitev slik) čim boljše. Naloga diskriminatorja $D$ je ločevanje med resničnimi slikami in umetnimi slikami, torej tistimi, ki jih je ustvaril generator. Naloga generatorja $G$ je prelisičiti diskriminator tako, da bo ta domneval, da gre za resnični vzorec. Tako $G$ kot $D$ sta običajno visoko parametrizirani nelinearni parametrični funkciji, ki ju iterativno učimo s postopkom gradientnega spusta. V dani iteraciji bo funkcija $G$ vrnila nov vzorec, funkcija $D$ pa bo ovrednotila verjetnost, da je nek vzorec resničen. Z učenjem bo generator ustvarjal vse boljše slike, diskriminator pa vse bolje ločeval med resničnimi in umetnimi slikami. Po končanem učenju diskriminator zavržemo, generator pa uporabimo za ustvarjanje novih vzorcev.

Matematično lahko učni postopek zapišemo kot igro med dvema igralcema s kriterijsko funkcijo

$$\begin{aligned}&\min_G \max_D V(D,G) = \\ &\mathbb{E}_{x\sim p_{data}} \log D(x) + \mathbb{E}_{z\sim p_z} \log(1 - D(G(z))),\end{aligned} \tag{1}$$

kjer je $p_{data}$ verjetnostna porazdelitev resničnih podatkov, $z$ pa šumni vektor, vzorčen iz preproste verjetnostne porazdelitve $p_z$ (običajno Gaussova porazdelitev, $z \sim \mathcal{N}(0,1)$). V praksi se izkaže, da zaradi izginjajoče odvodne informacije pri gradientnem spustu ob predobro naučenem diskriminatorju generator boljše učimo z maksimizacijo $\log(D(G(z))$ namesto minimizacije $\log(1 - D(G(z))$.

## 3.2. Enodimenzionalna konvolucija

Konvolucija je pogosta operacija v globokih nevronskih omrežjih. Pretežno se uporablja na področju računalniškega vida, kjer uporabljamo dvodimenzionalno konvolucijo. Če za modeliranja teksta uporabljamo konvolucijo, se poslužimo enodimenzionalne različice. Definirana je kot

$$h(y) = \sum_{x=1}^{k} f(x) g(y \cdot d - x + c), \tag{2}$$

kjer sta $f$ in $g$ funkciji z diskretnih vhodom, $k$ velikost jedra, $d$ korak, $c = k - d + 1$ pa odmik. Kot je v nevronskih omrežjih običajno, ima celotno konvolucijsko omrežje množico funkcij $f_{ij}(x)$, imenovanih uteži, na množici vhodov $g_i(x)$ in izhodov $h_j(y)$.

Podobno kot pri nevronskih omrežjih v računalniškem vidu, tudi tu lahko definiramo združevanje aktivacij (angl. pooling), kot npr. maksimalno združevanje in povprečno združevanje, le da namesto na prostorski komponenti to naredimo na časovni komponenti.

## 3.3. Povratno nevronsko omrežje

Za modeliranje jezikovne informacije z nevronskimi omrežji se uporablja strukture, ki so primerne za diskreten tip podatkov. To so običajno povratna nevronska omrežja, ki se uporabljajo pri raznovrstnih nalogah obdelovanja naravnega jezika, npr. ustvarjanje teksta (Sutskever et al., 2011), prevajanje (Sutskever et al., 2014), analiza sentimenta (Tai et al., 2015), ... Povratna nevronska omrežja v primerjavi s klasičnimi nevronskimi omrežji niso omejena na podatke fiksne dolžine, temveč lahko delujejo z zaporedji poljubne dolžine.

Osnovna povratna nevronska omrežja imajo težave s pozabljanjem informacije in stabilnim učenjem, zato se večinoma uporabljata dva posebna modela povratnih nevronskih omrežij: Long Short-Term Memory (LSTM) (Hochreiter in Schmidhuber, 1997) in Gated Recurrent Unit (GRU) (Chung et al., 2014).

Model LSTM je matematično opisan z naslednjimi operacijami

$$\begin{aligned}
i_t &= \sigma(W_{xi}^T x_t + W_{hi}^T h_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}^T x_t + W_{hf}^T h_{t-1} + b_f) \\
o_t &= \sigma(W_{xo}^T x_t + W_{ho} h_{t-1} + b_o) \\
g_t &= \tanh(w_{xg}^T x_t + W_{hg}^T h_{t-1} + b_g) \\
c_t &= f_t \otimes c_{t-1} + i_t \otimes g_t \\
y_t &= h_t = o_t \otimes \tanh(c_t),
\end{aligned} \tag{3}$$

model GRU pa z

$$\begin{aligned}
z_t &= \sigma(W_{xz}^T x_t + W_{hz} h_{t-1} + b_z) \\
r_t &= \sigma(W_{xr}^T x_t + W_{hr} h_{t+1} + b_r) \\
g_t &= \tanh(W_{xg}^T x_t + W_{hg}^T (r_t \otimes h_{t-1}) + b_g) \\
h_t &= z_t \otimes h_{t-1} + (1 - z_t) \otimes g_t,
\end{aligned} \tag{4}$$

kjer matrike $W$ predstavljajo učene uteži modela, $b$ učeni pristranski vektor modela, $\sigma$ sigmoidno funkcijo, tanh hiperbolični tangens, $\otimes$ hadamardov produkt, $x_t$ in $y_t$ pa vhod in izhod modela ob času $t$. Vektorja $c_t$ in $h_t$ predstavljata notranje skrito stanje modela, ki omogoča prenos informacije čez daljše časovno obdobje.

## 3.4. Podatkovne zbirke

Modeli, ki pogojujejo izhodno sliko z jezikovno informacijo, so vizualno prepričljive rezultate dosegli na podatkovnih bazah z omejeno raznolikostjo. Dve taki podatkovni bazi, ki sta obširno uporabljeni, sta Caltech-UCSD Birds (Welinder et al., 2010), ki vsebuje 11.788 slik 200 različnih vrst ptic s desetimi opisi na sliko in Oxford-102 Flowers (Nilsback in Zisserman, 2008), ki vsebuje 8.189 slik 102 različnih vrst rož s petimi opisi na sliko. Podatkovna zbirka, ki vsebuje vsakdanje prizore na slikah, je Microsoft COCO (Lin et al., 2014) s 82.783 slikami, ki vsebuje pet tekstovnih opisov na sliko. Primeri slik s pripadajočimi opisi vsake podatkovne zbirke so prikazani na sliki 1.

Slika 1: Primer para tekstovnega opisa in pripadajoče slike za zbirko Oxford-102 Flowers (na vrhu), Caltech-UCSD Birds (na sredini) in Microsoft COCO (spodaj).

### 3.5. Vrednotenje

Za vrednotenje generativnih modelov se običajno uporabljata metriki Inception Score (Salimans et al., 2016) in Fréchet Inception Distance (Heusel et al., 2017), ki sta namenjeni ocenjevanju vizualne kakovosti in raznovrstnosti nastalih slik. Za ocenjevanje ustreznosti semantične informacije na izhodni sliki glede na tekstovni opis pa se v člankih najpogosteje zatekajo k človeškemu vrednotenju preko raznih spletnih platform. Pomanjkanje avtomatskih metod za vrednotenje semantične skladnosti v tekstovnem opisu in na ustvarjeni sliki je velik problem, saj onemogoča ovrednotenje prispevkov posameznih komponent modela h končnem rezultatu.

Izmed dveh omenjenih metrik se v tekstovno pogojenem ustvarjanju slik večinoma uporablja metrika Inception Score. Definirana je kot

$$\mathrm{IS} = \exp(\mathbb{E}_{\boldsymbol{x}} D_{KL}(p(y|x), p(y))), \tag{5}$$

kjer $x = G(z)$ predstavlja umetno ustvarjeno sliko iz naključnega vektorja $z \sim p_z$, $p(y|x)$ verjetnostna porazdelitev razvrščevalnika Inception (Szegedy et al., 2016) na posamezne razrede, $p(y) = \int p(y|x = G(z))dz$ pa mejna porazdelitev razredov. Običajno se validacijska množica podatkov razdeli na deset podmnožic, IS pa oceni na vsaki podmnožici. Rezultat nato podamo v obliki povprečja in standardne deviacije.

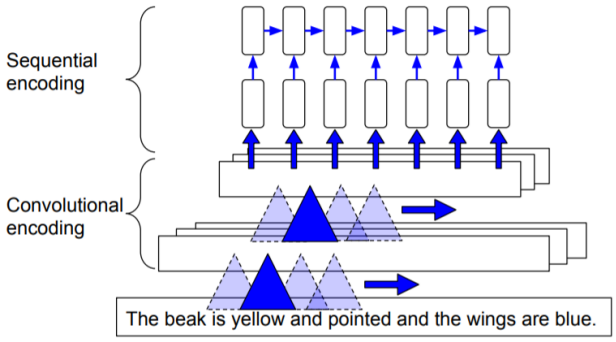
## 4. Jezikovne komponente v generativnih modelih

V nadaljevanju bosta opisani dve najpomembnejši metodi, ki sta uporabljeni kot jezikovna komponenta generativnega nasprotniškega modela. Globoke simetrične strukturne skupne vloženke je metoda, ki za osnovno enoto teksta vzame posamezno črko, globok pozornostni multimodalno podobnostni model pa za osnovno enoto teksta vzame besedo. Po opisu teh metod na kratko opišemo še doprinose najnovejših člankov s tega področja, možnosti uporabe slovenščine, rezultate trenutno najboljših modelov in možne izboljšave na tem področju.

### 4.1. Globoke simetrične strukturne skupne vloženke

Prvi članek, ki je prikazal prepričljivo ustvarjanje slik, pogojenih na njihovih tekstovnih opisi, je članek (Reed et al., 2016b). Za modeliranje tekstovne komponente so predlagali postopek, ki je združeval konvolucijska in povratna nevronska omrežja, imenovan globoke simetrične strukturne skupne vloženke (angl. Deep symmetric structured joint embedding). Metoda učenja tekstovnega modela je osnovana na članku (Reed et al., 2016a).

Postopek uči tekstovni model $\varphi$, slikovni model $\theta$ pa je prednaučen na nalogi razvrščanja podatkovne zbirke ImageNet (Deng et al., 2009) in je v postopku učenja zamrznjen. Tekstovni model $\varphi$ je sestavljen iz začetnih konvolucijskih slojev, ki jim sledi višjenivojski LSTM sloj. Arhitektura modela je prikazana na sliki 2. Motivacija za tako arhitekturo izhaja iz dejstva, da je konvolucijska operacija hitra z nizko računsko kompleksnost, vendar pa se ne zaveda celotnega zaporedja tekstovnega opisa. S kombinacijo obeh arhitektur dosežemo hitro procesiranje zaradi začetnih konvolucijskih operacij, za časovno komponento pa skrbi LSTM, ki obdeluje višjenivojske značilke. Tekstovni model $\varphi$ za osnovno enoto v tekstu vzame posamezno črko, ki jo kodira kot višjedimenzionalni vektor.




Slika 2: Arhitektura konvolucijskega in sekvenčnega kodiranja teksta. Za osnovno enoto vzamemo posamezne črke, ki jih kodiramo s konvolucijsko operacijo, višje-nivojske predstavitve pa s povratnim nevronskim omrežjem. Slika je povzeta po (Reed et al., 2016a).

Končno predstavitev teksta so modelirali kot

$$\varphi(t) = \frac{1}{L} \sum_{i=1}^{L} h_i, \tag{6}$$

kjer je $h_i$ skriti vektor končnega sloja povratnega nevronskega omrežja, $L$ pa dolžina zaporedja.

Ob danih podatkih $\mathcal{S} = \{(v_n, t_n, y_n), n = 1, ..., N\}$, kjer $v \in \mathcal{V}$ predstavlja slikovno informacijo, $t \in \mathcal{T}$ tekstovno informacijo, $y \in \mathcal{Y}$ pa razred, želimo ob učenju funkcij $f_v : \mathcal{V} \to \mathcal{Y}$ in $f_t : \mathcal{T} \to \mathcal{Y}$ minimizirati empirično tveganje, ki ga zapišemo kot

$$\frac{1}{N}\sum_{n=1}^{N} \Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n)), \tag{7}$$

kjer je $\Delta$ 0-1 kriterijska funkcija, z $N$ pa smo označili število vseh parov v podatkovni zbirki.

Funkcijo kompatibilnosti slikovne in tekstovne informacije so definirali kot

$$F(v, t) = \theta(v)^T \varphi(t), \tag{8}$$

razvrščevalnika $f_v$ in $f_t$ pa sta definirana kot

$$\begin{aligned} f_v(v) &= \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{t \sim \mathbb{T}(y)} F(v, t) \\ f_t(t) &= \arg\max_{y in \mathcal{Y}} \mathbb{E}_{v \sim \mathbb{V}(y)} F(v, t), \end{aligned} \tag{9}$$

kjer je $\mathcal{T}(y)$ podmnožica tekstovne informacije $\mathcal{T}$ razreda $y$, $\mathcal{V}(y)$ pa podmnožica slikovne informacije $\mathcal{V}$ razreda $y$.

Ker je 0-1 kriterijska funkcija v enačbi 7 nezvezna in nima definiranega odvoda, namesto te optimiziramo pomožno funkcijo, ki je zvezna:

$$\begin{aligned} &\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} l_v(v_n, t_n, y_n) + l_t(v_n, t_n, y_n) \\ &l_v(v_n, t_n, y_n) = \\ &\sum_{y \in \mathcal{Y}} \max(0, \Delta(y_n, y) + \mathbb{E}_{t \sim \mathcal{T}(y)}[F(v_n, t) - F(v_n, t_n)] \\ &l_t(v_n, t_n, y_n) = \\ &\sum_{y \in \mathcal{Y}} \max(0, \Delta(y_n, y) + \mathbb{E}_{t \sim \mathcal{T}(y)}[F(v, t_n) - F(v_n, t_n)] \end{aligned} \tag{10}$$

Z učenjem tako dobimo naučeno nevronsko omrežje $\varphi$, ki je ustrezno za kodiranje jezikovne informacije.

V članku (Reed et al., 2016b) so kodirali tekstovni opis s kodirnikom $\varphi$, nato pa so zakodiran opis s polnopovezano plastjo zmanjšali na 128-dimenzionalni vektor in dodali nelinearno aktivacijo. Ta vektor so zaporedno dodali naključnemu vektorju $z$, vse skupaj pa nato tvori začetni signal za generator $G$.

Diskriminator prav tako potrebuje informacijo o tekstovnemu opisu, saj želimo z diskriminatorjem oceniti, ali se dana slika in njen opis ujemata ali ne. Četudi je generator ustvaril prepričljivo sliko, moramo preveriti, ali ta slika ustreza našemu tekstovnemu opisu. Zato so v tem članku prej omenjeni 128-dimenzionalni vektor prostorsko ponovili in dodali v konvolucijski sloj prostorske dimenzije $4 \times 4$. Celotna arhitektura je prikazana na sliki 3.

Za učenje, ki bo upoštevalo semantično skladnost opisa in slike, moramo spremeniti kriterijsko funkcijo. Namesto enačbe 1 se diskriminator uči na parih podatkov. Označimo z $x$ resnično sliko, s $h$ kodiran tekst, ki semantično pripada trenutni sliki, z $\hat{x}$ sliko, ki jo je ustvaril generator in s $\hat{h}$ kodiran tekst, ki semantično ne ustreza trenutni sliki. Kriterijska funkcija za diskriminator, ki jo ta skuša maksimizirati, se tako glasi

$$\mathcal{L}_D = \log(D(x, h)) + \frac{\log(1 - D(x, \hat{h}) + \log(1 - D(\hat{x}, h))}{2}, \tag{11}$$

kriterijska funkcija za generator, ki jo ta skuša maksimizirati, pa se sedaj glasi

$$\mathcal{L}_G = \log(D(\hat{x}, h). \tag{12}$$

V članku (Zhang et al., 2017), ki je prikazal vizualno prepričljivejše in višje-resolucijske slike, kot (Reed et al., 2016b), so uporabljali isti način kodiranja $\varphi$, vendar skušajo doseči večjo razpršenost tekstovnih kodiranj. Ob upoštevanju, da je količina tekstovnih opisov omejena in da so dimenzije kodiranega tekstovnega opisa običajno visokodimenzionalne (običajno $> 100$ dimenzij), se v prostoru mnogoterosti kodiranega teksta pojavlja nezveznost. Zaradi tega v članku predlagajo tehniko pogojene avgmentacije (angl. Conditioning Augmentation), ki je zelo podobna tehniki kodiranja pri variacijskem avtokodirniku (Kingma in Welling, 2013).

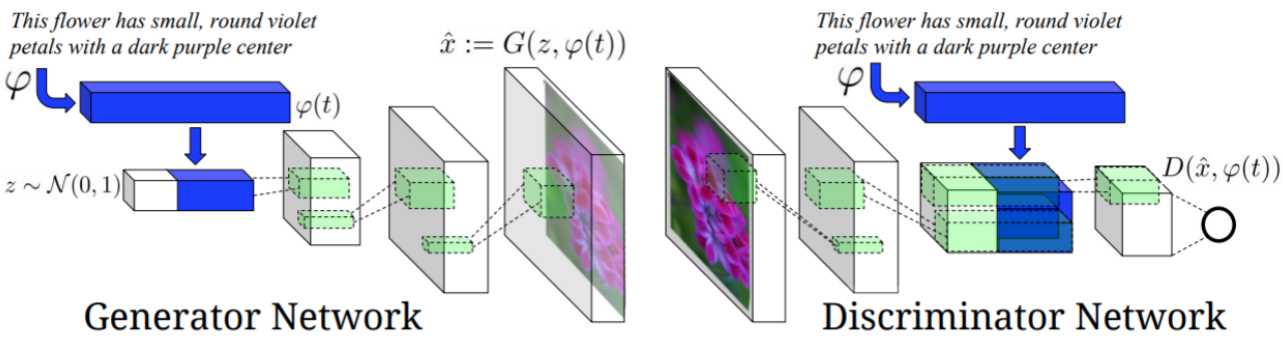
Metoda tekstovno kodiranje $c$ vzorči iz Gaussove porazdelitve $\mathcal{N}(\mu(\varphi(t)), \Sigma(\varphi(t)))$, kjer je kovariančna matrika $\Sigma$ diagonalna. Tako vektor povprečij $\mu$ kot diagonalna kovariančna matrika $\Sigma$ sta funkciji tekstovnega kodiranja $\varphi(t)$. Za preprečevanje prenaučenosti modela h kriterijski funkciji dodajo še izraz za regularizacijo te verjetnostne porazdelitve s Kullback-Leibler divergenco $D_{KL}(\mathcal{N}(\mu(\varphi(t)), \Sigma(\varphi(t))), \mathcal{N}(0, I))$. S to tehniko zagotovimo, da posamezni kodirani tekst zavzame večji prostor na mnogoterosti kodiranih tekstov, obenem pa generativnimi model spodbuja k robustnosti na manjše perturbacije kodiranega teksta. Tehnika pogojne avgmentacije je postala popularna metoda za vse nadaljnje generativne modele, ki so zunanji signal pogojili s tekstovnim opisom.

Poleg tega so v (Zhang et al., 2017) namesto arhitekture enega generativnega modela uporabili dva GAN modela, zložena en na drugega. Prvi je ustvaril nižjeresolucijsko $64 \times 64$ sliko, drugi pa je glede na to sliko ustvaril višjeresolucijsko $256 \times 256$ sliko.

### 4.2. Globok pozornostni multimodalni podobnostni model

Motivacija za modeliranje po črkah namesto modeliranja po besedah je predvsem omogočanje večje robustnosti na tipkarske napake v opisih. Ob zadostni količini lepo označenih podatkov pa imajo prednost modeli, ki kot samostojno enoto modelirajo posamezno besedo. Besede so modelirane kot vektorske zloženke, ki so pogosto prednaučene na nenadzorovan način, kot npr. GloVe (Pennington et al., 2014). Celotni tekstovni opis nato običajno modeliramo s povratno nevronsko mrežo.

Model AttnGAN (Xu et al., 2018) je besede modeliral z učenimi vektorskimi zloženkami. Niz besed je bil opisan z besednimi značilkami $e \in \mathbb{R}^{D \times T}$, kjer je $T$ dolžina niza, $D$ pa dimenzionalnost značilk. Iz niza besed modeliramo




Slika 3: Predlagana arhitektura modela iz članka (Reed et al., 2016b). Kodiran tekst $\varphi(t)$ se zaporedno doda naključnemo vektorju $z$, skupaj pa tvorita začetni signal za naslednje sloje generatorskega omrežja. Kodiran tekst $\varphi(t)$ se prav tako doda diskriminatornem omžreju, s čimer lahko prisilimo generatorsko omrežje k semantični povezanosti tekstovnega opisa in nastale slike. Slika je povzeta po (Reed et al., 2016a).

globalno tekstovno predstavitev $\bar{e}$ z uporabo dvosmernega LSTM modela. Dvosmerni LSTM model je sestavljen iz dveh LSTM komponent, kjer vsaka komponenta modelira niz besed v svoji smeri, prva od začetka povedi proti koncu, druga pa od konca povedi proti začetku. Globalni vektor $\bar{e}$ dobimo z združenjem končnega skritega vektorja $h$ obeh komponent.

Podobno kot pri modelu StackGAN, tudi ta model uporabi več diskriminatorjev, kjer je vsak diskriminator zadolžen za svojo resolucijo. Bistveno se razlikuje v tem, da AttnGAN tekst ne modelira le z globalno tekstovno predstavitev $\hat{e}$, temveč uporabi tudi posamezne besede $e$ za implementacijo mehanizma pozornosti v modelu. Pozornost predstavlja mehanizem, prvič predstavljen v članku (Bahdanau et al., 2014), kjer se model sam skuša naučiti, koliko posamezna vhodna komponenta prispeva h končnemu izhodu.

Tekstovni kodirnik $\varphi$ v nasprotju s prejšnimi modeli učijo po postopku, imenovanem globok pozornostni multimodalni podobnostni model (angl. deep attentional multimodal similarity model). Model maksimizira podobnost slike in posameznih besed v pripadajočih tekstovnih opisih. Označimo z $v \in \mathbb{R}^{D \times M}$ vizualne značilke, pridobljene iz prednaučene konvolucijske mreže, kjer je $D$ njihova dimenzionalnost, $M$ pa število prostornih elementov globokih značilk. Posamezno kodirano besedo označimo z $e_i$. Izračunamo podobnostno matriko

$$s = e^T v, \tag{13}$$

kjer velja $s \in \mathbb{R}^{T \times M}$, element matrike $s_{i,j}$ pa ponazarja podobnost med $i$-to besedo in $j$-to regijo slike. To matriko normaliziramo

$$\bar{s}_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=0}^{T-1} \exp(s_{k,j})}. \tag{14}$$

Vektorsko predstavitev slike glede na $i$-to besedo dobimo z

$$c_i = \sum_{j=0}^{M-1} \alpha_j v_j, \tag{15}$$

kjer je

$$\alpha_j = \frac{\exp(\gamma_1 \bar{s}_{i,j})}{\sum_{k=0}^{M} \exp(\gamma_1 \bar{s}_{i,k})}. \tag{16}$$

S faktorjem $\gamma_1$ uravnavamo prispevke drugih regij na vektorsko predstavitev posamezne regije.

Ustreznost med $i$-to besedo $e_i$ in predstavitvijo celotne slike $c_i$ definiramo z mero kosinusne podobnosti. Funkcijo ustreznosti definiramo kot

$$U(c_i, e_i) = \frac{c_i^T e_i}{||c_i|| ||e_i||}, \tag{17}$$

ustreznost med celotno sliko $Q$ in celotnim tekstovnim opisom pa kot

$$U(Q, D) = \log(\sum_{i=0}^{T-1} \exp(\gamma_2 U(c_i, e_i)))^{\frac{1}{\gamma_2}}, \tag{18}$$

kjer je $\gamma_2$ faktor, s katerim uravnavamo prispevke vektorskih predstavitev drugih besed.

Funkcija $U(Q, D)$ nam torej vrne podobnost med tekstovnim opisom in besedo. Edino, kar nam še preostane, je definiranje kriterijske funkcije za učenje. Učenje definiramo tako, da povečujemo logaritmsko verjetnost ustreznih parov opisov in slik. Konkretno lahko za sveženj parov $\{(Q_i, D_i)\}_{i=1}^{M}$ zapišemo, da je verjetnost ustreznosti opisa $D_i$ in slike $Q_i$

$$P(D_i|Q_i) = \frac{\exp(\gamma_3 U(Q_i, D_i))}{\sum_{j=1}^{M} \exp(\gamma_3 U(Q_i, D_i))}. \tag{19}$$

V danem svežnju označimo, da si $Q_i$ in $D_i$ ustrezata, vsi ostali opisi pa ne ustrezajo. Kriterijska funkcija, ki jo minimiziramo, je definirana kot

$$L_1 = -\sum_{i=1}^{M} \log P(D_i|Q_i), \tag{20}$$

simetrično pa minimiziramo tudi

$$L_2 = -\sum_{i=1}^{M} \log P(Q_i|D_i). \tag{21}$$

S to optimizacijo si zagotovimo primerne vloženke $e_i$, ki se uporabljajo v sklopu mehanizma pozornosti v konvolucijskih slojih modela GAN. Prav tako si želimo še kvalitetne globalne predstavitve celotnega opisa, za kar poskrbi model LSTM. Globalna predstavitev tekstovnega opisa $\bar{e}$ in globalna predstavitev celotne slike kot povprečenje aktivacij konvolucijskih filtrov na zadnjem sloju $\bar{v}$ se uporabita kot funkciji enačbe 17, nato pa postopek sledi vsem nadaljnjim enačbam. S tem dobimo še globalni doprinos h kriterijski funkciji.

### 4.3. Ostali modeli

Trenutno najboljše rezultate daje model MirrorGAN (Qiao et al., 2019), kjer skušajo modelu zagotoviti konsistentnost originalnih tekstovnih opisov in tekstovnih opisov ustvarjenih slik. Postopek se zgleduje po konceptu ciklične konsistence modela CycleGAN (Zhu et al., 2017). Konsistentnost tekstovnih opisov zagotovijo z predučenjem modela, ki glede na sliko ustvari ustrezen tekstovni opis. Celotni sistem učenja generativnega modela nato temelji na tem, da mora slika generatorja, pogojena na določenem tekstu, dobiti isti tekst kot rezultat novo predstavljenega modela ustvarjanja tekstovnega opisa. Ostali deli arhitekture se močno zgledujejo po modelu AttnGAN.

Opazimo, da vsi novejši modeli sliko gradijo postopno, torej z uporabo več generatorjev in diskriminatorjev za posamezne resolucije. V nasprotju z njimi pa v članku (Souza et al., 2020) model učijo direktno na najvišji resoluciji $256 \times 256$ brez uporabe vmesnih generatorjev in diskriminatorjev. To storijo z modifikacijo modela BigGAN (Brock et al., 2019), ki ga posodobijo na tak način, da namesto informacije o razredu izhodnega modela dobi kodiran tekstovni opis $\varphi(t)$.

### 4.4. Uporaba slovenščine

Domnevamo, da je zaradi odsotnosti podatkovnih zbirk in večje narave pregibnosti besed slovenščina problematičen jezik za učinkovito tekstovno pogojeno ustvarjanje slik. Mogoča rešitev se skriva v skrbnem predprocesiranjem podatkov ali uporabi predhodno naučenih besednih zloženk s postopki, kot je npr. GloVe (Pennington et al., 2014) in FastText (Bojanowski et al., 2017). Slednji tudi ponuja prednaučen model za slovenščino (Grave et al., 2018). Ti postopki temeljijo na nenadzorovanem učenju na velikanskih bazah podatkov (npr. Wikipedija nekega jezika), s čimer pridobijo besedne vloženke, ki besede s podobnim pomenom preslika bližje v višjedimenzionalnem prostoru. To lahko s pridom izkoristijo različni modeli za kodiranje tekstovnega opisa (Collobert et al., 2011).

### 4.5. Rezultati

V tabeli 1 so prikazani rezultati v obliki metrike Inception Score za nekaj najboljših modelov na področju tekstovno pogojenega ustvarjanja slik. Rezultati so ovrednoteni na validacijski množici podatkovnih zbirk, kjer ustvarimo 30.000 slik glede na tekstovni opis. Glede na ustvarjene slike izračunamo metriko Inception Score. Rezultati so prikazani za podatkovni zbirki Caltech-UCSD Birds (CUB) (Welinder et al., 2010) in Microsoft COCO (Lin et al., 2014). Ovrednoteni modeli so GAN-INT-CLS (Reed et al., 2016b), StackGAN (Zhang et al., 2017), AttnGAN (Xu et al., 2018) in MirrorGAN (Qiao et al., 2019).

Tabela 1: Inception Score posameznih modelov. Višji rezultat pomeni kvalitetnejši model.

| Model | CUB | COCO |
|---|---|---|
| GAN-INT-CLS | 2.88 | 7.88 |
| StackGAN | 3.70 | 8.45 |
| AttnGAN | 4.36 | 25.89 |
| MirrorGAN | 4.56 | 26.47 |

Na sliki 4 pa so prikazani primeri rezultatov posameznih modelov, ko izhod modela pogojimo z nekim tekstovnim opisom.

Slika 4: Rezultati posameznih modelov GAN na podatkovni zbirki Caltech-UCSD Birds. Od zgoraj navzdol: GAN-INT-CLS (Reed et al., 2016b), StackGAN (Zhang et al., 2017), AttnGAN (Xu et al., 2018) in MirrorGAN (Qiao et al., 2019).

### 4.6. Možnosti nadgradnje

Potencialne možnosti nadgradnje vidimo predvsem v naslednjih izboljšavah:

- **Tekstovni model**
  Trenutni tekstovni modeli delujejo na principu nenadzorovanega učenja za učenje značilk posameznih be-

sed v kombinaciji s povratno nevronsko mrežo za modeliranje celotnega stavka. Novejši tekstovni modeli (Devlin et al., 2018) delujejo brez povratnega mehanizma, temevač na podlagi mehanizma pozornosti z nenadzorovanem učenjem na velikanskih podatkovnih bazah pridobijo značilke, ki ustrezajo celotni povedi. Z doučevanjem na posameznih podatkovnih zbirkah so ti modeli na klasičnih podatkovnih zbirkah za vrednotenje razumevanje teksta presegli rezultate, ki so jih dosegale povratne nevronske mreže. Z ustreznim načinom doučevanja domnevamo, da lahko izkoristimo predznanje, zajeto v prednaučenem tekstovnem modelu, s tem pa pridobimo boljšo pogojenost glede na tekstovni opis.

- **Bogatenje podatkov**
  Podatkovne zbirke, opisane v članku, so zaradi zamudnega zbiranja podatkov manj obsežne kot podatkovne zbirke, namenjene drugim izključno eni modaliteti. Za generativna nasprotniška omrežja je znano, da boljše delujejo večanju količine podatkov. Nedavni članek (Zhao et al., 2020) pa je pokazal, da je mogoče kvalitetne generativne modele naučiti z veliko manj podatki ob uporabi odvodljivih transformacij za bogatenje podatkov.

- **Način učenja**
  V članku (Brock et al., 2019) so pokazali, da generativni modeli dosežejo boljše rezultati ob večanju velikosti svežnja (angl. batch size). Slabost te tehnike pa je velika računska potratnost.

## 5. Zaključek

Tematika tekstovno pogojenega ustvarjanja slik je aktivna in privlačna veja raziskovanja, ki v zadnjih letih dobiva velik zagon. V članku smo opisali metode, podatkovne zbirke in trenutne rezultate na področju tekstovno pogojenega generiranja slik z generativnimi nasprotniškimi omrežji. Posvetili smo se arhitekturi modelov za kodiranje tekstovnega opisa, načinom kodiranja tekstovnega opisa in ustreznim kriterijskim funkcijam. Na koncu smo predstavili še možnosti uporabe slovenščine in nekaj obetajočih novejših metod.

## 6. Zahvala

Predstavljeno delo je delno financirano s strani ARRS v okviru raziskovalnega programa P2-0250 Metrologija in biometrični sistemi ter aplikativnega raziskovalnega projekta L7-9406 OptiLEX.

## 7. Literatura

Martin Arjovsky, Soumith Chintala in Léon Bottou. 2017. Wasserstein generative adversarial networks. V: *International Conference on Machine Learning*, str. 214–223.

Dzmitry Bahdanau, Kyunghyun Cho in Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Piotr Bojanowski, Edouard Grave, Armand Joulin in Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

A. Brock, J. Donahue in K. Simonyan. 2019. Large scale GAN training for high fidelity natural image synthesis. V: *ICLR*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho in Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu in Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li in L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. V: *CVPR09*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee in Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio in Graham W Taylor. 2019. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. V: *Proceedings of the IEEE International Conference on Computer Vision*, str. 10304–10312.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville in Yoshua Bengio. 2014. Generative adversarial nets. V: *Advances in neural information processing systems*, str. 2672–2680.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin in Tomas Mikolov. 2018. Learning word vectors for 157 languages. V: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin in Aaron C Courville. 2017. Improved training of wasserstein gans. V: *Advances in neural information processing systems*, str. 5767–5777.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler in Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. V: *Advances in neural information processing systems*, str. 6626–6637.

Sepp Hochreiter in Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou in Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. V: *Proceedings of the IEEE conference on computer vision and pattern recognition*, str. 1125–1134.

T. Karras, T. Aila, S. Laine in J. Lehtinen. 2018. Progressive growing of GANs for improved quality, stability, and variation. V: *ICLR*.

Tero Karras, Samuli Laine in Timo Aila. 2019a. A style-based generator architecture for generative adversarial networks. V: *Proceedings of the IEEE Conference on*

*Computer Vision and Pattern Recognition*, str. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen in Timo Aila. 2019b. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*.

Diederik P Kingma in Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár in C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. V: *European conference on computer vision*, str. 740–755. Springer.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang in Stephen Paul Smolley. 2017. Least squares generative adversarial networks. V: *Proceedings of the IEEE International Conference on Computer Vision*, str. 2794–2802.

Maria-Elena Nilsback in Andrew Zisserman. 2008. Automated flower classification over a large number of classes. V: *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec.

Sebastian Nowozin, Botond Cseke in Ryota Tomioka. 2016. f-gan: Training generative neural samplers using variational divergence minimization. V: *Advances in neural information processing systems*, str. 271–279.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang in Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. V: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, str. 2337–2346.

Jeffrey Pennington, Richard Socher in Christopher D Manning. 2014. Glove: Global vectors for word representation. V: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, str. 1532–1543.

Tingting Qiao, Jing Zhang, Duanqing Xu in Dacheng Tao. 2019. Mirrorgan: Learning text-to-image generation by redescription. V: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, str. 1505–1514.

A. Radford, L. Metz in S. Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Scott Reed, Zeynep Akata, Honglak Lee in Bernt Schiele. 2016a. Learning deep representations of fine-grained visual descriptions. V: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, str. 49–58.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele in Honglak Lee. 2016b. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford in Xi Chen. 2016. Improved techniques for training gans. V: *Advances in neural information processing systems*, str. 2234–2242.

Douglas M Souza, Jônatas Wehrmann in Duncan D Ruiz. 2020. Efficient neural architecture for text-to-image synthesis. *arXiv preprint arXiv:2004.11437*.

Ilya Sutskever, James Martens in Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. V: *Proceedings of the 28th international conference on machine learning (ICML-11)*, str. 1017–1024.

Ilya Sutskever, Oriol Vinyals in Quoc V Le. 2014. Sequence to sequence learning with neural networks. V: *Advances in neural information processing systems*, str. 3104–3112.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens in Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. V: *Proceedings of the IEEE conference on computer vision and pattern recognition*, str. 2818–2826.

Kai Sheng Tai, Richard Socher in Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie in P. Perona. 2010. Caltech-UCSD Birds 200. Tehnično poročilo CNS-TR-2010-001, California Institute of Technology.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang in Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. V: *Proceedings of the IEEE conference on computer vision and pattern recognition*, str. 1316–1324.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang in Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. V: *Proceedings of the IEEE international conference on computer vision*, str. 5907–5915.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu in Song Han. 2020. Differentiable augmentation for data-efficient gan training. *arXiv preprint arXiv:2006.10738*.

Jun-Yan Zhu, Taesung Park, Phillip Isola in Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. V: *Proceedings of the IEEE international conference on computer vision*, str. 2223–2232.

# Digitalizacija kulturne dediščine v dveh projektih programa Študentski inovativni projekti za družbeno korist

**Katja Mihurko Poniž,* Narvika Bovcon,‡ Aneta Trajanov,†¥ Tina Smrekar⁞**

* Fakulteta za humanistiko, Raziskovalni center za humanistiko; Univerza v Novi Gorici
Vipavska 13, 5000 Nova Gorica
katja.mihurko.poniz@ung.si
‡ Fakulteta za računalništvo in informatiko, Univerza v Ljubljani,
Večna pot 113, 1000 Ljubljana
narvika.bovcon@fri.uni-lj.si
† Poslovno-tehniška fakulteta, Univerza v Novi Gorici
Glavni trg 8, 5271 Vipava
¥Odsek za tehnologije znanja, Institut Jozef Stefan
Jamova cesta 39, 1000 Ljubljana
aneta.trajanov@ijs.si
⁞ Akademija umetnosti, Univerza v Novi Gorici
Vipavska 13, 5000 Nova Gorica
tina.smrekar@ung.si

**Povzetek**

Članek predstavi dva projekta digitalizacije in spletne predstavitve kulturne dediščine aleksandrink in različnih skupin žensk, ki so večinsko povezane s primorskim in notranjskim geografskim prostorom: v prvem primeru gre za bazo podatkov o aleksandrinkah, v drugem pa za digitalizirano pisemsko dediščino. Program Študentskih inovativnih projektov za družbeno korist je odlična platforma za pilotske projekte s področja digitalne humanistike, saj predvideva interdisciplinarno delovno skupino študentov in mentorjev ter v korist lokalne skupnosti usmerjene praktične rezultate projekta.

**Two projects of digitizing the cultural heritage from the Student innovative projects for social benefit initiative**

The paper presents two projects of digitizing the cultural heritage of alexandrinians and other groups of women from the Primorska and Notranjska regions. In the first project we have built an on-line data base with places and dates related to the life and travel of alexandrinians to Egipt, whereas in the second project we have prepared a web site with women's transcribed letters. Digital humanities projects can be successfully realized in the frameworks of the Student innovative projects for social benefit initiative: in an interdisciplinary team of students and mentors, with results that are relevant in the local community.

## 1. Uvod

V okolju, v katerem deluje Univerza v Novi Gorici, se nahaja bogata kulturna dediščina, povezana z zgodovino dejavnosti, vlog in razumevanja žensk, ki v veliki meri še ni bila obdelana, raziskana in ovrednotena. Zgodovina njihovega vsakdanjega življenja je bila dolgo zapostavljeno področje raziskovanja, zato so številni zgodovinski viri še neobdelani. Hkrati zahteva njihova obdelava t. i. množično zunanje izvajanje (*crowdsourcing*), ki v slovenskem prostoru še ni razvilo in uveljavilo vseh možnosti. Študentje lahko pomembno prispevajo k digitalizaciji in obdelavi tovrstnega gradiva, saj delno že imajo kompetence s področja obdelave pisnih virov, v tovrstnih projektih pa pridobijo še kompetence s področja digitalne humanistike (DH), ki so trenutno v študijskih programih prisotne v manjši meri ali pa sploh ne. Kot izjemo navedimo programe na Univerzi v Novi Gorici, kjer je ta predmet del študijev na vseh treh stopnjah, obstaja pa tudi samostojen interdisciplinarni magistrski študij Digitalna humanistika.[1]

Problem, ki ga oba predstavljena projekta – digitalizaciji kulturne dediščine aleksandrink ter pisemske dediščine Primork in Notranjk – razrešujeta, je, kako s sodobnimi pristopi s področja DH obdelati zgodovinske vire, da postanejo bolj dostopni za zainteresirano javnost in nadaljnje raziskave, kako se jim približati skozi kreativne in umetniške digitalne prakse ter kako jih posredovati v učnem procesu. S tem študente usposabljamo za sodelovanje v interdisciplinarni skupini, ki združuje področja tehnike, humanistike, družboslovja, vizualnih komunikacij in umetnosti.

## 2. Digitalna humanistika v izobraževanju

V mednarodni DH skupnosti so izkušnje izvajanja vsebin s področja DH v študijskih programih spodbudile raziskave o DH v izobraževanju. Prvi premisleki o poučevanju računalništva v povezavi s humanistiko so se pojavili že v osemdesetih in devetdesetih letih 20. stoletja (Hirsch, 2012). Z razvojem raziskovalnih centrov, v katerih so se izoblikovale ideje za pedagoške programe, so vsebine s področja DH vse bolj prodirale v akademsko okolje. V zadnjem desetletju se je odvilo nekaj odmevnih delavnic in prispevkov na temo pedagogike v DH: npr. *Teaching Literature and Language Online* (2009), *THATCamp Pedagogy* (2011), *Learning through Digital Media* (2011) (Hirsch, 2012). Najbolj odmevno delo s tega področja je verjetno zbornik *Digital Humanities*

[1] http://ung.si/dh

*Pedagogy* (2012), ki je prvi večji projekt, ki premišljuje DH v povezavi s konkretnimi pristopi poučevanja. V okviru panevropske mreže DARIAH.EU deluje delovna skupina #dariahTeach, ki zagotavlja in spodbuja ustvarjanje in razširjanje visokokakovostnih učnih gradiv za raziskovalce in študente, ki jih zanimajo digitalna humanistična orodja, metode in pristopi. Delovna skupina se osredotoča zlasti na: trajnost in nadaljnji razvoj platforme #dariahTeach, razvoj odprtih spletnih učnih gradiv, razvoj novih modelov za ocenjevanje in objavljanje gradiv za poučevanje in usposabljanje, ter na dejavnosti, ki spodbujajo boljše razumevanje poučevanja DH prek disciplinskih, institucionalnih, jezikovnih in kulturnih meja (#dariahTeach, 2020).

Med odmevnejšimi zapisi o DH v izobraževanju je članek Ryana Cordella *How Not to Teach Digital Humanities* (2016). Avtor v njem opozarja na prepad med dojemanjem samega pojma "digitalno" med profesorji in študenti: medtem ko se prvim zdi povezava med "digitalnim" in "humanistiko" nekaj revolucionarnega, študenti samo besedo digitalno dojemajo kot tako rekoč nevtralno. A medtem ko jim je vsakdanja uporaba raznih digitalnih orodij povsem domača, o samih orodjih ne vedo veliko in so nemalokrat zadržani do DH. Mnogi izmed njih želijo študirati humanistiko kot nasprotje tehnološkemu poudarku v ustroju družbe, v kateri živijo. Temeljno vprašanje je zato, kako nevsiljivo podati študentom vsebine, ki so jim blizu, a na nov način, ki ni že *a priori* odbijajoč. Cordell opozarja, da predmeti, pri katerih naj bi na poglobljen način razpravljali o pojmu DH na splošno, študentov običajno ne pritegnejo, nasprotno pa radi delajo projekte, v katerih razvijajo veščine, ki sodijo na področje DH, zato Cordell zagovarja vključitev DH v dodiplomske kurikule kot "naravni" del tega, kar literarni zgodovinarji počnejo: npr. veščino oddaljenega branja lahko poučujemo ob veščini natančnega branja. Pomembno je tudi lokalno osredotočeno razmišljanje, izvajanje in ustvarjanje DH, ki ni osmišljena samo na univerzalni ravni, temveč v akademskem okolju ustanove in regije, kjer jo razvijamo. Smiselno jo je povezati s poslanstvom ustanove in z lokalni knjižnicami, muzeji, raziskovalnimi centri ipd. Študentom je všeč, če lahko prispevajo s svojimi projekti k javnemu dobremu – s tem nenazadnje prestopajo meje učilnice, kar jim omogoča tudi delo s knjižničarji in drugimi kolegi, s katerimi bi se sicer med študijem le bežno srečali (Cordell, 2016).

S podobnimi izzivi smo se srečali tudi pri obeh projektih, ki jih predstavljamo, zato menimo, da naše pozitivne izkušnje pri posredovanju pristopov s področja DH študentski projektni skupini lahko prispevajo k intenzivnejšemu vključevanju DH vsebin v visokošolske kurikulume. Odločitev, da v projektih ustvarimo elektronski zbirki, je izšla tudi iz prepričanja, da so raziskave, ki so povezane z zgodovino žensk in še posebno z njihovimi korespondencami, vitalen del DH (Prim.: *Correspondance d'Isabelle de Charrière*,[2] *NEWW Women Writers*,[3] *Književnost – Theory and History of Women's Writing in Serbian until 1915*,[4] *WeChangEd Database*,[5] *Reading Experience Database*,[6] *Discovering American Women's History Online*[7]). Študentke in študentje so se z elektronskimi zbirkami, ki se navezujejo na tematike iz zgodovine žensk, seznanili v prvi fazi projekta in tako pridobili nova znanja s področja DH.

## 3. Prednosti lokalnih projektov in iniciativa ŠIPK

Prednosti lokalno zastavljenih projektov s področja digitalne humanistike je v slovenskem prostoru mogoče uresničevati s finančno podporo programa Študentski inovativni projekti za družbeno korist (ŠIPK, Javni štipendijski, razvojni, invalidski in preživninski sklad Republike Slovenije), ki povezuje visokošolske zavode z negospodarskimi organizacijami, predvsem društvi in zavodi (knjižnicami, muzeji, galerijami …). V projektih sodelujejo študentje, visokošolski učitelji in strokovnjaki na projektnih področjih. Tako se razvijajo dobre prakse in inovativne oblike učenja za razvoj kompetenc in praktičnih izkušenj. Projekti se izvajajo v skupinah od 6 do 10 študentov pod vodstvom pedagoškega mentorja in strokovnega sodelavca iz lokalnega okolja in lahko trajajo od 3 do največ 5 mesecev. V okviru izbranih projektov študentje proučujejo različne inovativne rešitve za izzive negospodarskega in neprofitnega sektorja.

V članku želimo predstaviti dobro prakso izvajanja dveh študentskih projektov s področja DH, ki smo jih izvedli pod okriljem omenjenega programa na Univerzi v Novi Gorici v sodelovanju z različnimi akademskimi in neakademskimi partnerji spomladi leta 2018 in spomladi leta 2020.

## 4. Projekt Digitalizacija kulturne dediščine aleksandrink (2018)

### 4.1 Izvedba projekta

V projektu smo zbirali podatke o aleksandrinkah, ekonomskih migrantkah, ki so odhajale od konca 19. stoletja do druge svetovne vojne pretežno iz spodnje Vipavske doline in tudi drugih krajev na delo v Egipt. Tam so večinoma opravljale delo gospodinjskih pomočnic, služabnic, varušk, kuharic, sobaric, spremljevalk bogatih dam in dojilj. Podatke smo zbrali iz strokovne literature, televizijskih oddaj, filmov in iz dokumentov, ki jih je zbralo Društvo za ohranjanje kulturne dediščine aleksandrink Prvačina med obiskom v Egiptu (fotografije grobov alekandrink) ter v arhivih na Dunaju in v Rimu. V teh dokumentih se nahajajo podatki o aleksandrinkah (odkod so odšle, kraj službovanja ipd.), ki jih je bilo najprej potrebno digitalizirati in urediti. Kot najbolj sistematično možnost urejanja smo prepoznali digitalno bazo podatkov, ki smo jo v okviru projekta postavili na spletni strežnik in vanjo vpisali podatke o 1788 aleksandrinkah. Fotografirali smo predmete iz muzejske zbirke društva, ki so bili v lasti aleksandrink, ter digitalizirali stare fotografije, ki jih hrani društvo. Tako

[2] https://charriere.huygens.knaw.nl
[3] http://resources.huygens.knaw.nl/womenwriters
[4] http://knjizenstvo.etf.bg.ac.rs/en
[5] https://www.wechanged.ugent.be/wechanged-database
[6] https://www.open.ac.uk/Arts/reading/UK
[7] http://digital.mtsu.edu/cdm/landingpage/collection/women

smo ustvarili dve elektronski zbirki: zbirko podatkov o aleksandrinkah in zbirko fotografij.

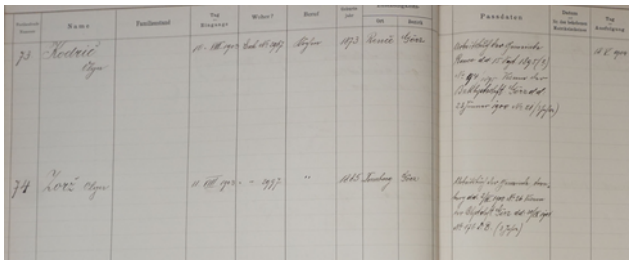
Digitalna baza je za javnost dostopna na spletnem naslovu http://green.fri.uni-lj.si:1869. Iz baze podatkov o aleksadrinkah je zaradi varovanja osebnih podatkov javnosti vidno le ime (prvi stolpec v bazi), druge podatke društvo posreduje, če je izkazan raziskovalni namen. Baza fotografij je v celoti dostopna. Podatkovna baza omogoča sistematično zbiranje in urejanje informacij o aleksandrinkah tudi po izteku projekta.

Na vstopni strani spletišča z bazo podatkov o aleksandrinkah je na kratko predstavljen pojem aleksandrinstva; k podrobnejši zgodovinski predstavitvi problema aleksandrinstva vodi gumb Zgodovina, ki uporabnika poveže s spletno stranjo Društva za raziskovanje kulturne dediščine aleksandrink, medtem ko do baze podatkov o aleksandrinkah, ki je osrednji del spletišča, vodi gumb Ogled arhiva. Na vstopni in na vseh podstraneh najdemo ob desnem zgornjem robu zaslona navigacijsko vrstico s tremi povezavami: do baze podatkov o aleksandrinkah (povezava Baza imen, deluje enako kot gumb Ogled arhiva, ki pa je viden samo na vstopni strani), do baze fotografij (povezava Baza slik) in do podrobnejših informacij o projektu[8] (O projektu).

Baza podatkov je predstavljena z razpredelnico, ki je oblikovana tako, da se prilagaja širini računalniškega zaslona. Ker je kategorij v razpredelnici veliko, je za ogled primeren večji zaslon. Spletišče je namenjeno raziskovanju s pregledovanjem velike zbirke podatkov, ki praviloma poteka na namiznem računalniku, zato prilagoditve prikaza na manjši zaslon pametnega telefona nismo posebej razvijali.

Kategorije, ki urejajo bazo zapisov, smo določili na podlagi zbranih podatkov iz različnih virov, ki niso imeli niti enotne oblike niti enotne vsebine. Podatki o aleksandrinkah (imena in priimki, poklici, ki so jih opravljale, datumi rojstev in smrti, dan prihoda v Egipt in odhoda, kraj prihoda, cilj potovanja, sorodniki, delodajalci) so bili pridobljeni iz naslednjih virov: različni članki o aleksandrinkah, knjige o aleksandrinkah (Peter Zorn: *Aleksandrinke so prišle domov*, Dorica Makuc: *Aleksandrinke*, Daša Koprivec: *Aleksandrinke*, *Otroci aleksandrink*, Mirjam Milharčič Hladnik: *Krila migracij*), filmi in dokumentarni filmi (*Pisma iz Egipta*, *Aleksandrinke*, *Aleksandrija, ki odhaja*), knjige prihodov in odhodov avstro-ogrskih državljanov v Egipt (nem. Passprotokole – v kategoriji Vir v preglednici so označene s kratico PP, letnice 1884–1895, 1896–1903, 1903–1905), mrliške knjige (označene so s kratico MK, datum in kraj pokopa v Egiptu) in nagrobniki na pokopališčih v Kairu. Fotografije knjig prihodov in odhodov ter mrliških knjig nam je posredovalo Društvo za ohranjanje kulturne dediščine aleksandrink.

[8] Projektna skupina: Vodja projekta: prof. dr. Katja Mihurko Poniž. Študentke in študentje: Vesna Bajc, Mija Horvat, Nika Kravos (Fakulteta za humanistiko, UNG), Saša Buti (Fakulteta za humanistične študije, UP), Manca Koren (Teološka fakulteta, UL), Blaž Štolar, Tiziano Biagi (Akademija umetnosti, UNG), Gregor Kerševan, Matej Kočevar (Fakulteta za računalništvo in informatiko, UL). Pedagoške mentorice: prof. dr. Katja Mihurko Poniž, izr. prof. dr. Narvika Bovcon, doc. mag. Tina Smrekar. Strokovna sodelavka in sodelavec: mag. Darinka Kozinc (Društvo za ohranjanje kulturne dediščine aleksandrink) in Rok Bavčar (Goriški muzej).



Slika 1. Primer vira.

Pri prepisovanju podatkov iz fotografiranih arhivskih dokumentov smo se srečali z dvema izzivoma: prvič, sistematično smo morali določiti kategorije za podatke, ki so bili v različnih virih zapisani na različne načine, in drugič, pravilno smo morali prepisati pogosto nečitljive, ročno in s starinsko pisavo napisane dokumente, ki so bili ponekod slabo ohranjeni in pogosto zapisani v tujem jeziku (običajno v nemškem ali italijanskem jeziku, kar je najbolj razvidno pri opredelitvi poklica aleksandrinke).

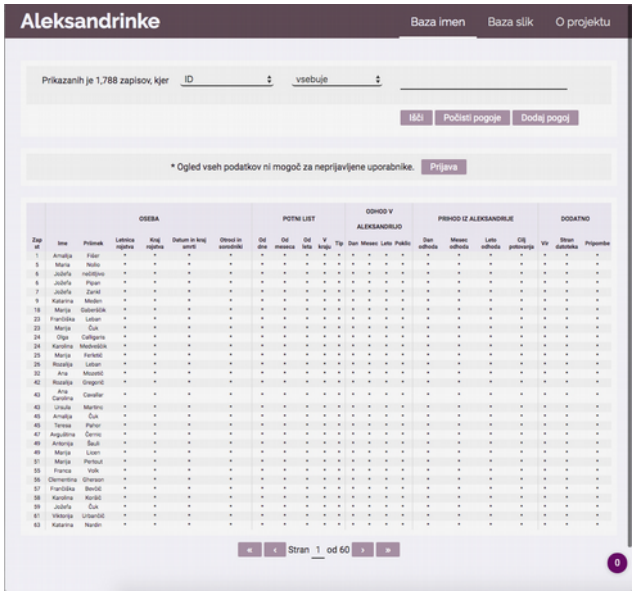
Urejanja in prepisovanja velikega števila dokumentov so se lotile študentke humanistike, ki so si dokumente iz več arhivov razdelile in hkrati vpisovale podatke v skupno preglednico v spletni storitvi Google Docs. S pomočjo skupne preglednice smo lažje nadzorovali in koordinirali delitev dela, saj so študentke že med vpisovanjem lahko prepoznale in preverile ponavljanje vnosov, ki so jih našle v različnih virih. Skupna preglednica je omogočila identifikacijo vseh kategorij podatkov iz različnih virov, na podlagi česar smo na sestanku študentov s pedagoškimi mentorji in strokovnimi sodelavci izluščili ključne kategorije, v katere smo v naslednjem koraku prerazporedili in ponekod združili podatke. To razpredelnico sta študenta računalništva pretvorila v končno datoteko Excel in nato v bazo MySQL. Študentke so vnose v nadaljevanju projekta urejale neposredno v urednikovem pogledu spletne strani in s tem pridobile tudi kompetence uporabe spletnega urejevalnika v sistemu za upravljanje z vsebinami WordPress.[9]

Študenta računalništva in informatike sta pri projektu v praksi uporabila in nadgradila svoje znanje spletnih tehnologij, postavila sta spletni strežnik na fakultetnem računalniku s pomočjo tehnologije Docker[10] ter smiselno in oblikovalsko pregledno prilagodila temo iz programa WordPress za namene spletne predstavitve našega projekta (določila sta elemente vmesnika spletne strani, primerno razporeditev vsebine in navigacijo, izbrala ustrezne črkovne vrste in barvno paleto, uredila velikostne odnose med vsemi elementi za optimalen prikaz, odzivala sta se na pripombe in pobude sodelavcev humanistov kot uporabnikov podatkovne baze in strokovnjakov s področja projekta, nazadnje sta pripravila tudi omejeni pogled na podatke in dostop prek gesla za raziskovalce).

Ob ogledu prikaza razpredelnice Baza imen na spletni strani in ob testiranju uporabniške izkušnje smo opazili pomembno pomanjkljivost pri poimenovanju posameznih kategorij, ki so morale biti za uporabnika čim bolj jasne in razumljive, hkrati pa je omejena širina zaslona – in preglednost – narekovala zelo kratko poimenovanje vsake kategorije. V podatkih namreč nastopa mnogo datumov in krajev, ki jih je bilo treba ustrezno in jedrnato

[9] http://wordpress.org
[10] http://docker.com

poimenovati: rojstva, smrti, izdaje potnega lista, odhoda in vrnitve. Problem smo rešili z združevanjem kategorij v nadkategorije: datume in kraje smo v preglednici razporedili v smiselno zaporedje in jih postavili v soseščino povezanih kategorij, saj se je sicer uporabnik lahko hitro zmedel med pregledovanjem obsežne razpredelnice.




Slika 2. Baza imen.

Baza omogoča iskanje in filtriranje podatkov po vseh kategorijah. S pomočjo iskalnika nad razpredelnico najprej izberemo kategorijo, po kateri bomo iskali, potem določimo tip iskanja (npr. da/ne vsebuje, je/ni enako, je/ni večje/manjše, je/ni eden izmed ...) ter nazadnje, kaj iščemo oz. s čim primerjamo. Določimo lahko več iskalnih kriterijev hkrati.

S pomočjo logičnih ukazov filtra iskanja po kategoriji Cilj potovanja v bazi podatkov lahko poiščemo aleksandrinke: 1) ki so se vrnile v domači kraj; 2) ki so ostale v Egiptu; 3) ki so se iz Egipta vrnile v drug kraj. Možnosti iz ene kategorije lahko prikažemo v povezavi z možnostmi iz drugih kategorij, ko nastavimo dva pogoja s pomočjo filtrov – npr. Cilj potovanja prikažemo v povezavi z izbranim domačim krajem, letom vrnitve, poklicem, v odvisnosti od otrok in sorodnikov itd. –, in na podlagi rezultatov posamezne primerjave ponudimo razlage o povezavah kategorij. Na ta način je mogoče spoznati nekatere značilnosti življenja aleksandrink v dokumentiranem zgodovinskem obdobju, ki v posameznih virih morda niso očitne, v okviru večje baze podatkov pa postanejo razvidne na ravni družbenega fenomena. S pomočjo vizualizacije povezovanja kategorij iz baze podatkov se razkrijejo vzorci v podatkih, ki so lahko osnova za nadaljnje interpretacije, saj predstavljajo utemeljene spodbude in smernice za iskanje in preverjanje družbeno-zgodovinskih vzrokov za izpričana dogajanja.

Datume smo v bazi zapisali z ločenimi kategorijami za dan, mesec in leto in s tem omogočili iskanje ne samo po času (letu) odhoda, ampak tudi po dnevih in mesecih, torej v katerih mesecih v letu so aleksandrinke najpogosteje odhajale (poletnih, zimskih, kot posledica drugih kontekstualnih družbeno-političnih dogodkov …), kdaj v mesecu (npr. na začetku, ob katerih dnevih, saj se odhodi ladij niso dogajali vsak dan).

Podatke o aleksandrinkah v spletni bazi lahko pretvorimo v CSV datoteko in na podlagi nje izdelamo vizualizacije razporeditev po kategorijah, kar načrtujemo v nadaljevanju projekta. Podatke bomo na ta način lahko pregledali in razporedili po npr.: 1) starosti, ko so odšle v Egipt; 2) starosti, ko so odšle iz Egipta; 3) kraju odhoda; 4) kraju rojstva; 5) kraju smrti; 6) po otrocih in sorodnikih; 7) po poklicu. Naštete kategorije lahko povežemo med seboj in ugotovimo, npr. katera je bila običajna starost pri odhodu v Egipt v določenem kraju, ali so odhajale pretežno ženske, ki so že imele otroke, ali ne, v katerem obdobju so odhajale za določeno delo, iz katerih krajev so prihajale izobražene ženske po poklicu učiteljice in iz katerih krajev so prihajale dobre kuharice ali šivilje.[11]

V pogledu Baza slik si lahko ogledamo digitalizirane fotografije materialne kulturne dediščine aleksandrink, tj. dokumentov, predmetov in historičnih fotografij oseb, ki jih hrani društvo. Spletna razstava gradiva deloma rešuje problem oddaljene dostopnosti in tudi pomanjkanja prostora za predstavitev muzejske zbirke v prostorih Društva.

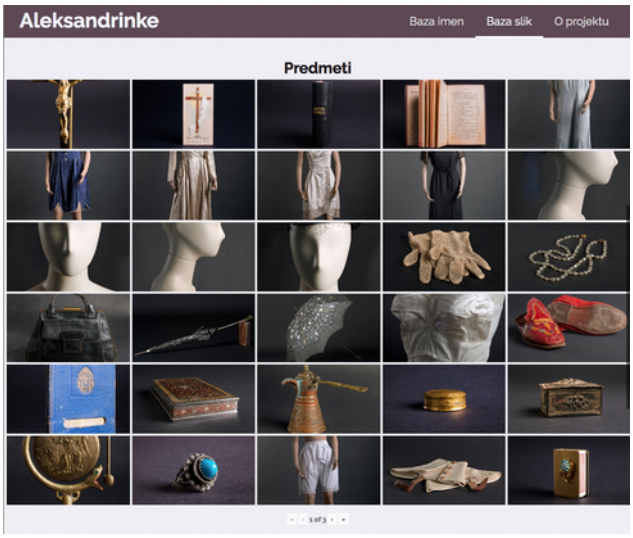
Študenta Digitalnih umetnosti in praks z Akademije umetnosti Univerze v Novi Gorici sta fotografirala materialno dediščino in zasnovala virtualno razstavo. Osnove studijske objektne fotografije, tako snemanja kot obdelave, sta spoznala na Akademiji na posebej prirejeni delavnici z mentorjem Lukom Deklevo, z delom na projektu pa sta pridobljena znanja še poglobila. Posebno pozornost sta posvetila specifikam fotografiranja muzejskih predmetov, ki so zaradi starosti zelo občutljivi. Pomemben pri ustvarjanju fotografskega izraza je bil dejanski stik s predmeti ob poznavanju njihove zgodovine,[12] saj je v detajlih predmetov materializiran marsikateri zgodovinski podatek, nenazadnje se v njih odraža tudi preplet svetov, v katerem so aleksandrinke živele – poleg večjezičnih dokumentov, kjer najdemo italijanščino, arabščino, francoščino itd. sta študenta bližnje fotografirala tudi našitka z blagovno znamko v klobuku in torbici (Paulette – Paris, Anna Staudinger).

Pri organizaciji fotografiranja smo se srečali z nekaj praktičnimi izzivi, npr. z organizacijo transporta dragocenih predmetov iz zbirke Društva in iz Goriškega muzeja v fotografski studio na Akademiji (fotografiranje na lokaciji namreč ni bilo dovolj kakovostno). Ker transport vseh predmetov ni bil mogoč, so pedagoška mentorica in strokovna sodelavca naredili ustrezen izbor predmetov. Izziv je bilo najti tudi pomožne predmete, npr. primerno lutko, na kateri bi lahko verodostojno obeležili klobuke, oblačila, čevlje, torbico ipd.

Študentke humanistike so fotografije aleksandrink opremile s podatki. Pripravile so tudi spremljevalno besedilo za celotno spletišče.

[11]Vizualizacije naštetega so bile narejene pri predmetu Osnove oblikovanja na Fakulteti za računalništvo in informatiko v letu 2020 v obliki interaktivne info grafike in digitalnega videa.

[12]Večina študentov je aleksandrinstvo spoznavala prvič, zato je bil zanje uvod v projekt najprej študij povezane literature, pisem in avdiovizualnih gradiv.




Slika 3. Baza slik – Predmeti.

Slika 4. Baza slik – dokument.

## 4.2 Družbena koristnost projekta

Serija projektov ŠIPK je namenjena vzpostavljanju povezav, ki imajo poleg študijskega poudarka ter povezovanja izobraževalne sfere z zaposlitvami s področja tudi namen izdelave rešitev in produktov, ki jih okolje in družba potrebujeta. Društvo za ohranjanje kulturne dediščine aleksandrink je pridobilo digitalno bazo podatkov in digitalno bazo fotografij o aleksandrinkah, ki omogočata in spodbujata nadaljnje raziskave. Gradivo, ki je bilo do projekta samo zbrano, ne pa obdelano, je sedaj digitalizirano in dostopno na spletu v obliki, ki omogoča analizo in vizualizacijo podatkov za namene raziskav in seznanjanja različnih občinstev s fenomenom aleksandrinstva. Fotografije predmetov so pripravljene za nadaljnjo obdelavo tako v smislu virtualnih razstav kot informativnega ali promocijskega gradiva društva.

Fenomen aleksandrinstva je do neke mere še vedno aktualen problem skupnosti. S tem ko smo ga zajeli v spletno podatkovno bazo, smo ga naredili vidnejšega in bolj preglednega, omogočili smo dodatno refleksijo in raziskave posameznih usod, kar lahko obogati védenje o problematiki migracij ne le v lokalnem-regionalnem, temveč tudi širšem prostoru, ki so ga zaznamovale in ga še zaznamujejo migracije.

# 5. Projekt Digitalizacija kulturne dediščine Primork in Notranjk (2020)

## 5.1 Izvedba projekta

Študentke in študentje so se v okviru drugega ŠIPK projekta,[13] ki se ukvarja z digitalizacijo pisemske dediščine Primork in Notranjk, uvodoma seznanili z različnimi spletnimi zbirkami pisem: *COST akcija IS1310 - Reassembling the Republic of Letters, 1500-1800*,[14] *The ePistolarium*,[15] *Encyclopedia of Romantic Nationalism in Europe*,[16] *Mapping the Republic of Letters*,[17] *Letters 1916-1923,*[18] *Visual Correspondence*,[19] *Early Modern Letters Online.*[20] Tako so pridobili širši vpogled v načine, kako se tovrstne podatke zbira, shrani in na koncu tudi predstavi. Pri pregledu področja smo analizirali vmesnike spletnih baz pisem, kar nam je pomagalo pri snovanju spletnega vmesnika za našo bazo.

Študentke humanistike so prepisale in uredile pisma iz obsežne korespondence družine Kremenšek z Notranjskega (medsebojno dopisovanje različnih članov), pisma aleksandrinke Felicite Peric svojemu možu ter pisma Zofke Kveder Marici Nadlišek Bartol in Ivanu Cankarju (gre za pisma vodilne slovenske intelektualke tistega časa svojima kolegoma) in jih vnesle v elektronsko zbirko. Kot posebej zahtevna se je izkazala transkripcija pisem Felicite in Franca Perica, v kateri so določena mesta zaradi starosti težko berljiva, veliko pa je tudi izrazov, za razumevanje in pravilen prepis katerih sta študentki potrebovali pomoč strokovne mentorice iz Društva za ohranjanje kulturne dediščine aleksandrink. Pri vnašanju pisem so se soočile z različnimi problemi, kot so nove ključne besede, različna spola pošiljateljev ali naslovnikov – zato so v sodelovanju s študentom računalništva dodajale nove kategorije v zbirko.

Študent s kompetencami s področja računalništva s Poslovno tehniške fakultete je za shranjevanje digitaliziranih pisem in podatkov o njih pripravil podatkovno bazo s pomočjo prosto dostopnega orodja phpMyAdmin. Poleg prepisov pisem hranimo v bazi še podatke o pošiljatelju, kraju, datumu, naslovniku (prejemniku pisma), naslovu in ključnih besedah v pismu. Ključne besede smo določili na začetku projekta, a smo jih še dodajali, ko smo naleteli na določene tematike.

[13] Vodja projekta: prof. dr. Katja Mihurko Poniž. Pedagoški mentorji: prof. dr. Stane Granda (Fakulteta za humanistiko, UNG), prof. dr. Katja Mihurko Poniž, doc. mag. Tina Smrekar, doc. dr. Aneta Trajanov. Strokovna sodelavca: dr. Megi Rožič (Društvo za ohranjanje kulturne dediščine aleksandrink) in Borut Koloini (Goriški muzej). Študentke in študenta: Vinka Kovačević, Maja Orel, Lara Vončina (Fakulteta za humanistiko, UNG), Domen Vinko, Anja Paternoster (Akademija umetnosti, UNG), Oliver Angelov (Poslovno-tehniška fakulteta, UNG), Tjaša Moškon (Teološka fakulteta, UL), Lučka Majcenovič, Ina Poteko, Pia Rednak (Filozofska fakulteta, UL).

[14] http://www.republicofletters.net/index.php/about/

[15] http://ckcc.huygens.knaw.nl/epistolarium/

[16] https://ernie.uva.nl

[17] http://republicofletters.stanford.edu/

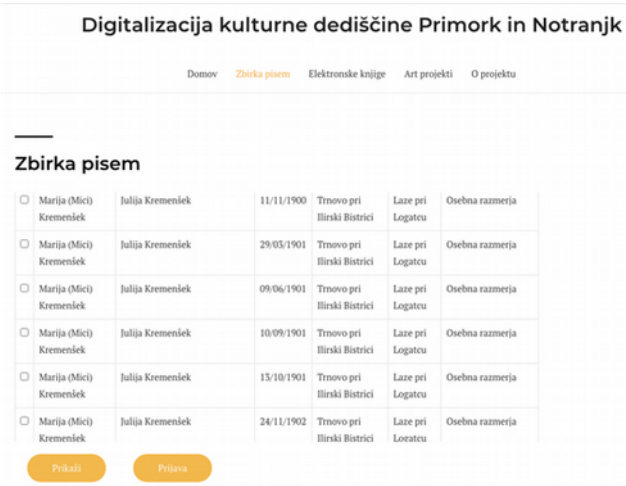
[18] http://letters1916.maynoothuniversity.ie

[19] http://letters.nialloleary.ie/index.php

[20] http://emlo-portal.bodleian.ox.ac.uk/collections/?page_id=907#context

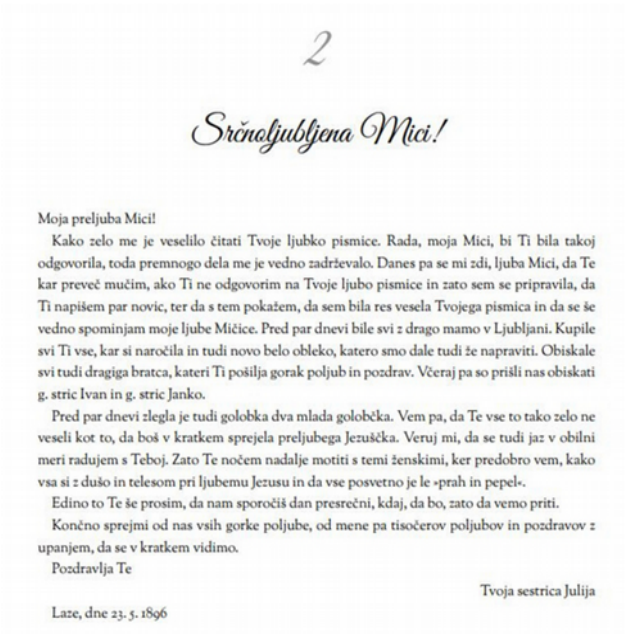
Do baze uporabniki dostopajo prek spletišča projekta, ki ga je študent razvil z orodjem WordPress. Odločili smo se, da je vsebina pisem v celoti vidna, nekatere kategorije pa niso (npr. spol pošiljatelja/pošiljateljice, ki je razviden iz imena). S prijavo je mogoče vstaviti pismo, dodati priponko (sliko pisma) in vnesti podatke v naslednja polja, ki smo jih skupaj določili na začetku projekta po preučitvi podobnih elektronskih zbirk: Pošiljatelj/Pošiljateljica, Prejemnik/Prejemnica, Poslano iz, Poslano v, Spol pošiljatelja/pošiljateljice, Spol prejemnika/prejemnice, Ključne besede in Vir. Ključne besede in spol v obeh kategorijah izbiramo iz spustnega menija. Spletna stran je postavljena na univerzitetnem spletnem strežniku http://sipk-pisma.ung.si/ in omogoča tako dostop do že naloženih vsebin projekta kot tudi vnašanje novih digitaliziranih pisem.

Slika 5. Spletišče projekta *Digitalizacija kulturne dediščine Primork in Notranjk*




Slika 6. Baza digitaliziranih pisem.

Del obravnavanega pisemskega korpusa, tj. korespondenca družine Kremenšek, je bil obdelan v obliki elektronske knjige z orodjem Pressbooks,[21] medtem ko korespondenca Felicite in Franca Perica ni bila obdelana s tem orodjem, saj je bil prepis dolgotrajen zaradi slabe kvalitete ohranjenih pisem in številnih težko čitljivih mest, tako da je transkripcija predstavljala večino ur, ki sta jih študentki imeli na projektu. Orodje Pressbooks smo izbrali, ker omogoča oblikovati knjigo, kar je za študentke in študente humanistike pomembna kompetenca, ki so jo v projektu tudi pridobile. Pri tem smo se soočili s težavo, da je kazalo naslovov v angleščini in tega ni mogoče spremeniti. Zato je študentka kazalo naredila posebej in ga nato dodala v pdf datoteko. Oblikovanje e-pub izdaje bi zahtevalo dodatno delo, za katero nam je v projektu zmanjkalo časa, prav tako bi bilo potrebno še delo na knjigi, do katere je mogoče dostopati neposredno na spletu. Ker načrtujemo nadgradnjo izdaje knjige o korespondenci družine Kremenšek z opombami in dodatnimi strokovnimi besedili, smo se odločili, da več časa namenimo urejanju korespondence, tako da smo na koncu projekta imeli zaključeni elektronski knjigi v pdf obliki. Na spletišču jima je namenjen poseben zavihek.




Slika 7. Stran iz elektronske knjige korespondence družine Kremenšek.

Inovativen dodatek v projektu, prav tako dostopen na spletišču projekta, sta dve umetniški interpretaciji arhivskega gradiva v povezavi s humanistično znanstvenoraziskovalno prakso, ki sta ju ustvarila študenta Digitalnih umetnosti in praks.

Študentka Anja Paternoster je ustvarila avtorsko delo *Ivankin dnevnik*, v katerem je na podlagi pisem družine Kremenšek sestavila kolaže iz lastnega in arhivskega vizualnega gradiva (ilustracije, fotografije) ter jih z odlomki iz pisem povezala v smiselno celoto v obliki knjige, tako v fizični kot tudi v elektronski obliki.

Po uvodnem študiju povezanih avdiovizualnih gradiv in literature je Anja Paternoster pregledala obe korespondenci, ki sta bili del projekta, ter se odločila, da bo obravnavala eno korespondenco in iz tega ustvarila umetniško knjigo. Zanimala sta jo predvsem kontekst in

[21] http://pressbooks.com

emocionalna plat, ki jo je moč razbrati iz pisem. Preko različnih vizualnih tehnik se je nato poglobila tako v vsebino pisem kot v kontekst časa in prostora, v katerih so bila pisma pisana. Večinoma je najprej izbrala del pisma, na katerega je nato ustvarila vizualno delo. Poleg nekaj arhivskih fotografij in samih pisem je risala in slikala analogno (barvice, čopič, praskanke) in digitalno (s pomočjo različnih programskih orodij), uporabila tudi lastne avtorske fotografije ter nato različne tehnike med seboj tudi digitalno kombinirala in kolažirala. V programu Adobe InDesign je vizualna dela nato združila še s tekstom. Študentka se je odločila za rabo raznoterih tipografij. Kot jukstapozicijo lastne sodobnosti je avtorica vključila fotografsko-slikarski avtoportret na koncu knjige. Publikacijo je izvozila v pdf formatu, v katerem je dostopna na spletni strani projekta.

Slika 8. *Ivankin dnevnik*

Študent Domen Vinko se je teme sprva nameraval lotiti kot fotograf, a je bil zavoljo izrednih razmer ob koronavirusni krizi primoran spremeniti fokus projekta. Svoje zanimanje je usmeril v sámo dejanje pisanja pisem in komunikacije prek njih. Uporabil je digitalizirana pisma družine Kremenšek in s pomočjo algoritmov umetne inteligence ustvaril program, ki na njihovi podlagi generira nova pisma. Uporabil je jezik Python ter JQuerry za branje teksta iz datotek. Veliko se je ukvarjal tudi z oblikovanjem izpisa pogovora umetnih inteligenc: odločil se je za videz, ki je podoben starejšemu računalniškemu terminalu. Raziskoval je, kako bi generirana pisma povezal med seboj, da bi bila odgovori drug na drugega. Njegova prvotna ideja je bila, da se pisma generirajo v živo, postopoma, in bi lahko bila odgovor na prejšnje pismo. Vendar to ni bilo mogoče zaradi omejitve univerzitetnega strežnika, kjer gostuje spletna stran projekta. Odločil se je, da generira večjo količino pisem v naprej in jih analizira glede na ključne besede. Po teh ključnih besedah se lahko pisma razvrstijo v red, kjer delujejo kot odgovori drug na drugega. Poskusi pogovorov umetnih inteligenc se pojavljajo vedno znova oz. so ves čas v teku: npr. v Laboratoriju kreativnih strojev (Creative Machines Lab) na Univerzi Cornell v ZDA (2011),[22] v podjetju Facebook, kjer različne umetne inteligence učijo pogajanja (2017),[23] ali pa globoko učenje za predvidevanje tekstov v stari Grščini na poškodovanih kamenih dekretih (2019).[24] Poskus študenta drugega letnika umetniške akademije seveda nikoli ni imel namena preboja v komuniciranju umetnih inteligenc, je pa gotovo pomenil zanimiv pristop z vidika vsebine, ki jo umetna inteligenca dobi v t. i. fazi urjenja. Algoritmu je namreč podal zbirko digitaliziranih pisem s konca 19. stoletja, vse generirane vsebine in jezik pa izhajajo iz tega, kar je bilo algoritmu podano. Gledalcu prinaša projekt nek drugačen pogled na to, kar je bilo (v času, ko je bila pisana originalna korespondenca) ter deloma tudi drugačen pogled na principe, preko katerih bodo našo sodobnost morda videli v prihodnosti, ko bodo soočeni s koščki naše realnosti, ki jo bodo morali razumeti, transkribirati, sestaviti.

Slika 9. *Pisma umetne inteligence.*

Domen Vinko je naredil tudi predstavitveni video projekta. Delo na predstavitvenem videu je, drugače kot delo na umetniških projektih, kjer sta imela študenta več svobode, potekalo bolj po principu (umetniške) storitve,

[22] https://www.youtube.com/watch?v=WnzlbyTZsQY
[23] https://engineering.fb.com/ml-applications/deal-or-no-deal-training-ai-bots-to-negotiate/
[24] https://arxiv.org/abs/1910.06262

torej v odnosu naročnik – izvajalec. Za študenta je bilo to eno prvih srečanj z (deloma) zunanjim naročnikom, kjer je bilo potrebno v digitalni obliki na koncizen način čim bolje predstaviti projekt, upoštevati večje število danih parametrov in vsebin, jih kreativno in privlačno interpretirati ter nato slediti še nekaj iteracijam popravkov. Študent je moral pri tem razumeti in zaobjeti celoto projekta, ki je sicer morda ne bi. Iz obstoječega materiala je nato pridobil, izvlekel oz. mestoma kreiral posamezne vizualno zanimive dele, jim dodal ozadja ter jih preko vključitve animacij in primerne glasbene podlage združil v kratko 40-sekundno avdiovizualno delo, ki uravnoteženo predstavlja vse dele projekta.

### 5.2 Družbena koristnost projekta

Študenti so za projekt ustvarili podatkovno bazo z bogatim korpusom ženskih pisem (skupno 475 pisem), dve elektronski knjigi, didaktično gradivo in umetniške izdelke, ki povezujejo umetnost z znanstvenimi raziskavami v hibridne tvorbe, upoštevajoč digitalne komunikacijske tehnologije. Da bi bili rezultati projekta čim bolje predstavljeni in vključeni v življenje v širšem regionalnem okolju Univerze v Novi Gorici, je študentka pedagoškega programa Slovenistike iz obdelanega gradiva pripravila interaktivno didaktično gradivo. Projekt je bil že ob začetku predstavljen v medijih, tudi končna predstavitev korespondence Felicite in Franca Peric je imela velik odmev.[25]

Na ta način smo ohranili in širšemu krogu zainteresirane strokovne in nestrokovne javnosti približali življenja žensk, kakor se razkrivajo skozi njihove lastne artikulacije. Pričevanja o vsakdanjem življenju žensk tako v ruralnem kot urbanem okolju so študentom predstavila nove tematike ter načine življenja in mišljenja, zaradi česar bodo v svojem nadaljnjem poklicnem življenju bolj pozorni in dovzetni za dosežke prezrtih družbenih skupin, ki imajo vedno svoje posebnosti. Ob obdelavi gradiva so študenti pridobili kompetence s področja DH, umetniških praks in pedagogike ter razširili svoje vedenje o zgodovini žensk iz primorskega in notranjskega prostora.

Spletišče s pismi omogoča dodajanje novih pisem in ima potencial, da postane digitalni arhiv kulturne dediščine tako vsakdanjega življenja kot tudi življenja pomembnih Slovenk in Slovencev, katerih pisma se bodo dodajala na spletišče.

## 6. Zaključek

V članku opisana projekta preizkušata metode poučevanja DH z dveh zornih kotov, prvi v reduciranem in oddaljenem pogledu na družboslovne podatke, drugi v poglobljenem pogledu na humanistične zgodovinske dokumente. Pri prvem projektu, *Digitalizacija kulturne dediščine aleksandrink*, so študenti in študentke sodelovali pri izdelavi digitalne zbirke podatkov o aleksandrinkah, ki je omogočila statistične raziskave nekaterih vidikov življenja aleksandrink. Ta izkušnja je spodbudila premislek o tem, ali je mogoče digitalizacijo razširiti in vključiti nove oblike arhivskega gradiva in kaj to prinese v zasnovi novega študentskega projekta. Drugi projekt, *Digitalizacija kulturne dediščine Primork in Notranjk*, se tako problematike, deloma tudi aleksandrink, loteva z nasprotnega konca: namesto redukcije njihovega življenja na nekaj kategorij podatkov, ki jih zajema baza, slednji prinaša digitalizirana pisma Primork in Notranjk, ki razkrivajo njihove kompleksne in individualne življenjske usode – te pa so osrednje zanimanje klasičnih humanističnih pristopov in niso reduktibilne na nekaj statističnih posplošitev. Zbirka pisem ne omogoča le vizualizacije metapodatkov, temveč tudi lingvistične analize korespondenc, saj je pisma mogoče raziskovati tudi kot korpus (četudi zaenkrat še zelo majhen). Z določitvijo ključnih besed, ki zadevajo različna področja, je pisma mogoče selekcionirati in raziskovati glede na različne topike in fenomene, kar odpira nove raziskave.

Oba projekta sta študentom in študentkam približala področje digitalne humanistike, ki je skoraj vsem med njimi bilo do tedaj povsem neznano. Ker je bila sestava skupine interdisciplinarna, so pridobili nove izkušnje sodelovanja in komuniciranja raziskovalne problematike s kolegi in kolegicami, ki ne obvladajo terminologije in metodologije novih področij. Prav tako je bila tovrstna komunikacija izziv za mentorice in mentorje, saj je bilo potrebno delo študentov in študentk ves čas pozorno spremljati in jih usmerjati. Izkušnje iz prvega projekta so bile pomemben dejavnik pri prijavi in izvedbi drugega. S stališča digitalizacije gradiva je bilo smiselno vključiti več študentk oz. študentov s področja humanistike, vendar je to pomenilo, da je elektronska zbirka bolj preprosta, kot je bila pri prvem projektu, kjer sta sodelovala dva študenta računalništva. Če bi projekt ponovno prijavljali, bi v projektno skupino vključili študenta druge stopnje, ki bi imel več kompetenc s področja računalništva in bi zastavljene naloge hitreje in lažje opravil.

Številni projekti DH so povezani z zbiranjem podatkov, ki jih raziskovalci in raziskovalke v okviru svojih znanstvenih izzivov ne morejo sami izpeljati. Morda tudi zato v slovenskem prostoru kar zadeva digitalizacijo korespondenc nismo v koraku z mednarodno DH skupnostjo. S prispevkom smo želele posredovati primere dobre prakse posredovanja DH znanj in veščin študentom in študentkam izven kurikuluma, z namenom, da bi jih navdušile za DH, saj je za našo skupnost pomembno, da mladi vanjo aktivno vstopajo.

## 7. Literatura

Ryan Cordell. 2016. How Not to Teach Digital Humanities. V: M. K. Gold, L. F. Klein, ur., *Debates in Digital Humanities 2016*.

Brett D. Hirsch. 2012. </Parentheses>: Digital Humanities and the Place of Pedagogy. V: B. D. Hirsch, ur, *Digital Humanities Pedagogy: Practices, Principles and Politics*, str. 3–30. Open Book Publishers.

#dariahTeach. https://www.dariah.eu/activities/working-groups/dariahteach/

Spletišče *Aleksandrinke*. http://green.fri.uni-lj.si:1869

Spletišče *Digitalizacija kulturne dediščine Primork in Notranjk*. http://sipk-pisma.ung.si/

[25] http://sipk-pisma.ung.si/o-projektu/

# Extracting correctly aligned segments from unclean parallel data using character n-gram matching

**Maja Popović, Alberto Poncelas**

ADAPT Centre, Dublin City University, Ireland
name.surname@adaptcentre.ie

**Abstract**

Training of Neural Machine Translation systems is a time- and resource-demanding task, especially when large amounts of parallel texts are used. In addition, it is sensitive to unclean parallel data. In this work, we explore a data cleaning method based on character n-gram matching. The method is particularly convenient for closely related language since the n-gram matching scores can be calculated directly on the source and the target parts of the training corpus. For more distant languages, a translation step is needed and then the MT output is compared with the corresponding original part. We show that the proposed method not only reduces the amount of training corpus, but also can increase the system's performance.

## 1. Introduction

Training of Neural Machine Translation (NMT) systems is a time- and resource-demanding task, especially when large amounts of parallel texts are used. In addition, NMT models are sensitive to unclean parallel data – they tend to perform better with more training data, however it is better to provide less data of better quality. In this work, we explore a data cleaning method based on character n-gram matching. The method is particularly convenient for closely related languages since the n-gram matching scores can be calculated directly on the source and the target parts of the training corpus. For more distant languages, a translation step is needed and then the MT output is compared with the corresponding original part. We show that our method can not only reduce the amount of training data and speed up the training process, but also can lead to better translation quality.

### 1.1. Related work

As the data has a big impact on the performance of the MT, several publications explored how to identify and remove sentences that could hurt the translation quality of an MT system. Usually, these are noisy parallel sentences where the meaning of the source sentence is different from the meaning of the target sentence. This may often happen in corpora where sentences are automatically crawled. In other cases, although the meaning is the same, the source and target parts are substantially diverging from a literal translation. For example, in the case of subtitling, due to the time limitation, the translations are sometimes simplified in order to be adapted to a particular time frame.

Usual approaches for filtering noisy sentence pairs include comparing characteristics of the sentences such as similarity of lengths, e.g. length difference (Taghipour et al., 2010) or length proportion (Khadivi and Ney, 2005), checking amounts of punctuation or name entities, or evaluating whether the alignment of words in the sentences are well distributed (Taghipour et al., 2010).

Many approaches also involve comparing sentences to their MT-translated version. First, an MT system is trained on the clean data, then it is used to translate the non-English part into English, and then mono-lingual matching methods are used to compare it against the original English side of the parallel corpus. Different matching metrics were used: METEOR (Erdmann and Gwinnup, 2019), Levenshtein distance (Sen et al., 2019), or BLEU (Parcheta et al., 2019) or TER (Freitag et al., 2017).

Other approaches use the word alignments (Bouamor and Sajjad, 2018; Kurfalı and Östling, 2019) or teacher-student frameworks (Freitag et al., 2017). Junczys-Dowmunt (2018) proposed Dual Conditional Cross-Entropy Filtering, where sentence pairs are scored based on the disagreement between MT-translated sentences in both directions. Axelrod et al. (2019) expanded this approach by iteratively improving the MT models with good-quality sentence pairs selected from the comparable set. Since 2018, a shared task on parallel corpus filtering[1] is included in the conference on machine transation (WMT) (Koehn et al., 2018; Koehn et al., 2019), where some of these methods were proposed and tested

While all these methods can be applied to similar language pairs, too, the method proposed in this work is much faster and less resource-intensive.

## 2. Goal of the paper

The goal of our experiments is to improve a parallel corpus used for training a NMT system and reduce its size by removing misaligned segments without significant loss in translation quality. The results show that the translation quality can even be improved. We carried out systematic experiments on two groups of closely related languages: Slovenian, Serbian and Croatian, as well as Spanish and Portuguese. The criterion for keeping or removing a parallel segment is character n-gram matching score between the source and the target languages. Different thresholds of the score were explored in order to find out the optimal one. For South Slavic languages, the experiments were carried out with Slovenian on one side and Serbian or Croatian on another side in all translation directions. For Spanish and Portuguese, we explored both translation directions.

[1] https://www.statmt.org/wmt18/parallel-corpus-filtering.html

In addition, a preliminary experiment with a distant language pair, English–Serbian, has been carried out. In this case, the source and the target part of the corpus cannot be compared directly, so MT is used as an intermediate step, and then the MT output of the training corpus is compared with the corresponding original part. This type of filtering has already been investigated (Freitag et al., 2017; Erdmann and Gwinnup, 2019; Sen et al., 2019; Parcheta et al., 2019), however using another matching metrics.

## 3. Cleaning method

Our cleaning method is based on the character n-gram based F-score CHRF[2] (Popović, 2015) which is usually used for MT evaluation to compare MT output with a human reference translation. The score is based on matched character n-grams between the two texts, with n-gram lengths n ranging from 1 to 6. When combining precision and recall into F-score, twice more weight goes to recall, since it is shown to be the best option for MT evaluation (Popović, 2016). It is fast, language-independent, tokenisation-independent, it can be used directly without training and parameter optimisation, and it shows good correlations with human judgments. In 2018, it became one of the baseline automatic metrics in the WMT shared task.[3] Recently, in a study encompassing a number of automatic MT metrics (Mathur et al., 2020). the CHRF score is recommended as a replacement for the widely used BLEU and TER scores.

**Closely related langugages** While designed for comparing two translations in the same language, the CHRF score can provide useful information about the degree of matching between two texts written in closely related languages. The intuitive idea is that the properly aligned segments will have higher CHRF scores than misaligned ones.

Therefore, for the Slovenian–Croatian/Serbian and Spanish–Portuguese corpora, we calculate the CHRF score for each parallel segment in the training corpora. Examples from the Slovenian–Croatian/Serbian corpora can be seen in Table 3., ordered according their CHRF scores (from highest to lowest). Due to similarity between the languages, the scores between the properly aligned segments are indeed relatively high (top part of the table) in contrast to misaligned segments (10,11,12). Nevertheless, the languages are sufficiently different so that some properly aligned short segments (words) have low scores (8, 9) and might be removed in the process. Still, if those words appear in longer sentences ("Nihče" in 4 and "Veš" in 5), these sentences will get high scores and will not be removed.

Similar tendencies can be found in the Spanish–Portuguese corpus.

**Distant languages** While the CHRF score can be used for more distant languages, too, the method becomes complicated. The segments cannot be compared directly, so that translation of one part of the training corpus is necessary, which imposes certain challenges. First, the process is not fast anymore since translation of a training corpus requires its time. Second, the translation process introduces more divergences so that more properly aligned segments might end up with lower scores. Third, quality of the MT system used for cleaning has to be taken into account. Generally, better systems result in higher automatic evaluation scores, so the optimal threshold for one MT system is not necessarily the same as for another system. Also, it should be investigated whether lower-quality systems can perform well by increasing the threshold, or the best option would be to use high quality system. Finally, translation direction might play an important role, too, because the performances of the systems in two translation directions usually differ.

Taking all this into account, we carried out a preliminary experiment on the English→Serbian translation. We analysed a system trained on a small training corpus (2M segments), and we cleaned the data using a better system trained on large training corpus (30M segments). The English part was first translated by the "cleaner" system, and then the CHRF scores were calculated between the translated and the original Serbian corpus.

Examples for English–Serbian are shown in Table 3.. It can be noted that the correlation between the CHRF score and the alignment quality became more blurred. Properly aligned segments often yield high scores (1, 2) whereas misaligned segments often yield lower scores (6, 8). Nevertheless, less literal translations with more shifts and divergences from the source text such as segments 5 and 7 also exhibit low scores. Another unnecessarily low score can be seen in a properly aligned segment 9 where the MT system left the word untranslated.

## 4. Experimental set-up

**Data** One of the largest sources of publicly available parallel data are *OpenSubtitles* from the OPUS[4] data collection (Tiedemann, 2012). Before the training of the baseline systems, the original parallel data were filtered in order to eliminate noisy parts: too long segments (more than 100 words), segment pairs with disproportional sentence lengths, segments with more than 1/3 of non-alphanumeric characters, as well as duplicate segment pairs were removed. The statistics of the remaining subtitles together with the development and test sets is shown in Table 3.

The development and test data for the closely related language pairs were provided by organisers of the WMT 2020 shared task on translation between similar languages.[5]

For the experiment with English–Serbian, the development test was separated from the training data and consists mainly of subtitles with a small portion of news. The test set consists of publicly available *IMDb* movie reviews[6] used in Lohar et al. (2019).

**MT systems** All NMT systems are built using the Sockeye implementation (Hieber et al., 2017) of the Transformer architecture (Vaswani et al., 2017). The systems operate on sub-word units generated by byte-pair encoding (BPE) (Sennrich et al., 2016). We set the number of BPE merging operations at 32000. We use shared vocabularies

[2] https://github.com/m-popovic/chrF
[3] http://www.statmt.org/wmt18/
[4] http://opus.nlpl.eu/
[5] http://www.statmt.org/wmt20/similar.html
[6] https://github.com/m-popovic/imdb-corpus-for-MT

| | properly aligned | Slovenian | Croatian/Serbian | CHRF score |
|---|---|---|---|---|
| 1) | yes | Hvala. | Hvala. | 100 |
| 2) | yes | Pa njegova moč? | A njegova moć? | 63.34 |
| 3) | yes | Zahtevam, da me izpustite. | Zahtijevam da me pustite. | 50.29 |
| 4) | yes | Nihče ne govori o iztrebljanju. | Nitko ne govori o istrebljenju. | 47.74 |
| 5) | yes | Veš, na kaj mislim. | Znaš na što mislim. | 37.51 |
| 6) | yes | Kaj je narobe s tem norim mačkom? | Što je s tim ludim mačkom? | 34.10 |
| 7) | yes | Bi rad vedel, kje je? | Zanima te gdje je? | 20.51 |
| 8) | yes | Veš. | Znaš. | 13.14 |
| 9) | yes | Nihče! | Nitko! | 8.89 |
| 10) | no | Seveda je bilo nekaj občudovanja vrednih kotičkov ob Temzi. | Poznat je i kao kavez. | 7.54 |
| 11) | no | Comprenez vous? | Razumiješ li? | 6.13 |
| 12) | no | Uh, uh, pogumno, naprej! | Kak' ste, šefe? | 2.58 |

Table 1: Examples from the corpus with closely-related languages: Slovenian segments (left), Serbian/Croatian segments (middle), and character n-gram matching scores between them (right). Correlation between alignment quality and the score can be observed.

| | properly aligned | English original | Serbian MT from English | Serbian original | CHRF score |
|---|---|---|---|---|---|
| 1) | yes | The interior of Texas at that time was completely unknown to the outside world. | Unutrašnjost Teksasa u to vreme bila je potpuno nepoznata spoljnom svetu. | Unutrašnjost Teksasa u to vreme je bila potpuno nepoznata spoljnom svetu. | 90.44 |
| 2) | yes | But they found a strange and evil country, | Ali našli su čudnu i zlu zemlju, | Ali su naišli na čudnu i zlu zemlju. | 63.87 |
| 3) | yes | This is a cattail. | Ovo je mačji rep. | Ovo je rogoz. | 27.62 |
| 4) | no | I don't know whether she's genuinely interested in protecting her husband, or the money. | Ne znam da li je zaista zainteresovana da zaštiti svog muža, ili novac. | Moram doma. Treba me ujak Martin. Ne mogu te tako pustiti. | 15.75 |
| 5) | yes | It was easy pickings. | To je bio lak plen. | To su mogli lako da nađu. | 12.53 |
| 6) | no | I don't wanna know. | Ne želim da znam. | Voliš li me? | 11.47 |
| 7) | yes | Now, those are minerals. | To su minerali. | A sad evo ruda. | 9.56 |
| 8) | no | How did you find out about the apartment? | Kako si saznao za stan? | Ne bi ga dala zaručniku. | 8.51 |
| 9) | yes | (THUNDER) | (THUNDER) | (GRMLJAVINA) | 5.84 |

Table 2: Examples from the corpus with distant languages: English original segments (left), MT outputs generated from them (middle left), Serbian original segments (middle right), and character n-gram matching scores between original Serbian and MT outputs from original English (right). Correlation between alignment quality and the score is not obvious, and also might depend on the performance of the used MT system.

for translation systems between the similar languages, and separated vocabularies for English–Serbian systems.

For translation between the South Slavic languages, we tested the data cleaning on multisource and multitarget systems as proposed in Johnson et al. (2017). Multisource setting was used for translation from Serbian/Croatian into Slovenian, and multitarget setting for the other translation direction. We decided to use multilingual systems (with joint Serbian and Croatian data) as baselines because they perform better than bilingual systems, which can be seen in Table 6. For translation between Spanish and Portuguese, only bilingual systems were possible to build. For all pairs of similar languages, we investigated CHRF thresholds in the range from 10 to 50 under the assumption that the threshold of 10 will probably retain too many misaligned segments while the threshold of 50 will remove too many proper segment pairs resulting in too small parallel corpus, so that the optimal threshold lies something between these two values.

For translation between English and Serbian, the system used for cleaning is trained on all publicly available corpora consisting mainly of subtitles (about 30 million segments) as well as a small portion of news (200k segments) and other sources (100k segments). For the preliminary experiment described in this work, we took 2 million segments from subtitles as a starting point and explored the clean-

(a) training corpora

| languages | | domain | # sentences |
|---|---|---|---|
| close | sl–hr | subtitles | 11 213 386 |
| | sl–sr | subtitles | 11 780 062 |
| | pt–es | subtitles | 22 317 343 |
| distant | en–sr | subtitles | 2 000 000 |
| | “cleaner" | mainly subtitles | 30 002 892 |

(b) development and test sets

| languages | | set | domain | # sentences |
|---|---|---|---|---|
| close | sl–hr | dev | PR publications | 2457 |
| | | test | PR publications | 2582 |
| | sl–sr | dev | PR publications | 1259 |
| | | test | PR publications | 1260 |
| | pt–es | dev | web and technical | 1283 |
| | | test | web and technical | 1495 |
| distant | en–sr | dev | subtitles+news | 1000 |
| | | test | movie reviews | 485 |

Table 3: Corpora statistics.

ing of this corpus. The English part was translated by the system trained on the full corpus, and the CHRF scores are calculated between the translation and the original Serbian part. For this language pair, we investigated three CHRF thresholds, 30, 40 and 50 under the assumption that translations below the threshold of 30 are not of sufficient quality to be relied on. This assumption is based on a qualitative manual inspection of the corpus.

**Evaluation** We evaluate the outputs of our systems using the following three automatic overall evaluation scores: word n-gram precision metric BLEU (Post, 2018), character n-gram F-score chrF (used for cleaning) (Popović, 2015), and character-level translation edit rate characTER (Wang et al., 2016). The BLEU score is used because of the long tradition. The two character level scores are shown to correlate much better with human assessments (Bojar et al., 2017; Ma et al., 2018), especially for morphologically rich languages.

## 5. Results

### 5.1. Closely related languages

Table 4 shows results for the translation between the South Slavic languages. It can be seen that the initial hypothesis is true: overall, the best threshold is chrF=20, lying between 10 and 50. It reduced the training corpus by almost half, and surpassed the scores of the system trained on the full data. Larger thresholds remove too many useful segments therefore deteriorating the scores. Threshold of 10, while still yielding better automatic scores than the baseline system, performs worse than the system with the threshold 20.

Four examples of translation from Slovenian into Serbian can be seen in Table 5. For the sentence #1, setting the threshold to 40 results in a mistranslation, while CHRF-10 and CHRF-20 achieve the same performance as the baseline system. The sentence #2 is fully correctly translated only by the CHRF-10 system. The only problem with the CHRF-20 system are two named entitites which, howvever, do not change the meaning of the sentence. Named entities are also corrupted by the baseline system, one is unnecessarily transcribed and another one unnecessarily translated, thus deteriorating adequacy. The CHRF-40 system, apart from an error in the named entity, generates a mistranslation, a local word order error, as well as deletion of a phrase which significantly changes the meaning. In the third sentence, the worst performance can again be observed for the baseline and CHRF-40 systems. Other two systems contain errors which do not change the meaning: the CHRF-10 system deletes a temporal adverb, while the CHRF-20 system makes a sub-optimal lexical choice. For the fourth sentence, the baseline and the CHRF-20 systems generate a correct output while CHRF-10 and CHRF-20 systems generate a mistranslation.

Same tendencies, although less clearly, can be observed for the translation between Spanish and Portuguese presented in Table 6.

### 5.2. Distant languages

The results for English->Serbian are shown in Table 5.2.. For the sake of completeness, we also show the results of the “cleaner" system trained on the full uncleaned data (row 1). The cleaning is performed on the 2M subset (row 2). For the in-domain test set, removing 0.5M segments with chrF<30 yields better BLEU and characTER and comparable CHRF than when all 2M segments are used. For the movie reviews test set, however, it does not surpass the original performance, but there is no substantial loss.

These results confirm the findings reported in (Freitag et al., 2017), namely that knowledge distillation using a “big" translation system and an automatic evaluation metric is a promising concept for reducing resources necessary for NMT training. Nevertheless, more systematic experiments should be carried out in the future, which should include different cleaning systems with different CHRF thresholds.

## 6. Conclusions

We propose a method for cleaning parallel training data between closely related languages based on character n-gram matching. The results on South Slavic and Spanish–Portuguese language pairs show that our method can not only reduce the amount of training data and speed up the training process, but it can also lead to better translation quality. Overall, the best option is to remove all segments with the chrF score below 20.

Our preliminary experiment on a more distant language pair, English–Serbian, which also requires an additional translation step, also shows promising results. However, more systematic experiments should be carried out in this direction, such as using different MT systems with different CHRF thresholds for cleaning, using the other translation direction for cleaning, as well as cleaning larger amounts of training data.

## Acknowledgments

The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under

(a) Croatian→Slovenian

| training | | dev, hr→sl | | | test, hr→sl | | |
|---|---|---|---|---|---|---|---|
| system | size | BLEU↑ | chrF↑ | chrTER↓ | BLEU↑ | chrF↑ | chrTER↓ |
| HR→SL | 11.2M | 38.5 | 65.7 | 29.4 | / | / | / |
| HR+SR→SL | 20.2M | 38.8 | 65.9 | 29.5 | 34.7 | 62.2 | 34.5 |
| +CHRF-50 | 1.4M | 25.2 | 54.3 | 37.3 | 24.3 | 53.7 | 36.6 |
| +CHRF-40 | 2.8M | 31.4 | 60.4 | 32.1 | 30.1 | 59.5 | **32.5** |
| +CHRF-30 | 5.6M | 36.6 | 64.3 | **29.0** | **35.6** | **63.6** | **29.5** |
| +CHRF-20 | 10.7M | **39.7** | **66.5** | **27.0** | **37.1** | **65.2** | **28.2** |
| +CHRF-10 | 17.2M | ***39.3*** | ***66.1*** | ***28.4*** | **37.1** | ***65.0*** | ***28.8*** |

(b) Serbian→Slovenian

| training | | dev, sr→sl | | | test, sr→sl | | |
|---|---|---|---|---|---|---|---|
| system | size | BLEU↑ | chrF↑ | chrTER↓ | BLEU↑ | chrF↑ | chrTER↓ |
| SR→SL | 11.8M | 40.6 | 67.2 | 30.3 | / | / | / |
| HR+SR→SL | 20.2M | 42.1 | 68.3 | 28.5 | 37.7 | 64.1 | 33.5 |
| +CHRF-50 | 1.4M | 25.7 | 54.8 | 38.5 | 26.5 | 55.8 | 34.9 |
| +CHRF-40 | 2.8M | 33.9 | 62.0 | 32.2 | 33.0 | 62.2 | **30.4** |
| +CHRF-30 | 5.6M | 39.2 | 66.1 | 29.0 | **37.9** | **66.2** | **27.4** |
| +CHRF-20 | 10.7M | **42.2** | **68.6** | **26.9** | **41.2** | **68.1** | **26.5** |
| +CHRF-10 | 17.2M | **42.5** | **68.9** | ***28.0*** | ***40.0*** | ***67.5*** | ***27.8*** |

(c) Slovenian→Croatian

| training | | dev, sl→hr | | | test, sl→hr | | |
|---|---|---|---|---|---|---|---|
| system | size | BLEU↑ | chrF↑ | chrTER↓ | BLEU↑ | chrF↑ | chrTER↓ |
| SL→HR | 11.2M | 33.4 | 62.6 | 33.0 | / | / | / |
| SL→HR+SR | 20.2M | 36.0 | 63.8 | 32.6 | 30.3 | 58.9 | 40.0 |
| +CHRF-50 | 1.4M | 25.5 | 54.9 | 35.9 | 24.2 | 53.8 | **36.0** |
| +CHRF-40 | 2.8M | 31.2 | 61.1 | **30.8** | **30.5** | **60.7** | **30.9** |
| +CHRF-30 | 5.6M | 35.3 | **64.4** | **29.1** | **34.4** | **64.0** | **29.2** |
| +CHRF-20 | 10.7M | **36.9** | **65.2** | **28.6** | **35.7** | **64.4** | **28.8** |
| +CHRF-10 | 17.2M | ***36.3*** | ***64.8*** | **28.6** | ***34.8*** | ***64.2*** | ***29.1*** |

(d) Slovenian→Serbian

| training | | dev, sl→sr | | | test, sl→sr | | |
|---|---|---|---|---|---|---|---|
| system | size | BLEU↑ | chrF↑ | chrTER↓ | BLEU | chrF | chrTER↓ |
| SL→SR | 11.8M | 33.3 | 62.3 | 34.3 | / | / | / |
| SL→HR+SR | 20.2M | 34.8 | 63.4 | 33.4 | 32.0 | 60.0 | 36.4 |
| +CHRF-50 | 1.4M | 22.9 | 53.5 | 38.7 | 26.0 | 55.2 | **34.2** |
| +CHRF-40 | 2.8M | 29.2 | 59.9 | **32.9** | 30.2 | **60.6** | **30.2** |
| +CHRF-30 | 5.6M | 34.6 | **63.4** | **31.7** | **36.4** | **64.6** | **27.8** |
| +CHRF-20 | 10.7M | **35.5** | **64.2** | **31.5** | **37.0** | **65.1** | ***28.2*** |
| +CHRF-10 | 17.2M | ***35.1*** | ***63.8*** | ***31.9*** | **37.7** | **65.6** | **27.4** |

Table 4: Results for translation between South Slavic languages: Croatian→Slovenian (a), Serbian→Slovenian (b), Slovenian→Croatian (c) and Slovenian→Serbian: corpus size and the three automatic MT evaluation scores (BLEU, chrF and characTER).

the European Regional Development Fund (ERDF) through Grant 13/RC/2106. This research was partly supported by the European Association for Machine Translation under its programme "2019 Sponsorship of Activities".

## 7. References

Amittai Axelrod, Anish Kumar, and Steve Sloto. 2019. Dual monolingual cross-entropy delta filtering of noisy parallel data. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 245–251, Florence, Italy.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation (WMT 2017)*, pages 489–513, Copenhagen, Denmark, September.

Houda Bouamor and Hassan Sajjad. 2018. H2@ bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proc. Work-*

| | |
|---|---|
| source #1 | Strojno učenje omogoča analizo nepredstavljive količine podatkov hitreje kot kdaj koli prej. |
| SL→SR(+HR) | Mašinsko učenje omogućava analizu nezamislive količine podataka brže nego ikada *ranije*. |
| +CHRF-10 | Mašinsko učenje omogućava analizu nezamislive količine podataka brže nego ikada *pre*. |
| +CHRF-20 | Mašinsko učenje omogućava analizu nezamislive količine podataka brže nego ikada. |
| +CHRF-40 | **Strujno** učenje omogućava analizu nezamislive količine podataka brže nego ikad. |
| source #2 | IBM-ov super računalnik Deep Blue je že več kot 10 let nazaj premagal najboljšega šahista Garryja Kasparova. |
| SL→SR(+HR) | IBMov super kompjuter, **Dip Blu**, je pobedio najboljeg šahista Gerija **Iskra** pre više od 10 godina. |
| +CHRF-10 | IBMov super kompjuter "Deep Blue" je pobedio najboljeg šahista Gerija Kasparova pre više od 10 godina. |
| +CHRF-20 | **IBM** superkompjuter Deep Blue je pobedio najboljeg šahistu **Geri** Kasparova pre više od 10 godina. |
| +CHRF-40 | IBMov super kompjuter Deep Blue **već je prebio** najboljeg šahista **Garrya** Kasparova. |
| source #3 | Trenutno je največ govora o avtonomni vožnji – vožnji avtomobilov brez vpletenosti voznika. |
| SL→SR(+HR) | Trenutno se najviše **radi** o autonomnoj vožnji... vožnji automobila bez **ičije pomoći** vozača. |
| +CHRF-10 | **XXX** Najviše se priča o autonomnoj vožnji auta bez uključivanja vozača. |
| +CHRF-20 | Trenutno se najviše priča o autonomnoj vožnji automobila bez **umešanosti** vozača. |
| +CHRF-40 | Trenutno **ima najviše govora** o autonomnoj vožnji. **Voziranje** automobila bez **umešanosti** vozača. |
| source #4 | Toda Google je pri tem šel še korak dalje. |
| SL→SR(+HR) | Ali Google je otišao korak dalje. |
| CHRF-10 | Ali Google je **nastavio** dalje. |
| CHRF-20 | Ali Google je otišao korak dalje. |
| CHRF-40 | Ali Google je **prolazio** i korak dalje. |

Table 5: Four examples of translation from Slovenian into Serbian: baseline system and systems with three different thresholds: chrF=10, chrF=20 and chrF=40. Errors are marked in bold ("XXX" denotes a missing word).

(a) Spanish→Portuguese

| training | | dev, es→pt | | | test, es→pt | | |
|---|---|---|---|---|---|---|---|
| system | size | BLEU↑ | chrF↑ | chrTER↓ | BLEU↑ | chrF↑ | chrTER↓ |
| ES→PT | 22.3M | 30.4 | 59.1 | 37.2 | 36.9 | 60.8 | 35.7 |
| +CHRF-50 | 2.1M | 26.8 | 56.2 | 40.5 | 31.3 | 56.8 | 39.4 |
| +CHRF-40 | 4.5M | 29.4 | 58.4 | 38.3 | 34.8 | 59.5 | 36.8 |
| +CHRF-30 | 8.9M | 30.1 | 58.8 | 37.6 | 35.5 | 60.1 | 36.2 |
| +CHRF-20 | 14.8M | **30.8** | **59.3** | **36.8** | *36.7* | *60.7* | **35.7** |
| +CHRF-10 | 20.4M | *30.4* | 58.9 | ***37.1*** | *36.7* | 60.6 | **35.4** |

(b) Portuguese→Spanish

| training | | dev, pt→es | | | test, pt→es | | |
|---|---|---|---|---|---|---|---|
| system | size | BLEU↑ | chrF↑ | chrTER↓ | BLEU↑ | chrF↑ | chrTER↓ |
| PT→ES | 22.3M | 35.7 | 62.0 | 34.3 | 40.5 | 63.4 | 33.1 |
| +CHRF-50 | 2.1M | 30.6 | 58.6 | 37.7 | 35.3 | 60.2 | 35.8 |
| +CHRF-40 | 4.5M | 34.4 | 61.2 | 35.2 | 38.8 | 63.0 | 33.2 |
| +CHRF-30 | 8.9M | 35.3 | *61.9* | *34.4* | **40.7** | **64.0** | **32.4** |
| +CHRF-20 | 14.8M | **35.8** | **62.0** | 34.7 | **40.7** | ***63.8*** | **32.3** |
| +CHRF-10 | 20.4M | *35.7* | **62.1** | *34.4* | *40.5* | ***63.6*** | ***32.8*** |

Table 6: Results for translation between Western Romance languages: Spanish→Portuguese (a) and Portuguese→Spanish (b): corpus size and the three automatic MT evaluation scores (BLEU, chrF and characTER).

*shop on Building and Using Comparable Corpora*, pages 43–47, Miyazaki, Japan.

Grant Erdmann and Jeremy Gwinnup. 2019. Quality and coverage: The AFRL submission to the WMT19 parallel corpus filtering for low-resource conditions task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 267–270, Florence, Italy, August.

Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. 2017. Ensemble distillation for neural machine translation. *arXiv preprint arXiv:1702.01802*.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Mac-

| English->Serbian | | open subtitles + news | | | IMDb movie reviews | | |
|---|---|---|---|---|---|---|---|
| system | corpus size | BLEU↑ | chrF↑ | cTER↓ | BLEU↑ | chrF↑ | cTER↓ |
| FULL (“cleaner") | 30M | 35.5 | 57.6 | 46.1 | 31.5 | 56.9 | 39.3 |
| OPENSUB-2M | 2M | 19.2 | 43.1 | 55.0 | **17.1** | **43.4** | **52.9** |
| +CHRF-50 | 943k | 17.9 | 40.6 | 56.7 | 15.5 | 40.0 | 55.7 |
| +CHRF-40 | 1.2M | 19.1 | 42.2 | 55.4 | 16.1 | 41.6 | 54.6 |
| +CHRF-30 | 1.5M | **20.3** | **43.0** | **54.6** | 17.0 | 42.5 | 53.2 |

Table 7: Results for translation between distant languages: corpus size and the three automatic MT evaluation scores (BLEU, chrF and characTER) for translation from English into Serbian on in-domain (mainly subtitles) and out-of-domain (movie reviews) test sets.

duff Hughes, and Jeffrey Dean. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Brussels, Belgium.

Shahram Khadivi and Hermann Ney. 2005. Automatic filtering of bilingual corpora for statistical machine translation. In *International Conference on Application of Natural Language to Information Systems*, pages 263–274.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the wmt 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Belgium, Brussels, October.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the wmt 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 56–74, Florence, Italy, August.

Murathan Kurfalı and Robert Östling. 2019. Noisy parallel corpus filtering through projected word embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 277–281, Florence, Italy.

Pintu Lohar, Maja Popović, and Andy Way. 2019. Building english-to-serbian machine translation system for imdb movie reviews. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 105–113, Florence, Italy.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 671–688, Belgium, Brussels, October.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4984–4997, Online, July.

Zuzanna Parcheta, Germán Sanchis-Trilles, and Francisco Casacuberta. 2019. Filtering of noisy parallel corpora based on hypothesis generation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 282–288, Florence, Italy, August.

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

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany, August.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Sukanta Sen, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Parallel corpus filtering based on fuzzy string matching. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 289–293, Florence, Italy, August.

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 (ACL 2016)*, pages 1715–1725, Berlin, Germany, August.

Kaveh Taghipour, Nasim Afhami, Shahram Khadivi, and Saeed Shiry. 2010. A discriminative approach to filter out noisy sentence pairs from bilingual corpora. In *Proceedings of 5th International Symposium on Telecommunications (IST 2010)*, pages 537–541, Tehran, Iran.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Eight International Conference on Language Resources and Evaluation, LREC 2012*, pages 2214–2218, Istanbul, Turkey.

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, Long Beach, USA.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany.

# Evaluation of related news recommendations using document similarity methods

**Marko Pranjić*§, Vid Podpečan†, Marko Robnik-Šikonja‡, Senja Pollak†**

* Jožef Stefan International Postgraduate School
Jamova cesta 39, 1000 Ljubljana, Slovenia

§ Trikoder d.o.o.
Ulica Miroslava Miholića 2, 10010 Zagreb, Croatia
marko.pranjic@styria.ai

† Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
vid.podpecan@ijs.si senja.pollak@ijs.si

‡ University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana, Slovenia
marko.robnik@fri.uni-lj.si

**Abstract**

A set of related articles is a useful addition to the newly published news. Such news articles contain more context and background information and provide a richer experience to the reader. Currently, the work of finding related articles is often done manually by the journalists writing the news story. The process can be automatized by suggesting relevant articles based on the similarity with the new article. We compare several link recommendation methods on the news archive of popular Croatian website 24sata. Our results show that the tf-idf weighting applied to bag-of-words document representation offers better matching with manually selected links by journalist than more sophisticated approaches, such as latent semantic indexing, doc2vec, and multilingual contextual embeddings BERT and XLM-R.

## 1. Introduction

Modern technology is changing the journalism and readers are developing new habits and expectations. Media houses throughout the world are embracing the opportunities of digital publishing and distributing the content on the web and through mobile applications. A common feature of online news are references (i.e. links) to other relevant news articles that provide more context or relevant background information. This makes other content, relevant for the current story, more accessible to readers, while media houses benefit from more efficient use of existing content and improved business metrics such as user engagement and the total time spent on a site.

Large media houses employ modern natural language processing (NLP) technologies to ease the task of content production and improve readers' satisfaction. Currently, advanced technological support to journalists is limited to large media houses in high-resource languages like English. In other cases, the task of finding and recommending related news articles remains the job of journalists writing the news story. The work described in this paper was done within the context of the EU H2020 EMBEDDIA project[1], whose objective is to develop and adapt state-of-the-art NLP technologies to less-resourced languages, and test them in real-world news and media production contexts.

In this paper, we compare several existing NLP methods that can be used to automate search for the related news. As the working dataset we use articles from 24sata[2], a major Croatian news web portal. We approach the problem with several text embedding methods and evaluate semantic similarity between articles using the cosine similarity on the selected representations.

The paper consists of seven sections. In Section 2, we present the related work on linking related news. In Section 3, we present the dataset of news articles used in this evaluation and its preprocessing. Section 4 describes the methods used. In Section 5, we describe the experimental settings, and in Section 6, we present the results. In Section 7, we conclude the work and suggest ideas for further work.

## 2. Related Work

Various research has explored linking news articles with the additional information. A task of finding background news given a Twitter message is introduced in Guo et al. (2013) and is performed using a named-entity recognition, matrix factorization methods, and graph analysis. In Gamon et al. (2008) and Ikeda et al. (2006), systems for linking news article with blog posts are described.

In recent years, there was an increase in the interest for document linking in the context of the news articles. In

[1] http://embeddia.eu/

[2] http://www.24sata.hr

2018, NIST in a partnership with the Washington Post, created a new TREC track known as the News Track (Huang et al., 2018; Soboroff et al., 2018). The main task of the News Track is the Background Linking task, defined as a problem of retrieving news articles that provide important context or background information that helps readers to better understand the query article. A number of researchers evaluated approaches to linking the query article with the background information. They attempted keyword extraction coupled with reranking on the number of matched paragraphs (Lu and Fang, 2018), graph-based analysis (Essam and Elsayed, 2019), or named-entity recognition (Bimantara et al., 2018). In Foley et al. (2019), authors evaluated approaches using keyphrase detection, clickbait classification and topic detection. In Lu and Fang (2019) and van der Sluis et al. (2018), authors describe systems for background linking of the news articles based on the entity linking. Most of those methods extract parts of text and use them as a query instead of comparing the whole document. An alternative approach is to represent the entire document as a vector in a semantic vector space. In this way, related documents can be found in the local neighborhood of the query (Soloshenko et al., 2015; Sitikhu et al., 2019). This is the approach we focus on in this paper.

## 3. Dataset description

The data used for analysis contains news articles from *24sata*, the biggest Croatian news publisher. The dataset contains 546,801 articles published online between 2007-03-12 and 2019-04-24. The length of news articles in the datasets vary from 11 to 19,695 words, with an average length of 272 words. The distribution of article lengths is depicted in Figure 1. All articles are written in Croatian language and each entry contains the **title**, **lead_text**, **content** of the article, **author**, content **tags**, the **section** of the newspaper where the article appears, **published_date** (the date when the article was published on the website), **created_date**, (the date when the article was originally written), and **related_articles** containing a list of references to other articles. The *published_date* and *created_date* can differ as some articles are written in advance to be published at a later time. The list of related articles is chosen by the journalist when the article is written and links to related articles are embedded in the content of the article when the article is published. In this paper, this information about related articles is used as a ground truth for evaluation of different related news recommendation methods.

Articles in our collection published before 2017-08-16 do not have any related article assigned, and also many later articles are missing this information. In total, there are 35,289 articles that have some *related_articles* assigned to them.

The dataset was split in such a way that 75% of the articles with the oldest *created_date* are available for training, and the rest form the testing set. The *related_articles* links for items that are found in the test set but reference news articles in the training set are discarded. Among 410,101 articles in the training set, there are 2,654 with related articles. Those related articles reference only articles from the training set and we use them to finetune the mBERT

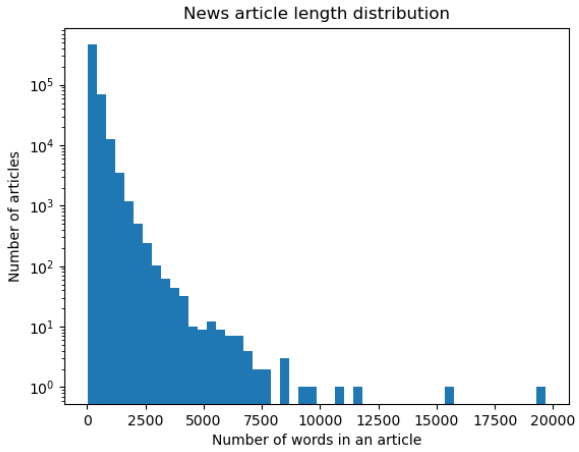



Figure 1: Distribution of number of words for the news articles in the dataset. The number of articles is shown on the logarithmic scale.

and XLM-R models as well as to determine the hyperparameters for Doc2Vec model. In the remaining 25% of the data used as the test set, there are 32,635 items with related articles and all their references are part of the test set.

### 3.1. Data preprocessing

The data contains several attributes such as tags, sections, author names etc. which might provide relevant additional information and improve document retrieval. For example, focusing the search to a subset of articles from the same section or to articles containing matching tags should improve results if such information is available during the model inference. Another requirement present in the real system is to take into account the date of the article and preferably return newer articles. In this paper, our goal is to compare the methods working on text and not on evaluating the whole system for linking the background information so this supporting metadata is ignored.

The text of an article is spread between title, lead and content fields and the first step was to concatenate these three fields. We used three different preprocessing settings, suitable for three data representations used by document matching algorithms (see Section 4.), bag-of-words representation, paragraph embeddings, and contextual embeddings.

1. Bag-of-words text representation benefits from preprocessing which removes noise and performs normalization. Following tokenization based on regular expression that preserves alphanumeric characters, we filtered out tokens of length 1 and all numbers, performed lemmatization with the updated Lemmagen lemmatizer[3] (Juršič et al., 2010), and filtered stopwords using a list of 325 Croatian stopwords.

2. The paragraphs which serve as an input to the Doc2Vec model are tokenized with the regular expression that preserves only alphanumeric characters and the obtained tokens are lemmatized.

[3] https://github.com/vpodpecan/lemmagen3

3. While the input to contextual embedding models (mBERT and XLM-R) may be lightly preprocessed (e.g., removing the URLs), in our case we performed no preprocesing and used tokenizers provided with the implementation of these models.

## 4. Document comparison methods

In our approach, the search for the most similar documents compares vectors of documents using cosine similarity. We compare two classical pre-neural approaches using bag-of-words representation, tf-idf and latent semantic indexing (LSI), with approaches using dense vector embeddings computed with neural networks. In the later group, we use the Doc2Vec (Le and Mikolov, 2014a) model, as well as the contextual multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R) models. Below we shortly describe each of the approaches.

### 4.1. Tf-idf

A standard pre-neural approach to document retrieval is to transform the documents into a bag-of-words (BOW) representation which is essentially a collection of sparse integer vectors (word counts). As BOW vectors do not take into account the relative importance of words in a document with respect to the whole corpus, they are typically re-weighted. A successful word reweighting method, term-frequency inverse-document-frequency (*tf-idf*), aims to reflect word importance in a document (*tf)* relative to its importance in the corpus (*idf*).

The preprocessing of a dataset (see Section 3.1.) returns a list of tokens for every document. These lists are transformed into sparse numeric vectors by first extracting the corpus vocabulary, computing word frequencies for each document (*tf*), and computing *tf-idf* weighted vectors using the vocabulary and overall word counts.

When compiling the corpus vocabulary, additional filtering parameters can be set. We used the default settings where tokens which appear in less than 5 documents or in more than 50% of all documents are filtered out. In addition, we experimented with setting these limits to 2 and 25%, respectively. The effects of different settings on the performance are presented in Section 6.

In our implementation, we used the `TfidfModel` from the Gensim library (Řehůřek and Sojka, 2010) and the default *nfc* SMART setting[4] for *tf-idf* weighting and normalization which used raw frequency for term frequency weighting, inverse document frequency for document frequency weighting and cosine document length normalization. This translates into the following set of equations for term frequency, document frequency, and normalization function:

$$tf(f_{i_k}) = f_{i_k}$$

$$idf(N, n_k) = log_2 \frac{N}{n_k}$$

$$g(G, D_i) = \sqrt{\sum_{k=1}^{t} w_{i_k}^2}$$

where $D_i = \{w_{i_1}, w_{i_2}, ..., w_{i_t}\}$ represents a document vector with $t$ unique terms in $D_i$ and $w_{i_k}$ and $f_{i_k}$ are weight and frequency of each term $T_k$ in $D_i$, $N$ is the number of documents, $n_k$ is the number of documents containing term $T_k$, and $g(G, D_i)$ describes a cosine document length normalization factor expressed as a function of some global collection statistics $G$, and document $D_i$.

### 4.2. LSI

Latent semantic indexing (Deerwester et al., 1990) performs singular value decomposition (SVD) on the weighted term-document matrix which is typically composed of BOW or *tf-idf* vectors. The computed SVD is truncated which has the effect of retaining only the most important semantic information while the noise and other artifacts are reduced. The result of LSI is a dense matrix where each document is represented with a fixed dimensional numeric vector (with a few hundred dimensions). In this respect, LSI is similar to neural embedding methods although the elements of the resulting vectors have a very different meaning.

We used the LSI implementation available in Gensim to transform the vectors from the *tf-idf* representation. We tested number of dimensions set to 100, 300, and 500.

### 4.3. Doc2Vec

Le and Mikolov (2014b) introduced two variants of Doc2Vec approach for dense document representations, called PV-DM (Paragraph Vector-Distributed Memory) and PV-CBOW (Paragraph Vector-Continuous Bag-of-Words). In the PV-DM model, a training context is defined with a sliding window moving over the text. The training task is defined as a prediction of the central word of the context window with a shallow neural network. An input to this neural network are embeddings for all the other words from the context together with the embedding of the whole document. During training, the network learns both the word embeddings and the embedding for the whole document. The simpler PV-CBOW model does not have the access to the context words from the document. The neural network predicts randomly sampled words from the document given only the embedding of the whole document.

Both Doc2Vec models have a number of tuneable hyperparameters that can significantly impact the performance of the model. In order to determine those parameters, we used the Bayesian optimisation. This optimisation strategy optimizes black-box functions by placing a prior belief about the function, and updating it with each evaluation. The parameters of the function for the next evaluation are selected based on a predefined criterion that takes into account previous evaluations of the function. We optimized the evaluation metric MAP@10 (introduced in Section 5.) on related articles contained in the training set. To guide the search for parameters, we used a Gaussian Process (GP) prior and Expected Improvement (EI) criterion

[4]The SMART (System for the Mechanical Analysis and Retrieval of Text) Information Retrieval System defines notation for term weighting and normalization where different formulas are allowed for computing term frequency, document frequency and document length normalization.

that maximize the evaluation metric. After 120 evaluations, we selected the best performing hyperparameters. The size of the resulting vector was set to 180 dimensions, the context window covered 5 words left and right from the central word, the vocabulary size was set to 36,000 words, and we ignored words with with less than 35 occurences. The training procedure used negative sampling with 30 negative words and a threshold for downsampling of frequent words set to $3.7 \times 10^{-4}$. In Section 6., we report only results using those hyperparameters.

We used the Doc2Vec (PV-DM and PV-CBOW) implementation available in Gensim library and the Bayesian optimisation from the Scikit-Optimize[5] library.

### 4.4. mBERT

Multilingual BERT (mBERT) is a 12 layer Transformer model (Vaswani et al., 2017) proposed by Devlin et al. (2019). The mBERT was simultaneously trained on Wikipedia pages of 104 languages using masked language modelling (MLM) and next sentence prediction (NSP) objectives. All 104 languages for the mBERT model use shared word piece vocabulary without an explicit way to denote different languages. Maximum length of the input sequence for the model is 512 tokens and each token is represented with 768 dimensions. News articles that require more tokens than the maximum length are trimmed and only the first 512 tokens are kept. The input to BERT begins with *[CLS]* token and ends with a *[SEP]* token denoting the end of a sequence. Additionally, model receives an attention mask to avoid performing attention on padding token indices in case of shorter articles. Running the model produces a context dependant representations of the input tokens. The whole input sequence can be represented with a single vector by using the context dependant representation of the *[CLS]* token. In Devlin et al. (2019), this representation is used as an aggregate sequence representation for classification tasks. Another way to represent the whole sequence, used in Reimers and Gurevych (2019), is to take an average of all input tokens. The BERT model allows for an input to contain a pair of sequences and it is possible to finetune the model with a classification objective that will classify the news article as related or not. At the inference time, in order to find related articles for a newly written article, this approach would require running the model for each article from the archive paired with the new article. Due to a very large number of articles, this is not feasible - determining which articles are related would take too much time and prolong the publication of the new article. Instead, we finetune the model using the Siamese BERT-Networks, a network architecture proposed by Reimers and Gurevych (2019) that avoids this problem. We hold out 10% of the training data and use it to evaluate the finetuning progress. Once the loss on the evaluation data starts increasing, the training is stopped. We use Adam optimizer with a learning rate of $1 \times 10^{-5}$ with linear warmup on 10% of the training data and a batch of 16 examples. The objective that is minimized is the triplet margin loss:

$$L(a, p, n) = \max\{\|\mathbf{a} - \mathbf{p}\|_2 - \|\mathbf{a} - \mathbf{n}\|_2 + \textit{margin}, 0\}$$

In our experiments, the margin for the loss function was set to 1. Anchor and positive example of the triplet objective is the pair of related articles, while a negative example is a related article of a randomly chosen article. We evaluate the model that is finetuned using the averaged token representation, as well as the one finetuned on the initial *[CLS]* token.

The pretrained model used for finetuning was the *bert-base-multilingual-cased* variant available in the Transformers (Wolf et al., 2019) library.

### 4.5. XLM-RoBERTa

The XLM-RoBERTa (XLM-R) (Conneau et al., 2019) is a large multilingual BERT-like model based on RoBERTa (Liu et al., 2019). It uses the sentence piece tokenizer and it is trained with the masked language model objective (MLM) on the CommonCrawl data in 100 languages, including Croatian. Similar to the mBERT, all languages share the same vocabulary (a larger one than mBERT) and the model does not need an explicit marker to denote the language of the input. The maximum size of the input is 512 tokens and each token is represented by a vector with 1024 dimension. Long news articles that do not fit within 512 tokens are trimmed and only the first 512 tokens are kept. All tokenized sequences begin with the '$<s>$' token that denotes a beginning of the sequence and end with the '$<\backslash s>$' token. The initial token can be used as a whole sequence representation in the similar way to the *[CLS]* token of the mBERT model. This model also accepts a pair of sequences as input, and the same caveats apply as with the mBERT model. Finetuning uses the same Siamese BERT-Networks architecture with the Adam optimizer, $4 \times 10^{-5}$ learning rate with a linear warmup on 10% of the data and batches of 16 examples. A triplet margin objective is optimized on the related articles with the negative sampling as described for the mBERT model. We evaluate the model that is finetuned on the averaged token representation, as well as the one finetuned on the initial token.

The pretrained model used for finetuning is the *xlm-roberta-large* variant of the model available in the Transformers library.

## 5. Evaluation setting

As described in Section 3., the evaluation data consists of 25% of the latest articles from the whole dataset. This mimics the realistic scenario, where all older articles are considered when a new article is published. When considering eligible articles in our document retrieval task, we considered their age stored in the *published_date* attribute. This is consistent with the real world scenario where a journalist must not link older but unpublished articles in order to avoid dead links.

The algorithms were trained on the older 75% of articles. The *tf-idf* model thus discarded any newly introduced tokens (words) and used *idf* estimates from the training data when computing *tf-idf* vectors. The same *tf-idf* model was used for the LSI model. All embeddings-based models used trained models to infer vectors of the training data. We used cosine similarity for all document retrieval operations[6].

[5] `https://scikit-optimize.github.io`

[6] Note that the cosine similarity might not be the best choice

The performance of all algorithms was assessed using the mean average precision (MAP) score. The evaluation score for a single article query returns all other articles sorted by decreasing similarity from the query. The score for this query is the average precision of elements of the ranked result list. Evaluation over all articles produces a list of such average precisions. The final score for the model is a mean across the average precisions. In our experiments, we considered only the top ten retrieved documents when computing the MAP score, i.e. MAP@10. A model that would return articles at random (i.e. the random baseline model) would have a 0.0 score in MAP@10.

## 6. Results

The performance of compared document linking methods is presented in Table 1. The results are surprising, as the best result is achieved with the simple *tf-idf* model. This is disappointing concerning the state-of-the-art neural embeddings. However, we evaluated different representations using the links between articles selected by the journalists. This does not necessarily mean that the retrieved articles are not good recommendations, but verifying this hypothesis would require human evaluation.

The *tf-idf* representation consistently performs best across all evaluated hyperparameters of the model, which do not significantly influence the score. Nevertheless, results in Table 1 suggest that including rare words ($m = 2$) improves the performance, while excluding frequent and rare words ($M = 25\%, m > 2$) decreases the performance. While the difference is small it suggests that *tf-idf* models trained on this domain may benefit from preprocessing settings that do not remove what is typically considered as artifacts or noise. From a journalist's perspective this corresponds to linking articles based on a few rare keywords. This may also offer an explanation why LSI does not achieve scores comparable to *tf-idf*. Since LSI is designed to retain only the most important semantic information, the rare words which could improve the models in this particular domain are filtered out.

Both Doc2Vec representations perform similarly are much better then LSI. They approach the performance of *tf-idf* but cannot match it. LSI is less successful with the reduction of dimensions but is competitive with much larger mBERT and XLM-R embeddings that use 768 and 1024 dimensions to represent a document. For both mBERT and XLM-R, using an average of contextual token embeddings shows better results than using the result of only the *[CLS]* token, which is consistent with the conclusions of Reimers and Gurevych (2019).

## 7. Conclusions and future work

We evaluated several document representations to recommend related news articles. The results show that the *tf-idf* representation produces results more consistent with manual selection of journalists compared to more sophisticated representations. We do not yet have a definite explanation for this result and plan to explore it in the future work. Our current belief is that journalists use a keyword search to locate related articles. The articles found in this way contain the same words as the query and bias the evaluation in favor of the *tf-idf* method.

It is possible that the dataset contains a significant amount of noise and that better results cannot be achieved. For example, one possible source of noise would be journalists that misuse the related article information by including their own articles as related, even when they are not, in order to increase the view count of their own articles. In future work, we plan to perform a manual evaluation of the returned related articles on a selected subset.

Although mBERT and XLM-R models achieve state-of-the-art results in many NLP tasks, they did not fare well in this evaluation. Another reason for this might be that document representations created by those models are not suitable for comparison with the cosine similarity. Reimers and Gurevych (2019) offer this explanation when evaluating BERT representations and we plan to explore the impact of similarity measures on the performance of mBERT and XLM-R models.

| Model | MAP@10 |
|---|---|
| tf-idf (m=5, M=50%)[7] | 0.279 |
| tf-idf (m=2, M=50%) | **0.281** |
| tf-idf (m=2, M=25%) | **0.281** |
| tf-idf (m=10, M=50%) | 0.277 |
| tf-idf (m=10, M=25%) | 0.277 |
| LSI (d=500) | 0.186 |
| LSI (d=300) | 0.166 |
| LSI (d=100) | 0.124 |
| Doc2Vec (PV-DM) | 0.248 |
| Doc2Vec (PV-CBOW) | 0.240 |
| mBERT (AVG) | 0.142 |
| mBERT (CLS) | 0.094 |
| XLM-R (AVG) | 0.246 |
| XLM-R (CLS) | 0.186 |

Table 1: The performance of different approaches on the task of similar news article retrieval.

**Acknowledgements**

The research was supported by the Slovene Research Agency through research core funding no. P6-0411 and P2-103. This paper is supported by European Union's Horizon 2020 Programme project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media, grant no. 825153). The results of this publication reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

## 8. References

Agra Bimantara, Michelle Blau, Kevin Engelhardt, Johannes Gerwert, Tobias Gottschalk, Philipp Lukosz, Shenna Piri, Nima Saken Shaft, and Klaus Berberich. 2018. htw saar @ TREC 2018 news track. In *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018*.

for some embeddings models.

[7]This is the default setting for filtering extremes from the dictionary in Gensim.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, F. Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *ArXiv*, abs/1911.02116.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Marwa Essam and Tamer Elsayed. 2019. bigIR at TREC 2019: Graph-based analysis for news background linking. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019*.

John Foley, Ananda Montoly, and Mayeline Pena. 2019. Smith at TREC2019: learning to rank background articles with poetry categories and keyphrase extraction. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019*.

Michael Gamon, Sumit Basu, Dmitriy Belenko, Danyel Fisher, Matthew Hurst, and Arnd Christian König. 2008. Blews: Using blogs to provide context for news articles. In *ICWSM*.

Weiwei Guo, Hao Li, Heng Ji, and Mona T. Diab. 2013. Linking tweets to news: A framework to enrich short text data in social media. In *ACL*.

Shudong Huang, Ian Soboroff, and Donna Harman. 2018. TREC 2018 news track. In *Proceedings of the Second International Workshop on Recent Trends in News Information Retrieval co-located with 40th European Conference on Information Retrieval (ECIR 2018)*, pages 57–59.

Daisuke Ikeda, Toshiaki Fujiki, and Manabu Okumura. 2006. Automatically linking news articles to blog entries. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Matjaz Juršič, Igor Mozetič, Tomaž Erjavec, and Nada Lavrač. 2010. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *J. UCS*, 16(9):1190–1214.

Quoc Le and Tomas Mikolov. 2014a. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196.

Quoc Le and Tomas Mikolov. 2014b. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

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*.

Kuang Lu and Hui Fang. 2018. Paragraph as lead - finding background documents for news articles. In *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018*.

Kuang Lu and Hui Fang. 2019. Leveraging entities in background document retrieval for news articles. In *Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP/IJCNLP*.

Pinky Sitikhu, Kritish Pahi, Pujan Thapa, and Subarna Shakya. 2019. A comparison of semantic similarity methods for maximum human interpretability. *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, 1:1–4.

Ian Soboroff, Shudong Huang, and Donna Harman. 2018. TREC 2018 news track overview. In *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018*.

Anastasia Soloshenko, Yulia Orlova, Vladimir Rozaliev, and A.V. Zaboleeva-Zotova. 2015. Establishing semantic similarity of the cluster documents and extracting key entities in the problem of the semantic analysis of news texts. *Modern Applied Science*, 9, 04.

Dwane van der Sluis, Dyaa Albakour, and Miguel Martinez. 2018. Signal at TREC 2018 news track. In *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

# Cross-lingual Transfer of Twitter Sentiment Models Using a Common Vector Space

**Marko Robnik-Šikonja***, **Kristjan Reba***, **Igor Mozetič**†

* University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, SI-1000 Ljubljana, Slovenia
marko.robnik@fri.uni-lj.si kr3377@student.uni-lj.si

† Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana
igor.mozetic@ijs.si

**Abstract**

Word embeddings represent words in a numeric space in such a way that semantic relations between words are encoded as distances and directions in the vector space. Cross-lingual word embeddings map the vector space of one language to the vector space of another language, or vector spaces of multiple languages to the joint vector space where similar words are aligned. Cross-lingual embeddings can be used to transfer machine learning models between languages, thereby compensating for insufficient data in less-resourced languages. We use cross-lingual word embeddings to transfer machine learning prediction models for Twitter sentiment between 13 languages. We focus on two transfer mechanisms using the joint numerical space for many languages as implemented in the LASER library: the transfer of trained models and the expansion of training sets with instances from other languages. Our experiments show that the transfer of models between similar languages is sensible, while dataset expansion did not increase the predictive performance.

## 1. Introduction

Word embeddings are representations of words in numerical form, as vectors of typically several hundred dimensions. The vectors are used as input to machine learning models; for complex language processing tasks, these generally are deep neural networks. The embedding vectors are obtained from specialized neural network-based embedding algorithms, e.g., word2vec (Mikolov et al., 2013), or fastText (Bojanowski et al., 2017). Modern word embedding spaces exhibit similar structures across languages, even when considering distant language pairs like English and Vietnamese (Mikolov et al., 2013). This means that embeddings independently produced from monolingual text resources can be aligned, resulting in a common cross-lingual representation, called cross-lingual embeddings, which allows for fast and effective integration of information in different languages.

There exist several approaches to cross-lingual embeddings. The first group of approaches uses monolingual embeddings with the optional help from a bilingual dictionary to align the pairs of embeddings (Artetxe et al., 2018a). The second group of approaches uses bilingually aligned (comparable or even parallel) corpora to construct joint embeddings in all the involved languages (Artetxe and Schwenk, 2019). The third type of approach is based on large pretrained multilingual masked language models such as BERT (Devlin et al., 2019). In this work, we focus on the second group of approaches, i.e. a joint sentence representation for many languages (Artetxe and Schwenk, 2019) as implemented in the LASER library[1] and available for 93 languages.

Sentiment annotation is a costly and lengthy operation, with relatively low inter-annotator agreement (Mozetič et al., 2016). Large annotated sentiment datasets are therefore rare, especially for low-resourced languages. The transfer of already trained models or datasets from other languages would be useful and would increase the ability to study sentiment-related phenomena for many more languages than possible today.

Using a collection of 13 large Twitter sentiment datasets, annotated in the same manner, we study two modes of cross-lingual transfer based on projections of sentences into a joint vector space. The first approach transfers trained models from source to target languages, where the model is trained on source language(s), and used for classification in the target language(s). This model transfer is possible because texts in all involved languages are embedded into the common vector space. The second approach expands the training set with instances from other languages, and then all instances are mapped into the common vector space during neural network training. Additionally, we analyse the quality of representations for the Twitter sentiment classification and compare the common vector space for several languages constructed by LASER library, multilingual BERT, and traditional bag-of-words approach. The results show a relatively low decrease in predictive performance when transferring trained sentiment prediction models between languages.

The paper is divided into four sections. In Section 2, we present background on different types of cross-lingual embeddings: alignment of monolingual embeddings, building a fixed common vector space for several languages, and multilingual contextual models. In Section 3, we present a large collection of tweets from 13 languages used in our empirical evaluation, the implementation details of our deep neural network prediction models, and the evaluation metrics used. Section 4 contains four series of experiments. We first analyse the transfer of trained models between lan-

[1] https://github.com/facebookresearch/LASER

guages from the same language group and from a different language group, followed by the expansion of datasets with instances from other languages. We end the experimental part with the evaluation of representation spaces and compare the common vector space with the multilingual BERT model. In Section 5, we summarise the results, draw the conclusions, and present ideas for further work.

## 2. Background

Word embeddings represent each word in a language as a vector in a high dimensional vector space so that the relations between words in a language are reflected in their corresponding embeddings. Cross-lingual embeddings attempt to map words represented as vectors from one vector space to the other so that the vectors representing words with the same meaning in both languages are as close as possible. Søgaard et al. (2019) present a detailed overview and classification of cross-lingual methods.

Cross-lingual approaches can be sorted into three groups, described in the following three subsections. The first group of methods uses monolingual embeddings with (an optional) help from bilingual dictionaries to align the embeddings. The second group of approaches uses bilingually aligned (comparable or even parallel) corpora for joint construction of embeddings in all involved languages. The third type of approach is based on large pretrained multilingual masked language models such as BERT (Devlin et al., 2019). The multilingual BERT is typically used as a starting model which is fine-tuned for a particular task, without explicitly extracting embedding vectors.

### 2.1. Alignment of monolingual embeddings

Cross-lingual alignment methods take precomputed word embeddings for each language and align them with optional use of bilingual dictionaries. Two types of monolingual embedding alignment methods exist. The first type of approaches map vectors representing words in one of the languages into the vector space of the other language (and vice-versa). The second type of approaches maps embeddings from both languages into a common vector space. The goal of both types of alignments is the same: the embeddings for words with the same meaning must be as close as possible in the final vector space. A comprehensive summary of existing approaches can be found in (Artetxe et al., 2018a). The open source implementation of the method described in (Artetxe et al., 2018a), named *vecmap*[2], is able to align monolingual embeddings either using supervised, semi-supervised, or unsupervised approach.

The supervised approach requires the use of a bilingual dictionary, which is used to match embeddings of equivalent words. The embeddings are aligned using the Moore-Penrose pseudo-inverse, which minimizes the sum of squared Euclidean distances. The algorithm always converges but can be caught in a local maximum when the initial solution is poor. To overcome this, several methods (stochastic dictionary introduction, frequency-based vocabulary cutoff, etc.) are used that help the algorithm to climb out of local maxima. A more detailed description of the algorithm is given in (Artetxe et al., 2018b).

[2]https://github.com/artetxem/vecmap

The semi-supervised approach uses a small initial seeding dictionary, while the unsupervised approach is run without any bilingual information. The latter uses similarity matrices of both embeddings to build an initial dictionary. This initial dictionary is usually of low but sufficient quality for later processing. After the initial dictionary (either by seeding dictionary or using similarity matrices) is built, an iterative algorithm is applied. The algorithm first computes optimal mapping using the pseudo-inverse approach for the given initial dictionary. Then the optimal dictionary for the given embeddings is computed, and the procedure is repeated with the new dictionary.

When constructing mappings between embedding spaces, a bilingual dictionary can be helpful as its entries can be used as anchors for the alignment map for supervised and semi-supervised approaches. However, lately, researchers have proposed methods that do not require the use of a bilingual dictionary, but rely on adversarial approach (Conneau et al., 2018) or use the frequencies of the words (Artetxe et al., 2018b) to find a required transformation. These are called unsupervised approaches.

### 2.2. Projecting into a common vector space

To construct a common vector space for all involved languages, one requires a large aligned bilingual or multilingual parallel corpus. The constructed embeddings must map the same words in different languages as close as possible in the common vector space. The availability and quality of alignments in training set corpus may present an obstacle. While Wikipedia, subtitles, and translation memories are good sources of aligned texts for large languages, less-resourced languages are not well-presented and building embeddings for such languages is a challenge.

LASER (Language-Agnostic SEntence Representations) is a Facebook research project focusing on joint sentence representation for many languages (Artetxe and Schwenk, 2019). Similarly to machine translation architectures, it uses an encoder-decoder architecture. The encoder is trained on a large parallel corpus, translating a sentence in any language or script to a parallel sentence in either English or Spanish (whichever exists in the parallel corpus), thereby forming a joint representation of entire sentences in many languages in a shared vector space. The project focused on scaling to a large number of languages; currently, the encoder supports 93 different languages. Using LASER, one can train a classifier on data from just one language and use it on any language supported by LASER. A vector representation in the joint embedding space can be transformed back into a sentence using a decoder for the specific language.

### 2.3. Multilingual BERT

BERT (Bidirectional Encoder Representations from Transformers) embedding (Devlin et al., 2019) generalises the idea of language model (LM) to masked language models, inspired by the cloze test, which tests understanding of a text by removing a certain portion of words, which the participant is asked to replace. The masked language model randomly masks some of the tokens from the input, and the task of LM is to predict the missing token

based on its neighbourhood. BERT uses transformer neural networks (Vaswani et al., 2017) in a bidirectional sense and further introduces the task of predicting whether two sentences appear in a sequence. The input representation of BERT are sequences of tokens representing sub-word units. The input is constructed by summing the embeddings of corresponding tokens, segments, and positions. Some widespread words are kept as single tokens, others are split into sub-words (e.g., frequent stems, prefixes, suffixes—if needed down to single letter tokens). The original BERT project offers pre-trained English, Chinese and multilingual model. The latter, called mBERT, is trained on 104 languages simultaneously.

To use BERT in classification tasks only requires adding connections between its last hidden layer and new neurons corresponding to the number of classes in the intended task. The fine-tuning process is applied to the whole network, and all of the parameters of BERT and new class-specific weights are fine-tuned jointly to maximise the log-probability of the correct labels.

## 3. Datasets and experimental settings

In this section, we present the evaluation metrics, experimental data and implementation details of the used neural prediction models.

### 3.1. Evaluation metrics

Following Mozetič et al. (2016) we report $\overline{F_1}$ score which takes positive and negative sentiment into account, and classification accuracy $CA$. $F_1(c)$ score for class value $c$ is the harmonic mean of precision $p$ and recall $r$ for the given class $c$, where the precision is defined as the proportion of correctly classified instances from the instances predicted to be from the class $c$, and the recall is the proportion of correctly classified instances actually from the class $c$.

$$F_1(c) = \frac{2p_c r_c}{p_c + r_c}.$$

The $F_1$ score returns values from $[0, 1]$ interval, where 1 means perfect classification and 0 completely wrong predictions. We use $F_1$ score averaged over positive $(+)$ and negative $(-)$ sentiment class:

$$\overline{F_1} = \frac{F_1(+) + F_1(-)}{2}.$$

As the sentiment labels are ordered, the neutral sentiment label is implicitly taken into account in $\overline{F_1}$.

The classification accuracy CA is defined as the ratio of correctly predicted tweets $N_c$ to all the tweets $N$:

$$CA = \frac{N_c}{N}$$

### 3.2. Datasets

We use a corpus of Twitter sentiment datasets (Mozetič et al., 2016), consisting of 15 languages, with over 1.6 million annotated tweets. The languages covered are Albanian, Bosnian, Bulgarian, Croatian, English, German, Hungarian, Polish, Portuguese, Russian, Serbian, Slovak, Slovene, Spanish, and Swedish. Authors studied the annotators' agreement on the labelled tweets. They discovered that for some languages (English, Russian, Slovak) the SVM classifier achieves significantly lower score than the annotators. This hints that for these languages, there might be a room for improvement using better classification model or larger training set.

We cleaned the above datasets by removing the duplicated tweets, weblinks, and hashtags. Due to the low quality of sentiment annotations indicated by low self-agreement and low inter-annotator agreement, we removed Albanian and Spanish datasets. For these two languages, the self-agreement expressed with $\overline{F_1}$ score (i.e. $F_1(c)$ is the fraction of equally labelled tweets out of all the tweets with a given label $c$) is 0.60 and 0.49, respectively; the inter-annotator agreement is 0.41 and 0.42. The characteristics of the remaining 13 datasets are presented in Table 1.

### 3.3. Implementation details

In our experiments, we use two different types of prediction models, BiLSTM neural networks using joint vector space embeddings constructed with the LASER library, and multilingual BERT. The multilingual BERT model is case sensitive (i.e. bert_multi_cased), pretrained on 104 languages, has 12 transformer layers, and 110 million parameters. We fine-tune only the last layer of the network, using the batch size of 32, and 3 epochs.

The cross-lingual embeddings from LASER library are pretrained on 93 languages, using BiLSTM networks, and are stored as 1024 dimensional embedding vectors. Our classification models contain the embedding layer, followed by multilayer perceptron hidden layer of size 8, and an output layer with three neurons (corresponding to three output classes, negative, neutral, and positive sentiment) using the softmax. We use ReLU activation function and Adam optimizer. The fine-tuning uses a batch size of 32 and 10 epochs.

Further technical details are available in the freely available source code.

## 4. Experiments and results

Our experimental evaluation focuses on text representation using embeddings into a common vector space with the LASER library. We conducted several experiments reported below: transfer of models between languages from the same and different language family, expansion of training sets with varying amounts of data from other languages, and comparison of the joint space embeddings with the multilingual BERT. We did not systematically test all possible combinations of languages and language groups as this would require too much computational time and results would not fit into the paper. Instead, we arbitrary selected a few language combinations in advance. We leave a more systematic approach based on informative features (Lin et al., 2019) for further work.

### 4.1. Transfer to the same language family

We first test the transfer of prediction models between similar languages from the same language family. The transfer between similar languages is the most likely to be successful. As the source and target languages, we tried

| | Number of tweets | | | | Agreement | |
|---|---|---|---|---|---|---|
| Language | Negative | Neutral | Positive | All | Self- | Inter- |
| Bosnian | 12,868 | 11,526 | 13,711 | 38,105 | 0.81 | 0.51 |
| Bulgarian | 15,140 | 31,214 | 20,815 | 67,169 | 0.77 | 0.50 |
| Croatian | 21,068 | 19,039 | 43,894 | 84,001 | 0.81 | 0.51 |
| English | 26,674 | 46,972 | 29,388 | 103,034 | 0.79 | 0.67 |
| German | 20,617 | 60,061 | 28,452 | 109,130 | 0.73 | 0.42 |
| Hungarian | 10,770 | 22,359 | 35,376 | 68,505 | 0.76 | - |
| Polish | 67,083 | 60,486 | 96,005 | 223,574 | 0.84 | 0.67 |
| Portuguese | 58,592 | 53,820 | 44,981 | 157,393 | 0.74 | - |
| Russian | 34,252 | 44,044 | 29,477 | 107,773 | 0.82 | - |
| Serbian | 24,860 | 30,700 | 16,161 | 71,721 | 0.81 | 0.51 |
| Slovak | 18,716 | 14,917 | 36,792 | 70,425 | 0.77 | - |
| Slovene | 38,975 | 60,679 | 34,281 | 133,935 | 0.73 | 0.54 |
| Sweedish | 25,319 | 17,857 | 15,371 | 58,547 | 0.76 | - |

Table 1: The left-hand side reports the number of tweets from each of the category and the overall number of instances for individual languages. The right-hand side contains self-agreement of annotators, and inter-annotator agreement for languages where more than one annotator was involved.

| | | Transfer | | Both target | |
|---|---|---|---|---|---|
| Source | Target | $\overline{F_1}$ | CA | $\overline{F_1}$ | CA |
| German | English | 0.55 | 0.59 | 0.62 | 0.65 |
| Polish | Russian | 0.64 | 0.59 | 0.70 | 0.70 |
| Polish | Slovak | 0.63 | 0.59 | 0.72 | 0.72 |
| German | Swedish | 0.58 | 0.57 | 0.67 | 0.65 |
| German Swedish | English | 0.58 | 0.60 | 0.62 | 0.65 |
| Slovene Serbian | Russian | 0.53 | 0.55 | 0.70 | 0.70 |
| Slovene Serbian | Slovak | 0.59 | 0.52 | 0.72 | 0.72 |
| Average performance gap | | 0.09 | 0.11 | | |

Table 2: The transfer of trained models between languages from the same language family using common vector space (left-hand side) and comparison with both training and testing set from the target language (on the right-hand side).

| | | Transfer | | Both target | |
|---|---|---|---|---|---|
| Source | Target | $\overline{F_1}$ | CA | $\overline{F_1}$ | CA |
| Russian | English | 0.52 | 0.56 | 0.62 | 0.65 |
| English | Russian | 0.57 | 0.58 | 0.70 | 0.70 |
| English | Slovak | 0.46 | 0.44 | 0.72 | 0.72 |
| Polish, Slovene | English | 0.58 | 0.57 | 0.62 | 0.65 |
| German, Swedish | Russian | 0.61 | 0.61 | 0.70 | 0.70 |
| English, German | Slovak | 0.50 | 0.47 | 0.72 | 0.72 |
| Average performance gap | | 0.14 | 0.15 | | |

Table 3: The transfer of trained models between languages from different language groups using the common vector space representation (left-hand side), and comparison with both training and testing set from the target language (on the right-hand side).

several combinations of Slavic and Germanic languages. We report the results in Table 2.

In each experiment, we use the entire dataset(s) of the source language as the training set, and the whole dataset of the target language as the testing set. We compare the results with the training and testing set from the target language, where 70% of the dataset is used for training and 30% for testing. The latter results can be taken as an upper bound of what the transfer models could achieve in an ideal condition.

The results from Table 2 show that there is a gap between transfer learning models and native models from 4% to 20% (on average the decrease in performance for transfer learning is 9.3% for $\overline{F_1}$ and 11.1% for CA). For the direct transfer of models without additional target data, these results are encouraging.

### 4.2. Transfer to different language family

We repeat the experiments we did for languages from the same language family on languages from different language families. The transfer is less likely to be successful in this case, and we expect a lower performance in these unfavourable conditions.

The results from Table 3 show that there is a gap between transferred models and native models from 4% to 28% (on average the decrease of performance for transfer learning is 14% for $\overline{F_1}$ and 15.2% for CA). This gap is significant and makes the resulting transferred models less useful in the target languages. Another observation is that the differences between target languages are significant. It seems that the transfer to Slovak is much less successful than to Russian, while English is in between the two.

### 4.3. Increasing datasets with several languages

We test possible improvements in prediction performance if we increase the training sets with instances from several related and unrelated languages. The training set in each experiment consists of instances from several languages projected into the common vector space and also 70% of the target language dataset. The remaining 30% of target language instances are used as the testing set. As the text representation, we use projection into the common vector space computed with the LASER library.

The results from Table 4 show a gap between learning models using the expanded datasets and native models (the decrease for expanded models is from 2% to 7%, on average 3% for $F_1$ and 5.7% for CA). These results indicate that the tested expansion of datasets was unsuccessful, i.e. the provided amount of instances from the target language was

| | | Expanded | | Only target | |
|---|---|---|---|---|---|
| Source | Target | $\overline{F_1}$ | CA | $\overline{F_1}$ | CA |
| English, Croatian, Slovene | Slovene | 0.58 | 0.53 | 0.60 | 0.60 |
| English, Croatian, Serbian, Hungarian, Slovak | Slovak | 0.67 | 0.65 | 0.72 | 0.72 |
| English, Croatian, Russian | Russian | 0.67 | 0.65 | 0.70 | 0.70 |
| Russian, Swedish, English | English | 0.60 | 0.61 | 0.62 | 0.65 |
| Average improvement | | -0.03 | -0.06 | | |

Table 4: The expansion of training sets with instances from several languages projected into the common vector space using the LASER library (left-hand side) and comparison with training and testing set from the same language (on the right-hand side).

| | All other & Target | | Only Target | |
|---|---|---|---|---|
| Target | $\overline{F_1}$ | CA | $\overline{F_1}$ | CA |
| Bosnian | 0.64 | 0.59 | 0.67 | 0.64 |
| Bulgarian | 0.54 | 0.56 | 0.50 | 0.59 |
| Croatian | 0.63 | 0.57 | 0.73 | 0.68 |
| English | 0.58 | 0.60 | 0.62 | 0.65 |
| German | 0.52 | 0.59 | 0.53 | 0.65 |
| Hungarian | 0.59 | 0.61 | 0.60 | 0.67 |
| Polish | 0.67 | 0.63 | 0.70 | 0.66 |
| Portugal | 0.44 | 0.39 | 0.52 | 0.51 |
| Russian | 0.66 | 0.64 | 0.70 | 0.70 |
| Serbian | 0.52 | 0.49 | 0.48 | 0.54 |
| Slovak | 0.64 | 0.61 | 0.72 | 0.72 |
| Slovene | 0.54 | 0.50 | 0.60 | 0.60 |
| Swedish | 0.63 | 0.59 | 0.67 | 0.65 |
| Average improvement | -0.04 | -0.07 | | |

Table 5: The expansion of training sets with instances from all other languages mapped into the common vector space using the LASER library (left-hand side) and comparison with training and testing set from the same language (on the right-hand side).

already sufficient for successful learning. The additional instances from other languages are likely to be of lower quality than the native instances and therefore decrease the performance.

To test an even more extensive expansion of the training sets, we trained models on all other languages and 70% of the target language, while testing them on the remaining 30% of the target language. The results are presented in Table 5.

The results show that using many languages and significant enlargement of datasets can be successful. For Bulgarian and Serbian training on many languages gives higher $\overline{F_1}$ score (but not CA) than training only on the target language. For all other languages, the tried expansions of training sets are unsuccessful, and the difference to native models is on average 3.5% for $\overline{F_1}$ score and 6.8% for CA.

### 4.4. Comparing embeddings

In our final experiment, we compare embeddings into a common vector space obtained with LASER library with the multilingual BERT. Note that in this experiment, there is no transfer between different languages but only a test of the quality of the representation, i.e. embeddings. The training set in each experiment consists of randomly chosen 70% of the dataset for each language, while the remaining 30% of instances are used as the testing set. As a baseline, we report the results of the SVM model without neural embeddings that uses Delta TF-IDF weighted bag-of-ngrams representation with substantial preprocessing of tweets (Mozetič et al., 2016). These results are not entirely comparable with our setting as they were obtained with 10-fold stratified blocked cross-validation, while we use a single 70:30 split. Further, the datasets for Bosnian, Croatian, and Serbian language were merged in (Mozetič et al., 2016) due to the similarity of these languages; therefore, we report the performance on the merged dataset for the SVM classifier. Results are presented in Table 6.

| | LASER | | mBERT | | SVM | |
|---|---|---|---|---|---|---|
| Language | $\overline{F_1}$ | CA | $\overline{F_1}$ | CA | $\overline{F_1}$ | CA |
| Bosnian | **0.67** | 0.64 | 0.65 | **0.66** | 0.61 | 0.56 |
| Bulgarian | 0.50 | 0.59 | **0.58** | **0.60** | 0.52 | 0.54 |
| Croatian | **0.73** | **0.68** | 0.64 | **0.68** | 0.61 | 0.56 |
| English | 0.62 | 0.65 | **0.72** | **0.71** | 0.63 | 0.64 |
| German | 0.53 | 0.65 | **0.66** | **0.66** | 0.54 | 0.61 |
| Hungarian | 0.60 | 0.67 | **0.65** | **0.69** | 0.64 | 0.67 |
| Polish | **0.70** | 0.66 | **0.70** | **0.73** | 0.68 | 0.63 |
| Portugal | 0.52 | 0.51 | **0.66** | **0.67** | 0.55 | 0.51 |
| Russian | 0.70 | 0.70 | **0.74** | **0.75** | 0.61 | 0.60 |
| Serbian | 0.48 | 0.54 | 0.56 | 0.54 | **0.61** | **0.56** |
| Slovak | **0.72** | 0.72 | 0.70 | **0.75** | 0.68 | 0.68 |
| Slovene | 0.60 | 0.60 | **0.66** | **0.64** | 0.55 | 0.54 |
| Swedish | **0.67** | 0.65 | 0.64 | **0.66** | 0.66 | 0.62 |
| Average | 0.62 | 0.64 | **0.66** | **0.67** | 0.61 | 0.59 |

Table 6: Comparison of different representations: supervised mapping into a common vector space with the LASER library, multilingual BERT, and bag-of-ngrams with the SVM classifier. The best score for each language and metric is in bold.

The SVM baseline using bag-of-words representation achieves lower predictive performance than the two neural embedding approaches. We speculate that the main reason is the knowledge about language structure contained in large precomputed embeddings used by the neural approaches. Together with the fact that standard feature-based machine learning approaches require much more preprocessing effort, it seems that there are no good reasons why to bother with this approach in text classification. The multilingual BERT is the best of the three tested methods, achieving the best average $\overline{F_1}$ and CA scores, as well as the best result in most languages (in bold). The downside is that the fine-tuning and execution of mBERT requires much more computational time compared to precomputed fixed embeddings. Nevertheless, with progress in optimization techniques for neural network learning and advent of computationally more efficient BERT variants, e.g., (You et al., 2020), this obstacle might disappear in the future.

## 5. Conclusions

We studied two approaches to the cross-lingual transfer of Twitter sentiment prediction models based on mappings of words into the common vector space: transfer of trained models, and expansion of datasets with instances

from other languages. Our empirical evaluation is based on relatively large datasets of labelled tweets from 13 European languages. As word representations, we used mappings into a common vector space produced by the LASER library. The results show that there is a significant transfer potential using the models trained on similar languages; compared to training and testing on the same language, we get on average 9.3% lower $\overline{F_1}$ score and 11.1% lower CA. Using models trained on languages from different language families produces larger differences (on average 14% for $\overline{F_1}$ and 15.2% for CA). Our attempt to expand training sets with instances from different languages was unsuccessful using either additional instances from a small group of languages or instances from all other languages. Finally, we tested the quality of text representations by comparing cross-lingual joint embedding space of LASER library, multilingual BERT embeddings, and classical bag-of-ngram representation coupled with SVM classifier. The results show that the multilingual BERT is the most successful of the three, followed by the common vector space of LASER library, while bag-of-ngrams is rarely competitive. The source code of our analyses is freely available[3].

In future work, we plan to expand the experiments with other embedding techniques, in particular, the ELMo contextual embeddings (Peters et al., 2018) together with non-isomorphic cross-lingual transformations that could produce better representations in the joint vector spaces.

**Acknowledgements**

The research was supported by the Slovene Research Agency through research core funding no. P6-0411 and P2-103. This paper is supported by European Union's Horizon 2020 Programme project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media, grant no. 825153), and Rights, Equality and Citizenship Programme project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, grant no. 875263). The results of this publication reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

## 6. References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

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.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of International Conference on Learning Representation ICLR 2018*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint 1309.4168*.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual Twitter sentiment classification: The role of human annotators. *PLOS ONE*, 11(5).

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Anders Søgaard, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui. 2019. *Cross-Lingual Word Embeddings*. Morgan & Claypool Publishers.

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.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training BERT in 76 minutes. In *International Conference on Learning Representations, ICLR 2019*.

[3]https://github.com/kristjanreba/cross-lingual-classification-of-tweet-sentiment

# Primerjava slovenskih besednih vektorskih vložitev z vidika spola na analogijah poklicev

**Anka Supej*, Matej Ulčar†, Marko Robnik-Šikonja†, Senja Pollak***

*Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana
a.supej@gmail.com, senja.pollak@ijs.si

†Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Večna pot 113, 1000 Ljubljana
{matej.ulcar, marko.robnik}@fri.uni-lj.si

**Povzetek**

V zadnjih letih je uporaba globokih nevronskih mrež in gostih vektorskih vložitev za predstavitve besedil privedla do vrste odličnih rezultatov na področju računalniškega razumevanja naravnega jezika. Prav tako se je pokazalo, da vektorske vložitve besed pogosto zajemajo pristranosti z vidika spola, rase ipd. Ena izmed metod za ocenjevanje kakovosti vložitev so izračuni analogij. Prispevek se osredotoča na evalvacijo vektorskih vložitev besed v slovenščini z vidika spola. Sestavili smo seznam moških in ženskih ustreznic poklicev (dostopen prek repozitorija CLARIN) in preko analogij ovrednotili spolno pristranost modelov vložitev fastText, word2vec in ELMo z različnimi konfiguracijami in pristopi k računanju analogij.

**Abstract**

In recent years, the use of deep neural networks and dense vector embeddings for text representation have led to excellent results in the field of computational understanding of natural language. It has also been shown that word embeddings often capture gender, racial and other types of bias. One of the methods for assessing the quality of word embeddings are analogies calculations. The article focuses on the evaluation of Slovene word embeddings in terms of gender. We compiled a list of male and female equivalents of occupations and evaluated the gender bias of fastText, word2vec and ELMo embeddings.

## 1. Uvod

Raziskave na stičišču spola in jezika so metodološko različne. Sociolingvistične študije poročajo o načinih, po katerih se uporaba jezika med ženskami in moškimi razlikuje (npr. širše besedišče, milejše izražanje, uporaba tipičnih slovničnih struktur pri ženskah) (Lakoff, 1973; Tannen, 1990; Argamon et al., 2003). Opažanja, da se uporaba jezika med spoloma razlikuje, so navdihnila študije profiliranja avtorjev na besedilih različnih jezikov in tipov besedil (Koolen in van Cranenburgh, 2017; Pardo et al., 2015; Martinc et al., 2017), tudi za slovenščino ((Verhoeven et al., 2017; Škrjanec et al., 2018).

Dimenzija spola v korpusih ni prisotna le kot jezikovna variacija, temveč tudi v obliki večplastne pristranosti, tako v posameznih besedilih kot tudi v večjih korpusih. Sorodne raziskave ugotavljajo:

- da se pristranost odraža kot pomanjkanje omemb žensk: korpusi, ki se pogosto uporabljajo v raziskavah, vsebujejo znatno manj zaimkov ženskega spola (Zhao et al., 2018) ali drugih nanašalnic na ženske (Caldas-Coulhard in Moon, 2010; Baker, 2010);
- da so ženske manj pogosto avtorice ali urednice (Hill in Shaw (2013): le 16% urednic Wikipedije je žensk);
- da korpusi zajemajo spolno stereotipne kolokacije (Pearce, 2008), ki npr. predstavljajo ženske predvsem skozi reproduktivno funkcijo (Gorjanc, 2007) in jih ne povezujejo z (družbeno) močjo (Baker, 2010).

V zadnjih letih je razmah na področju obdelave naravnega jezika povezan predvsem z uporabo globokih nevronskih mrež, ki se uporabljajo tudi za učenje predstavitev besedil v obliki gostih vektorskih vložitev besed. Izkaže se, da tudi vektorske vložitve besed pogosto zajemajo pristranosti z vidika spola, rase ipd. Pristranost se v besednih vložitvah kaže preko semantičnih asociacij in posledične bližine v vektorskem prostoru (Mikolov et al., 2013b). Računsko jo lahko ovrednotimo npr. s kosinusno podobnostjo med vložitvami, ki opisujejo nek širši pojem (npr. spol), ter stereotipnimi koncepti (npr. v Caliskan et al. (2017): asociacija žensk in umetnosti ter moških in znanosti) ali preko izračuna analogij (Bolukbasi et al., 2016), ki predpostavljajo odnos: $\overrightarrow{moški} - \overrightarrow{poklic_M} \approx \overrightarrow{ženska} - \overrightarrow{poklic_Ž}$. Poleg študij, ki so pokazale na pristranost samih vložitev, so različni avtorji pokazali tudi prenos pristranosti v algoritme za različne naloge obdelave naravnega jezika, od strojnega prevajanja (Prates et al., 2020; Vanmassenhove et al., 2018) do študij sentimenta (Kiritchenko in Mohammad, 2018)). Na drugi strani pa nekateri avtorji (Nissim et al., 2019) opozarjajo tudi na pretirano poudarjanje pristranosti pri zasnovi raziskav z analogijami.

Glavni doprinos prispevka je evalvacija slovenskih modelov besednih vektorskih vložitev z vidika spola, ki še ni dovolj raziskano (izjema je npr. analiza slovenskega modela w2v v Supej et al. (2019)). Prispevek se osredotoča na kvantitativno evalvacijo in primerjavo širokega nabora izbora slovenskih vložitev ter različnih pristopov k evalvaciji, s čimer nagovarja predvsem razvijalce jezikovnotehnoloških orodij, ki vložitve uporabljajo. Kljub temu da s tem indirektno problematiziramo pristranost v jeziku ter pokažemo tudi na nekaj stereotipnih povezav, pa je podrobnejša kritična interpretacija izven fokusa tega prispevka.

V prispevku najprej predstavimo sorodna dela (2. razdelek). V 3. razdelku opišemo seznam moških in ženskih poklicev ter uporabljenih besednih vektorskih vložitev. V 4. in 5. razdelku predstavimo metodologijo in rezultate, nato zaključimo z diskusijo in načrti za nadaljnje delo.

## 2. Sorodna dela

Korpusi odražajo jezikovne variacije (vključno z različnimi vrstami pristranosti) v odnosu do družbenih dejavnikov. Orodja procesiranja naravnega jezika, ki se učijo na korpusih, lahko variacije in pristranosti podedujejo: nekatere študije prikažejo, da so orodja procesiranja naravnega jezika uspešnejša, ko tovrstne variacije upoštevajo (Volkova et al., 2013; Hovy, 2015). Študija Hovy (2015) pokaže, da vključitev informacij o starosti in spolu avtorjev izboljša uspešnost treh nalog v petih različnih jezikih. Pristranost v korpusih ima lahko tudi negativne posledice, kar lahko podkrepimo z nekaj primeri.

Pogosto uporabljeni korpusi vsebujejo pristranosti do te mere, da so orodja procesiranja naravnega jezika uspešnejša pri vhodnih podatkih, kjer je besedilo napisala starejša oseba (Hovy in Søgaard, 2015). Študija Garimella et al. (2019) pokaže, da sta oblikoslovni označevalnik in skladenjski razčlenjevalnik uspešna na tekstih, ki so jih napisale ženske, ne glede na to, na katerih podatkih sta bila naučena. Teksti moških avtorjev so bolje razčlenjeni/označeni, v kolikor je v učnih podatkih na voljo dovolj besedil, ki so jih napisali moški. Uspešnost orodij, kot so razčlenjevalniki, za tekste moških avtorjev je torej lahko posledica neuravnoteženosti množice učnih podatkov v prid moškemu avtorstvu. Pristranost v korpusih ima poleg negativnega vpliva na orodja obdelave naravnega jezika (Sun et al., 2019) tudi druge negativne posledice, kot je npr. nepravilna razrešitev koreferenc (Zhao et al., 2018).

Vektorske vložitve besed, prav tako naučene na korpusih, poleg sintaktičnih značilnosti besed ujamejo tudi njihove semantične relacije. To se izraža v geometriji prostora vektorskih vložitev: semantično povezane vložitve so si v vektorskem prostoru bližje in razporejene v podobnih smereh. Zato je z njimi mogoče računati tudi odnose, ki presegajo enostavno sorodnost besed, npr. preko analogij. Npr. odnos *Madrid:Španija* je podoben odnosu *Pariz:Francija* (Mikolov et al., 2013b). Vektorske vložitve besed so naučene na korpusih z različnimi algoritmi in tako kot korpusi vsebujejo pristranosti. Beseda, ki je npr. stereotipno povezana z določenim spolom, bo v vektorskem prostoru tako bližje vektorski vložitvi besed *ženska* ali *moški* (Garg et al. (2018) npr. pokažejo, da je pridevnik *časten* v angleščini bližje besedi *moški* kot besedi *ženska*), pristranosti pa se kažejo tudi preko stereotipnih rešitev analogij (npr. Bolukbasi et al. (2016): rešitev analogije *moški:programer::ženska:x* je *gospodinja*). Nissim et al. (2019) opozarjajo, da tovrstne raziskave pretirano poudarjajo pristranost.

Ker se lahko pristranost z uporabo orodij, ki uporabljajo vložitve, ojača (Zhao et al., 2017), se več raziskovalnih skupin ukvarja z metodami "razpristranjevanja" (angl. *debiasing*) vektorskih vložitev. Primeri teh postopkov so izenačevanja oddaljenosti med spolno zaznamovanimi besedami in poklici (Bolukbasi et al., 2016; Bordia in Bowman, 2019), vstavljanje dodatnih omejitev v učni korpus (npr. zagotavljanje enake razporeditve poklicnih aktivnosti med spoloma v učnih podatkih) (Zhao et al., 2017), odstranjevanja tekstov, ki povzročajo pristranost (Brunet et al., 2018), in učenje spolno nevtralnih vektorskih vložitev (Zhao et al., 2018). Gonen in Goldberg (2019) opozarjata, da mnogi postopki "razpristranjevanja" pristranost le zakrijejo, medtem ko ta dejansko ostane prisotna v vložitvah.

Študije na področju raziskovanja pristranosti v vektorskih vložitvah besed so pogosto zasnovane na analogijah poklicev. Ker je v slovenščini spol besed morfološko izražen, kot rezultat analogije pričakujemo žensko oz. moško obliko poklica. Predhodna raziskava na besednih vložitvah (word2vec) v slovenščini (Supej et al., 2019) je pokazala, da je natančnost iskanja analogij dokaj visoka tako pri iskanju moškega kot tudi ženskega poklica. Rezultati kljub temu odražajo spolne pristranosti: rezultat analogije *ženska:tajnica :: moški:x* da rezultat $x =$ *šef*, prvih 10 najbližjih rezultatov različnih analogij pa odseva več spolnih neenakosti: asociacija žensk z domačimi opravili, moških s poklici višjega statusa itd. V nasprotju s predhodno raziskavo se naš prispevek ne ukvarja s sociološko problematizacijo rezultatov analogij enega tipa besednih vložitev (tj. word2vec), temveč preko analogij poklicev ovrednoti različne modele vložitev, njihove konfiguracije in morebiten vpliv filtriranja podatkov na rezultate. V prispevku torej razširimo študijo (Supej et al., 2019) z obširno analizo razpoložljivih modelov slovenskih vektorskih vložitev besed na razširjenem seznamu poklicev.

## 3. Podatki

V tem razdelku predstavimo sestavljen seznam poklicev ter opišemo različne vektorske vložitve besed.

### 3.1. Seznam poklicev

Naš izbor poklicev temelji na standardni klasifikaciji poklicev (Vlada RS, 1997), katere osnova je *Mednarodna standardna klasifikacija poklicev*. Večina poklicev v klasifikaciji je večbesednih zvez (npr. *upravljalec/upravljalka metalurškega žerjava*), ki so zaradi svoje specifičnosti in obsežnosti manj primerne za računske naloge. Za potrebe izračuna analogij smo se omejili na enobesedne poklice. Celotni seznam enobesednih poklicev zajema 422 parov, ki jih omejimo še glede na naslednje kriterije:

(1) Poklic ima žensko in moško obliko (spolno nevtralne besede, npr. *pismonoša*, niso vključene). (2) Vsaj ena izmed oblik poklica se pojavi v Slovenskem oblikoslovnem leksikonu *Sloleks 2.0* (Dobrovoljc et al., 2019), ki omogoča razlikovanje med lastnimi in občnimi imeni (nekateri poklici so namreč tudi lastna imena; npr. *kovač*), ali pa se v referenčnem korpusu standardne slovenščine Gigafida 2.0 (2020) pojavi 500- ali večkrat. (3) V primerih, kjer ocenimo, da za poimenovanje v standardni klasifikaciji poklicev obstajajo bolj uveljavljene različice, naboru podatkov dodamo sopomenko z istim korenom (npr. za izraz *fotografka*) iz standardne klasifikacije dodamo sopomenko *fotografinja*). Pri izračunih za izhodiščne besede upoštevamo obliko, ki je v korpusih bolj pogosta, pri pravilnosti izračunanih analogij pa upoštevamo katerokoli različico. (4) Če standardna klasifikacija ne vključuje

ženske (npr. *dramatik*) ali moške oblike poklica (npr. *prostitutka*), smo ročno dodali ustrezno različico, v kolikor ta obstaja (npr. za *postreščka* in za *hosteso* uveljavljene ženske oz. moške oblike ni) in je prisotna v Gigafidi. (5) Iz nabora smo izključili poklice, kjer je ženska oz. moška varianta poklica homofon (npr. *strežnik*, *detektivka*), oz. kjer je poklic možno asociirati s poklici nepovezanim kontekstom (npr. *čarovnik/čarovnica*).

Končni seznam vsebuje 234 parov poklicev, ki bo prosto dostopen na repozitoriju CLARIN [1].

### 3.2. Modeli vektorskih vložitev

V eksperimentih smo uporabili več različnih konfiguracij znanih vektorskih vložitev:

- fastText (Bojanowski et al., 2016):
  - 100-dimenzionalni vektorji, naučeni tekom projekta EMBEDDIA[2] na Gigafidi 2.0,
  - 300-dimenzionalni vektorji, naučeni kot v prejšnjem primeru,
  - 100-dimenzionalni vektorji besed s portala Sketch Engine (word),
  - 100-dimenzionalni vektorji s portala Sketch Engine, kjer so vektorji vložitve lem (lemma),
  - 100-dimenzionalni vektorji CLARIN.SI-embed.sl (Ljubešić in Erjavec, 2018) in
  - 300-dimenzionalni vektorji s portala fasttext.cc;
- word2vec (Mikolov et al., 2013a): 256-dimenzionalni vektorji, ki so bili naučeni za potrebe portala Kontekst.io (Plahuta, 2020) in so na voljo po dogovoru[3];
- ELMo (Peters et al., 2018): 1024-dimenzionalni vektorji kontekstnih vložitev projekta *EMBEDDIA*, naučeni na Gigafidi (Ulčar, 2019), kjer so vzete povprečne vrednosti 200.000 najpogostejših besed (izračunano na podlagi slovenske Wikipedije). Uporabljenih je bilo več različnih vrst vektorjev:
  - vektorji z izhoda prvega (CNN) nivoja mreže, ki so kontekstno neodvisni (tj. *layer 0*),
  - vektorji z izhoda drugega (prvega LSTM) nivoja mreže, ki so kontekstno odvisni (tj. *layer 1*),
  - vektorji z izhoda tretjega (drugega LSTM) nivoja mreže, ki so kontekstno odvisni (tj. *layer 2*).

## 4. Metodologija evalvacije

Analogije poklicev smo za vsako izmed vložitev izračunali na štiri različne načine. Jedro pristopa je pri vseh načinih enako: za vsako moško obliko poklica ($P_m$) iščemo ustrezno žensko obliko ($P_f$). Izračunamo vektor

$$\overrightarrow{d} = \overrightarrow{P_m} - \overrightarrow{m} + \overrightarrow{f},$$

kjer je $\overrightarrow{m}$ moški vektor in $\overrightarrow{f}$ ženski vektor. V idealnem primeru bi bil vektor $\overrightarrow{d}$ enak $\overrightarrow{P_f}$. Vektorju $\overrightarrow{d}$ poiščemo vektorje $N$ najbližjih besed glede na kosinusno razdaljo.

[1] http://hdl.handle.net/11356/1347
[2] http://embeddia.eu/
[3] https://kontekst.io/partnerstvo

| $m$ | $f$ | $m$ | $f$ |
|---|---|---|---|
| moški | ženska | brat | sestra |
| gospod | gospa | oče | mati |
| fant | dekle | sin | hči |
| fant | punca | dedek | babica |
| deček | deklica | mož | žena |
| stric | teta | on | ona |

Tabela 1: Pari inherentno moških in ženskih besed.

Pri iskanju najbližjih besed smo upoštevali vse besede, ki se nahajajo v vložitvah, razen besed *moški*, *ženska*, besede $P_m$ ter vseh besed, ki vsebujejo nečrkovne simbole (številke, vezaje, druga ločila, itd.) Če se beseda $P_f$ nahaja med $N$ najbližjimi besedami, ta primer štejemo kot pravilno določenega, sicer kot napačnega. Pri tem smo ignorirali velike in male začetnice, na primer besede *Zdravnik*, *zdravnik* in *ZDRAVNIK* upoštevamo kot isto besedo.

Postopek ponovimo za vsako žensko obliko poklica ($P_f$), kjer iščemo ustrezno moško obliko ($P_m$). Vektor $\overrightarrow{d}$ v tem primeru izračunamo kot

$$\overrightarrow{d} = \overrightarrow{P_f} - \overrightarrow{f} + \overrightarrow{m},$$

pri iskanju najbližjih besed pa namesto besede $P_m$ izpustimo besedo $P_f$. Končni rezultat predstavlja delež pravilno določenih primerov, oz. mera *natančnost pri* $N$ (angl. *precision at* $N$ oz. $P@N$).Višji $N$ omogoča primerjavo v širši okolici zadetka v vektorskem prostoru.

Za določitev moškega vektorja $\overrightarrow{m}$ in ženskega vektorja $\overrightarrow{f}$ smo uporabili dva pristopa. V prvem je $m$ kar beseda *moški* in $f$ beseda *ženska*. V drugem pristopu razliko $\overrightarrow{f} - \overrightarrow{m}$, oz. $\overrightarrow{m} - \overrightarrow{f}$ podobno kot Bolukbasi et al. (2016) predstavimo s povprečno razliko vektorjev parov besed, ki se specifično nanašajo na žensko oz. moškega (Tabela 1).

Pri iskanju najbližjih N besed smo uporabili tudi alternativen pristop, kjer smo vse besede v vložitvah lematizirali z orodjem LemmaGen[4]. S tem smo izničili vpliv pregibanja besed; na primer, besedi *zdravnico* in *zdravnice* sta pri tem postopku enaki, saj imata isto lemo *zdravnica*.

## 5. Rezultati

Rezultate predstavimo za vsak pristop, opisan v 4. razdelku, z mero natančnost pri $N$, kjer je $N$ enak 1, 5 in 10. Nekaterih poklicev z našega seznama ni v vseh vložitvah. Če iskane besede ni med $N$ najbližjimi, je primer označen kot napačen, četudi te besede sploh ni med vložitvami. Primere, ko poklica, ki ga imamo na vhodu, ni med vložitvami in tako ne moremo izračunati vektorja $\overrightarrow{d}$, obravnavamo na dva načina. V prvem načinu (`all`) tak primer štejemo kot napačen, v drugem načinu (`covered`) pa ga izločimo iz primerov in na končni rezultat ne vpliva.

Rezultati, kjer imamo na vhodu moški poklic $P_m$ in iščemo ustrezni ženski poklic $P_f$ so v Tabeli 2. Rezultati, kjer za ženski poklic $P_f$ na vhodu iščemo moški poklic $P_m$ so v Tabeli 3. Pristop, pri katerem smo vse besede lematizirali, ima pripono `_lem`. Pristop, kjer smo za moški in ženski vektor oz. njuno razliko uporabili povprečne razlike vektorjev besed iz tabele 1, ima pripono `_avg`.

[4] https://github.com/vpodpecan/lemmagen3/

| Vložitve | št. dimenzij in pristop | all P@1 | all P@5 | all P@10 | covered P@1 | covered P@5 | covered P@10 |
|---|---|---|---|---|---|---|---|
| ELMo Embeddia | 1024D_l0_avg | 0.128 | 0.278 | 0.299 | 0.166 | 0.359 | 0.387 |
| | 1024D_l0 | 0.162 | 0.291 | 0.308 | 0.210 | 0.376 | 0.398 |
| | 1024D_l0_lem_avg | 0.286 | 0.308 | 0.312 | 0.370 | 0.398 | 0.403 |
| | 1024D_l0_lem | 0.291 | 0.303 | 0.312 | 0.376 | 0.392 | 0.403 |
| | 1024D_l1_avg | 0.295 | 0.303 | 0.308 | 0.381 | 0.392 | 0.398 |
| | 1024D_l1 | 0.291 | 0.303 | 0.303 | 0.376 | 0.392 | 0.392 |
| | 1024D_l1_lem_avg | 0.295 | 0.303 | 0.308 | 0.381 | 0.392 | 0.398 |
| | 1024D_l1_lem | 0.291 | 0.303 | 0.303 | 0.376 | 0.392 | 0.392 |
| | 1024D_l2_avg | 0.291 | 0.308 | 0.308 | 0.376 | 0.398 | 0.398 |
| | 1024D_l2 | 0.286 | 0.308 | 0.308 | 0.370 | 0.398 | 0.398 |
| | 1024D_l2_lem_avg | 0.291 | 0.308 | 0.308 | 0.376 | 0.398 | 0.398 |
| | 1024D_l2_lem | 0.286 | 0.308 | 0.308 | 0.370 | 0.398 | 0.398 |
| fastText.cc | 300D_avg | 0.594 | 0.722 | 0.735 | 0.607 | 0.738 | 0.751 |
| | 300D | 0.436 | 0.688 | 0.718 | 0.445 | 0.703 | 0.734 |
| | 300D_lem_avg | 0.641 | 0.739 | 0.748 | 0.655 | 0.755 | 0.764 |
| | 300D_lem | 0.487 | 0.709 | 0.735 | 0.498 | 0.725 | 0.751 |
| fastText Embeddia | 100D_avg | 0.667 | 0.709 | 0.714 | 0.672 | 0.716 | 0.720 |
| | 100D | 0.632 | 0.709 | 0.714 | 0.638 | 0.716 | 0.720 |
| | 100D_lem_avg | 0.671 | 0.714 | 0.718 | 0.677 | 0.720 | 0.724 |
| | 100D_lem | 0.632 | 0.709 | 0.714 | 0.638 | 0.716 | 0.720 |
| | 300D_avg | 0.662 | 0.709 | 0.718 | 0.668 | 0.716 | 0.724 |
| | 300D | 0.679 | 0.714 | 0.714 | 0.685 | 0.720 | 0.720 |
| | 300D_lem_avg | 0.679 | 0.714 | 0.718 | 0.685 | 0.720 | 0.724 |
| | 300D_lem | 0.679 | 0.714 | 0.714 | 0.685 | 0.720 | 0.720 |
| fastText CLARIN.SI-embed.sl | 100D_avg | **0.761** | 0.868 | 0.880 | **0.761** | 0.868 | 0.880 |
| | 100D | 0.705 | 0.855 | 0.885 | 0.705 | 0.855 | 0.885 |
| | 100D_lem_avg | **0.761** | **0.880** | **0.902** | **0.761** | **0.880** | **0.902** |
| | 100D_lem | 0.709 | 0.859 | 0.885 | 0.709 | 0.859 | 0.885 |
| fastText Sketch Engine (word) | 100D_avg | 0.714 | 0.765 | 0.774 | 0.717 | 0.768 | 0.777 |
| | 100D | 0.688 | 0.765 | 0.774 | 0.691 | 0.768 | 0.777 |
| | 100D_lem_avg | 0.722 | 0.778 | 0.782 | 0.725 | 0.781 | 0.785 |
| | 100D_lem | 0.688 | 0.765 | 0.778 | 0.691 | 0.768 | 0.781 |
| fastText Sketch Engine (lemma) | 100D_avg | 0.598 | 0.786 | 0.821 | 0.598 | 0.786 | 0.821 |
| | 100D | 0.380 | 0.658 | 0.756 | 0.380 | 0.658 | 0.756 |
| word2vec Kontekst.io | 256D_avg | 0.402 | 0.543 | 0.585 | 0.407 | 0.550 | 0.593 |
| | 256D | 0.248 | 0.483 | 0.509 | 0.251 | 0.489 | 0.515 |
| | 256D_lem_avg | 0.402 | 0.543 | 0.585 | 0.407 | 0.550 | 0.593 |
| | 256D_lem | 0.248 | 0.483 | 0.513 | 0.251 | 0.489 | 0.519 |

Tabela 2: Rezultati za vse vložitve in variante, kjer je na vhodu moški poklic in iščemo ustrezen ženski poklic. Če za moški poklic na vhodu ne najdemo vložitve, tak primer štejemo kot napačno ugotovljen (all), oz. ga izpustimo iz rezultatov (covered). Najboljši rezultati v vsakem stolpcu so odebeljeni.

Rezultati kažejo, da dobimo boljše rezultate s fastText vložitvami, z izračunom, kjer namesto samega vektorja *moški* oz. *ženska* uporabimo povprečje besed z inherentno izraženim spolom ter z lematizacijo. Rezultate podrobneje razčlenimo v naslednji sekciji.

## 6. Diskusija

Vložitve, ki dosegajo največjo natančnost pri ugotavljanju analogij (z vhodnim moškim poklicem), so vložitve fastText CLARIN.SI-embed.sl (Tabela 2). Pri vhodnem ženskem poklicu dosegajo največjo natančnost, če upoštevamo le vložitve poklicev, ki so prisotni, fastText Embeddia, medtem ko so na vzorcu vseh vložitev najnatančnejše vložitve fastText CLARIN.SI-embed.sl (Tabela 3). V različnih modelih vložitev, pri različnih vhodnih podatkih velja, da lematizacija izhodnih podatkov in hkrati uporaba vektorja povprečne razlike med ženskimi in moškimi besedami (namesto uporabe le besed *ženska* oz. *moški*) izboljša natančnost analogije. Modeli, kjer je na vhodu poklic ženskega spola, v povprečju dosegajo višjo natančnost analogij v primeru `covered` rezultatov (če ženskega poklica ni v med vložitvami, tega ne štejemo kot napačno ugotovljeno analogijo). Rezultati `all` so podobni pri obeh tipih vhodnih podatkov.

Vložitve fastText Embeddia dosegajo zelo podobne rezultate s 100- in 300-dimenzionalnimi vložitvami, (glej Tabeli 2 in 3). (Druge vložitve so bile naučene na drugih jezikovnih virih, zato niso neposredno primerljive.) Vendar pa je iz Tabele 5 (vložitve fastText Embeddia) tudi razvidno, da igra dimenzionalnost veliko vlogo pri tem, kako pogosto je rezultat analogije sam vhodni poklic. Dimenzionalnost bi torej imela velik vpliv na nefiltrirane rezultate.

V vseh modelih vložitev je delež poklicev moškega spola večji kot delež poklicev ženskega spola (Tabela 4).

| Vložitve | št. dimenzij in pristop | all P@1 | all P@5 | all P@10 | covered P@1 | covered P@5 | covered P@10 |
|---|---|---|---|---|---|---|---|
| ELMo Embeddia | 1024D_l0_avg | 0.226 | 0.299 | 0.303 | 0.707 | 0.933 | 0.947 |
| | 1024D_l0 | 0.137 | 0.295 | 0.303 | 0.427 | 0.920 | 0.947 |
| | 1024D_l0_lem_avg | 0.291 | 0.299 | 0.303 | 0.907 | 0.933 | 0.947 |
| | 1024D_l0_lem | 0.286 | 0.303 | 0.303 | 0.893 | 0.947 | 0.947 |
| | 1024D_l1_avg | 0.291 | 0.303 | 0.303 | 0.907 | 0.947 | 0.947 |
| | 1024D_l1 | 0.282 | 0.303 | 0.303 | 0.880 | 0.947 | 0.947 |
| | 1024D_l1_lem_avg | 0.291 | 0.303 | 0.303 | 0.907 | 0.947 | 0.947 |
| | 1024D_l1_lem | 0.291 | 0.303 | 0.303 | 0.907 | 0.947 | 0.947 |
| | 1024D_l2_avg | 0.282 | 0.299 | 0.299 | 0.880 | 0.933 | 0.933 |
| | 1024D_l2 | 0.274 | 0.295 | 0.299 | 0.853 | 0.920 | 0.933 |
| | 1024D_l2_lem_avg | 0.282 | 0.299 | 0.299 | 0.880 | 0.933 | 0.933 |
| | 1024D_l2_lem | 0.274 | 0.295 | 0.299 | 0.853 | 0.920 | 0.933 |
| fastText.cc | 300D_avg | 0.291 | 0.590 | 0.675 | 0.393 | 0.798 | 0.913 |
| | 300D | 0.111 | 0.415 | 0.585 | 0.150 | 0.561 | 0.792 |
| | 300D_lem_avg | 0.453 | 0.654 | 0.701 | 0.613 | 0.884 | 0.948 |
| | 300D_lem | 0.338 | 0.637 | 0.679 | 0.457 | 0.861 | 0.919 |
| fastText Embeddia | 100D_avg | 0.654 | 0.705 | 0.709 | 0.900 | 0.971 | 0.976 |
| | 100D | 0.342 | 0.632 | 0.658 | 0.471 | 0.871 | 0.906 |
| | 100D_lem_avg | 0.658 | 0.705 | 0.709 | 0.906 | 0.971 | 0.976 |
| | 100D_lem | 0.534 | 0.671 | 0.684 | 0.735 | 0.924 | 0.941 |
| | 300D_avg | 0.607 | 0.705 | 0.709 | 0.835 | 0.971 | 0.976 |
| | 300D | 0.239 | 0.624 | 0.697 | 0.329 | 0.859 | 0.959 |
| | 300D_lem_avg | 0.688 | 0.709 | 0.714 | **0.947** | **0.976** | **0.982** |
| | 300D_lem | 0.594 | 0.705 | 0.709 | 0.818 | 0.971 | 0.976 |
| fastText CLARIN.SI-embed.sl | 100D_avg | 0.731 | 0.850 | 0.876 | 0.784 | 0.913 | 0.940 |
| | 100D | 0.077 | 0.547 | 0.726 | 0.083 | 0.587 | 0.780 |
| | 100D_lem_avg | **0.782** | **0.876** | **0.885** | 0.839 | 0.940 | 0.950 |
| | 100D_lem | 0.607 | 0.821 | 0.855 | 0.651 | 0.881 | 0.917 |
| fastText Sketch Engine (word) | 100D_avg | 0.701 | 0.761 | 0.769 | 0.886 | 0.962 | 0.973 |
| | 100D | 0.167 | 0.598 | 0.718 | 0.211 | 0.757 | 0.908 |
| | 100D_lem_avg | 0.735 | 0.761 | 0.769 | 0.930 | 0.962 | 0.973 |
| | 100D_lem | 0.641 | 0.752 | 0.761 | 0.811 | 0.951 | 0.962 |
| fastText Sketch Engine (lemma) | 100D_avg | 0.581 | 0.803 | 0.829 | 0.673 | 0.931 | 0.960 |
| | 100D | 0.440 | 0.701 | 0.769 | 0.510 | 0.812 | 0.891 |
| word2vec Kontekst.io | 256D_avg | 0.453 | 0.568 | 0.581 | 0.679 | 0.853 | 0.872 |
| | 256D | 0.244 | 0.393 | 0.479 | 0.365 | 0.590 | 0.718 |
| | 256D_lem_avg | 0.453 | 0.568 | 0.581 | 0.679 | 0.853 | 0.872 |
| | 256D_lem | 0.342 | 0.457 | 0.530 | 0.513 | 0.686 | 0.795 |

Tabela 3: Rezultati za vse vložitve in variante, kjer je na vhodu ženski poklic in iščemo ustrezen moški poklic. Če za ženski poklic na vhodu ne najdemo vložitve, tak primer štejemo kot napačno ugotovljen (all), oz. ga izpustimo iz rezultatov (covered). Najboljši rezultati v vsakem stolpcu so odebeljeni.

Največja pokritost je v vložitvah fastText CLARIN.SI-embed.sl, pri vložitvah modela ELMo pa se na primer pojavi le 75 od 234 izbranih poklicev ženskega spola. Razlog za mnogo manjšo zastopanost poklicev pri modelu ELMo je, da smo se zaradi tehničnih razlogov omejili le na 200 tisoč najpogostejših besed v Wikipediji (ELMo vložitve so v osnovi kontekstualne vložitve in je proces povprečenja računsko zahteven). Pri drugih tehnologijah vložitev smo imeli približno milijon besed. Moški poklici, ki se ne pojavljajo v vložitvah, so običajno poklici, ki so tipično povezani z ženskim spolom (npr. *šiviljec* ali *kozmetik*). Tudi poklici ženskega spola, ki se ne pojavljajo v vložitvah, so npr. tradicionalno povezani z moškimi (v vložitvah različnih tipov na primer ni *avtomehaničarke, tesarke* itd.) ali pa gre za kulturno pogojene izključno moške poklice (npr. *nadškof*). Slabo zastopanost poklicev ženskega spola lahko povežemo tudi z drugimi faktorji – Zhao et al. (2018) poročajo, da se različne zvrsti tekstov pogosteje nanašajo na moške v okviru njihovega poklica kot pa je to pri ženskah.

Modeli vložitev v konfiguraciji `lem_avg` (uporaba vektorja povprečnih razlik med spolno zaznamovanimi besedami in lematiziranje izhodnih podatkov) dajejo zelo različne rezultate. Rezultati analogij pri modelih ELMo in word2vec so večinoma poklici. Pri vložitvah fastText Embeddia, CLARIN.SI-embed.sl in Sketch Engine (word) so rezultati poklici in ostale besede, sorodne vhodnemu poklicu, ter besede z istim korenom kot vhodni poklic. Rezultati modelov fastText.cc in Sketch Engine (lemma) so večinoma besede z istim korenom kot vhodni poklic.

Po mnenju Nissim et al. (2019) je interpretacija večine študij, ki povezujejo analogije s pristranostjo, pretirana. Računanje analogij je namreč zastavljeno tako, da se izključi vhodni poklic, četudi bi bil to dejanski rezultat z

najvišjo kosinusno podobnostjo. Kljub temu, da smo pri rezultatih izločili vhodne poklice, kar je za samo računanje analogij standarden postopek, smo analizirali tudi rezultate pred filtriranjem. Pri analizi teh rezultatov smo opazili, da je med rezultati analogije z najvišjo kosinusno podobnostjo pogosto sam vhodni poklic (Tabela 5), kar pa zelo variira med posameznimi modeli.

Rezultati analogij so zanimivi z vidika semantike. Prva rezultata analogij (vložitev fT Embeddia 100D_lem_avg) *ženska:krojačica :: moški:x krojač* in *ženska:šivilja :: moški:x krojač* ponazarjata, da besedne vektorske vložitve upoštevajo tako slovnične kot tudi semantične elemente (vektorske vložitve besede *šiviljec* ni, *krojač* pa je semantično povezana beseda). Rešitve nekaterih analogij (predvsem v modelu w2v Kontekst.io lem_avg) z vhodnim poklicem niso povezane ali so stereotipne. Npr., rešitve analogije *moški:rudar :: ženska:x* v modelu w2v Kontekst.io lem_avg so npr.: *barbika, klovnesa, čarovnica, lutka, prostitutka, akrobatka, najstnica, opica, princeska, striptizeta*. Na stereotipne analogije v modelu w2v opozorijo tudi v Supej et al. (2019).

V okviru analize smo naredili tudi skupni frekvenčni seznam rezultatov analogij za vse ženske oz. moške vhodne poklice za posamezen model vložitev (upoštevajoč le konfiguracije `lem_avg`) (cf. Tabela 6). Opazimo vzorec, da se pri modelih ELMo l2_lem_avg in w2v Kontekst.io lem_avg najbolj pogosti ženski poklici/besede pojavljajo pogosteje kot najpogostejši moški poklici. Ena od možnih interpretacij je, da izbrana modela v primerjavi z nekaterimi drugimi vsebujeta relativno manj vektorskih besednih vložitev (200.000 oz. približno 600.000 za posamezen model). Oba modela imata tudi manjšo zastopanost ženskih oblik poklicev med besednimi vložitvami. Poklici, ki se kljub temu pojavljajo med vložitvami, se zato ponovijo večkrat. Poklicev moškega spola je v besednih vložitvah več, zato se posamezni poklici ne pojavljajo tako pogosto.

Med pogostimi ženskimi analogijami pri modelih ELMo l2_lem_avg in w2v Kontekst.io lem_avg zaznamo poklice nižjega družbenega statusa (*čistilka, perica, gospodinja*) ter zastarele poklice, kjer je bila ženska v podrejenem položaju (*služkinja*). Pri najpogostejših moških analogijah so poklici nižjega družbenega statusa izjemno redki.

Ugotavljamo tudi, da se nekatere besede (predvsem poklici ženskega spola) v rezultatih pojavljajo ne glede na semantično povezanost z vhodnim poklicem. V več primerih je rešitev analogije (predvsem ko gre za vhodni tipično moški poklic) nepovezana z vhodnim poklicem (npr. *bolničarka* kot prva rešitev analogije *moški:rudar :: ženska:x* in *šivilja* kot prva rešitev analogije *moški:avtomehanik :: ženska:x* v modelu fT Embeddia 100D_lem_avg). Možna razlaga, za potrditev katere bi bili potrebni dodatni testi, je, da so nekatere vektorske vložitve besed bolj 'centralne' od drugih in so najbližji sosed velikemu številu drugih besed, kar je v vektorskih vložitvah mogoč pojav. Možnost za nadaljnje delo je (delno) zmanjšati vpliv tovrstnih vložitev s pomočjo mere, alternativne kosinusni podobnosti, tj. CSLS (Conneau et al., 2018) oz. podobne mere, ki upošteva medsebojne razdalje najbližjih *n* sosedov).

| *vložitve* | *m* | *f* |
|---|---|---|
| ELMo | 0.774 | 0.321 |
| fastText_cc | 0.979 | 0.739 |
| fastText Embeddia | 0.991 | 0.726 |
| fastText CLARIN.SI-embedd.sl | **1.000** | **0.932** |
| fastText Sketch Engine (word) | 0.996 | 0.791 |
| fastText Sketch Engine (lemma) | 1.000 | 0.863 |
| word2vec Kontekst.io | 0.987 | 0.667 |

Tabela 4: Delež poklicev moškega (*m*) in ženskega (*f*) spola v vložitvah.

| Vložitve | št. dim. in pristop | delež |
|---|---|---|
| ELMo Embeddia | 1024D_l0_avg | 0.547 |
| | 1024D_l0 | 0.547 |
| | 1024D_l0_lem_avg | 0.547 |
| | 1024D_l0_lem | 0.547 |
| | 1024D_l1_avg | 0.423 |
| | 1024D_l1 | 0.483 |
| | 1024D_l1_lem_avg | 0.423 |
| | 1024D_l1_lem | 0.483 |
| | 1024D_l2_avg | 0.064 |
| | 1024D_l2 | 0.088 |
| | 1024D_l2_lem_avg | 0.064 |
| | 1024D_l2_lem | 0.088 |
| fT fastText.cc | 300D_avg | 0.831 |
| | 300D | 0.825 |
| | 300D_lem_avg | 0.831 |
| | 300D_lem | 0.825 |
| fT Embeddia | 100D_avg | 0.143 |
| | 100D | 0.141 |
| | 100D_lem_avg | 0.143 |
| | 100D_lem | 0.141 |
| | 300D_avg | 0.419 |
| | 300D | 0.513 |
| | 300D_lem_avg | 0.419 |
| | 300D_lem | 0.513 |
| fT CLARIN.SI-embed.sl | 100D_avg | 0.316 |
| | 100D | 0.310 |
| | 100D_lem_avg | 0.316 |
| | 100D_lem | 0.310 |
| fT Sketch Engine (word) | 100D_avg | 0.096 |
| | 100D | 0.135 |
| | 100D_lem_avg | 0.096 |
| | 100D_lem | 0.135 |
| fT Sketch Engine (lemma) | 100D_avg | 0.803 |
| | 100D | 0.927 |
| w2v Kontekst.io | 256D_avg | 0.483 |
| | 256D | 0.718 |
| | 256D_lem_avg | 0.483 |
| | 256D_lem | 0.718 |

Tabela 5: Delež primerov, pri katerih je rezultat analogije z najvišjo kosinusno podobnostjo sam vhodni poklic (pred filtriranjem za računanje rezultatov v Tabelah 2 in 3). Št. vseh primerov je 468 iz 234 parov poklicev.

## 7. Zaključki in nadaljnje delo

V prispevku smo na nalogi analogij moških in ženskih poklicev ovrednotili različne slovenske vektorske vložitve (z različnimi konfiguracijami in pristopi k računanju analogij). Ugotovili smo, da dobimo najboljše rezultate s fastText vložitvami. Pri ženskih analogijah za moške poklice

| ELMo Embeddia l2_lem_avg | | | | fastText CLARIN.SI_lem_avg | | | | word2vec Kontekst.io_lem_avg | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *m* vhod | | *f* vhod | | *m* vhod | | *f* vhod | | *m* vhod | | *f* vhod | |
| Rezultat | $n$ | Rezultat | $n$ | Rezultat | $n$ | Rezultat | $n$ | Rezultat | $n$ | Rezultat | $n$ |
| bolničarka | 47 | geograf | 9 | šivilja | 15 | mizar | 11 | kuharica | 44 | ortoped | 14 |
| biokemičarka | 39 | politolog | 8 | ključavničarka | 11 | biolog | 10 | gospodinja | 38 | pisatelj | 14 |
| frizerka | 39 | biolog | 7 | inštalaterka | 9 | ključavničar | 9 | šivilja | 33 | kardiolog | 13 |
| trgovka | 39 | dramaturg | 7 | keramičarka | 9 | zgodovinar | 9 | frizerka | 32 | nevrolog | 13 |
| čistilka | 34 | književnik | 7 | filologinja | 8 | internist | 8 | kozmetičarka | 30 | urolog | 13 |
| znanstvenica | 34 | scenarist | 7 | oftalmologinja | 8 | režiser | 8 | čistilka | 29 | psihiater | 12 |
| kuharica | 33 | animator | 6 | filozofinja | 7 | arheolog | 7 | fotografinja | 29 | ekolog | 11 |
| geologinja | 30 | esejist | 6 | geofizičarka | 7 | natakar | 7 | zdravnica | 29 | hišnik | 11 |
| perica | 28 | etnolog | 6 | kmetica | 7 | pisatelj | 7 | služkinja | 26 | biolog | 10 |
| služkinja | 28 | fotograf | 6 | nevrokirurginja | 7 | primarij | 7 | trgovka | 26 | korenjak | 10 |
| biologinja | 27 | ilustrator | 6 | strugarka | 7 | stomatolog | 7 | slikarka | 25 | maneken | 10 |
| gospodinja | 26 | lutkar | 6 | geologinja | 6 | tesar | 7 | tajnica | 25 | režiser | 10 |
| matematičarka | 26 | paleontolog | 6 | hematologinja | 6 | fotoreporter | 6 | veterinarka | 25 | akademik | 9 |
| mikrobiologinja | 26 | pravnik | 6 | kardiologinja | 6 | gostilničar | 6 | znanstvenica | 25 | akademski_slikar | 9 |
| arheologinja | 25 | režiser | 6 | paleontologinja | 6 | kardiolog | 6 | socialna_delavka | 24 | glasbenik | 9 |

Tabela 6: 15 najpogostejših besed, ki se pojavljajo med prvimi desetimi rezultati analogij v določenem modelu vložitev glede na vhodni poklic v analogiji (*m* ali *f*).

na vhodu se najbolje odreže model fastText CLARIN.SI-embed.sl, za ženske poklice na vhodu pa so to modeli fastText CLARIN.SI-embed.sl ter fastText Embeddia. Pristop, kjer za izračun namesto samega vektorja *moški* oz. *ženska* uporabimo povprečje besed z inherentno izraženim spolom, izboljša rezultate, enako velja za lematizacijo. Najboljši rezultati (P@10) so tako 0.885 za ženske iztočnice z modelom fastText CLARIN.SI-embed.sl-100D_lem_avg in 0.982 s fastText Embeddia 300D_lem_avg z upoštevanim pogojem, da so poklici v vložitvah. Za moške poklice na vhodu pa 0.902 z modelom fastText CLARIN.SI-embed.sl 100D_lem_avg (enak rezultat velja tudi za pogoj prisotnih vložitev). Modeli fastText CLARIN.SI-embed.sl imajo največji delež iskanih moških in ženskih poklicev. Med obravnavanimi vložitvami kažejo vložitve modela Kontekst.io pri kvalitativni analizi na največjo pristranost modela glede na spol (stereotipno ženski in moški poklici, ki se pojavljajo med analogijami ne glede na iztočnico). Prispevek se sicer osredotoča na kvantitativno evalvacijo in je s tem uporaben predvsem za razvijalce novih orodij, podrobnejši kvalitativni analizi in odnosu med vložitevami, jezikom in družbeno močjo pa se bomo posvetili v prihodnje. V nadaljnjem delu bomo obravnavali tudi kontekstne vložitve modela BERT, preizkusili metode za zmanjševanje vpliva vložitev, ki so bolj centralne od drugih, ter študijo razširili na druge jezike projekta EMBEDDIA. Poleg tega bomo preizkusili vpliv spolnih pristranosti v napovednih modelih na praktičnih nalogah, kot je analiza sentimenta.

## 8. Zahvala

Delo je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru programov P2-0103 (Tehnologije znanja) in P6-411 (Jezikovni viri in tehnologije za slovenski jezik) ter EU prek okvirnega programa za raziskave in inovacije Obzorje2020 - projekt EMBEDDIA (št. 825153).

## 9. Literatura

Gigafida 2.0. 2020. Gigafida 2.0: Korpus pisne standardne slovenščine. https://viri.cjvt.si/gigafida, 1. 5. 2020.

Shlomo Argamon, Moshe Koppel, Jonathan Fine in Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *TEXT*, 23:321–346.

Paul Baker. 2010. Will Ms ever be as frequent as Mr? a corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender & Language*, 4(1):125–149.

Piotr Bojanowski, Edouard Grave, Armand Joulin in Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama in Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. V: *30th Conference on Neural Information Processing Systems*.

Shikha Bordia in Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. *CoRR*, abs/1904.03035.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson in Richard S. Zemel. 2018. Understanding the origins of bias in word embeddings. *CoRR*, abs/1810.03611.

Carmen Rosa Caldas-Coulhard in Rosamund Moon. 2010. 'curvy, hunky, kinky': Using corpora as tools for critical analysis. *Discourse & Society*, 21(2):99–133.

Aylin Caliskan, Joanna J. Bryson in Arvind Narayanan. 2017. Semantics derived automatically from language corpora necessarily contain human biases. *Science*, 356(6334):183–186.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer in Hervé Jégou. 2018. Word translation without parallel data. V: *Proc. of International Conference on Learning Representation (ICLR)*.

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. CLARIN.SI. http://hdl.handle.net/11356/1230.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky in James Zou. 2018. Word embeddings quantify 100 years of gen-

der and ethnic stereotypes. *PNAS*, 115(16).

Aparna Garimella, Carmen Banea, Dirk Hovy in Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. V: *Proc. of the 57th Annual Meeting of the ACL*, str. 3493–3498. ACL.

Hila Gonen in Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. V: *Proc. of NAACL-HLT 2019*, str. 609–614.

Vojko Gorjanc. 2007. Kontekstualizacija oseb ženskega in moškega spola v slovenskih tiskanih medijih. V: I. Novak-Popov, ur., *Stereotipi v slovenskem jeziku, literaturi in kulturi: zbornik predavanj 43. seminarja slovenskega jezika, literature in kulture*, str. 173–180. Center za slovenščino kot drugi/tuji jezik, Ljubljana.

Benjamin Hill in Aaron Shaw. 2013. The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS One*, 8.

Dirk Hovy in Anders Søgaard. 2015. Tagging performance correlates with author age. V: *Proc. of the 53rd Annual Meeting of the ACL and the 7th IJCNLP*, str. 483–488.

Dirk Hovy. 2015. Demographic factors improve classification performance. V: *Proc. of the 53rd Annual Meeting of the ACL and the 7th IJCNLP*, str. 752–762. ACL.

Svetlana Kiritchenko in Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *CoRR*, abs/1805.04508.

Corina Koolen in Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. V: *Proc. of the First Ethics in NLP workshop*, str. 12–22. ACL.

Robin Lakoff. 1973. Language and woman's place. *Language in Society*, 2(1):45–80.

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.

Matej Martinc, Iza Škrjanec, Katja Zupan in Senja Pollak. 2017. PAN 2017: Author profiling - gender and language variety prediction. V: *Working Notes of CLEF 2017*.

Tomas Mikolov, Greg S. Corrado, Kai Chen in Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. V: *International Conference on Learning Representations*, str. 1–12.

Tomas Mikolov, Wen-tau Yih in Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. V: *Proc. of the 2013 Conference of the North American Chapter of the ACL: Human Language Technologies*, str. 746–751. ACL.

Malvina Nissim, Rik Noord in Rob van der Goot. 2019. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, str. 1–17.

Francisco M. Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein in Walter Daelemans. 2015. Overview of the 3rd author profiling task at PAN 2015. V: L. Cappellato, N. Ferro, G. J. F. Jones in E. SanJuan, ur., *Working Notes of CLEF 2015*, zvezek 1391 iz *CEUR Workshop Proceedings*. CEUR-WS.org.

Michael Pearce. 2008. Investigating the collocational behaviour of man and woman in the BNC using Sketch Engine. *Corpora*, 3:1–29, 05.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee in Luke Zettlemoyer. 2018. Deep contextualized word representations. V: *Proc. of NAACL-HLT 2018*, str. 2227–2237.

Marko Plahuta. 2020. O slovarju. https://kontekst.io/o-slovarju.

Marcelo O. R. Prates, Pedro H. Avelar in Luís C. Lamb. 2020. Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, 32:6363–6381.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang in William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. V: *Proc. of the 57th Annual Meeting of the ACL*, str. 1630–1640. ACL.

Anka Supej, Marko Plahuta, Matthew Purver, Michael Mathioudakis in Senja Pollak. 2019. Gender, language, and society: Word embeddings as a reflection of social inequalities in linguistic corpora. V: *Zbornik Slovenskega sociološkega srečanja 2019 - Znanost in družbe prihodnosti*, str. 75–83.

Deborah Tannen. 1990. *You Just Don't Understand: Women and Men in Conversation*. Ballantine Books, NY.

Matej Ulčar. 2019. ELMo embeddings model, Slovenian. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1257.

Eva Vanmassenhove, Christian Hardmeier in Andy Way. 2018. Getting gender right in neural machine translation. V: *Proc. of the EMNLP*, str. 3003–3008. ACL.

Ben Verhoeven, Iza Škrjanec in Senja Pollak. 2017. Gender profiling for Slovene Twitter communication: The influence of gender marking, content and style. V: *Proc. of the 6th BSNLP Workshop*, str. 119–125. ACL.

Vlada RS. 1997. 1641. uredba o uvedbi in uporabi standardne klasifikacije poklicev. *Uradni list RS*, 28:2217. https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina?urlid=199728&stevilka=1641.

Svitlana Volkova, Theresa Wilson in David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. V: *Proc. of the EMNLP*, str. 1815–1827. ACL.

Iza Škrjanec, Nada Lavrač in Senja Pollak. 2018. Napovedovanje spola slovenskih blogerk in blogerjev. V: D. Fišer, ur., *Viri, orodja in metode za analizo spletne slovenščine*, str. 356–373. Ljubljana: Znanstvena založba FF.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez in Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. V: *Proc. of the EMNLP*, str. 2979–2989. ACL.

Tianlu Zhao, Jieyu fand Wang, Mark Yatskar, Vicente Ordonez in Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. V: *Proc. of the NAACL-HLT*, str. 15–20. ACL.

# Uporaba jezikovnotehnoloških virov in postopkov pri razvoju učnega e-okolja Slovenščina na dlani

**Darinka Verdonik,* Simona Majhenič,* Špela Antloga,* Sandi Majninger,* Marko Ferme,* Kaja Dobrovoljc,* Simona Pulko,† Mira Krajnc Ivič,† Natalija Ulčnik†**

* Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška cesta 46, 2000 Maribor
darinka.verdonik@um.si; simona.majhenic@um.si; s.antloga@um.si; sandi.majninger@um.si; marko.ferme@um.si; kaja.dobrovoljc@um.si
† Filozofska fakulteta, Univerza v Mariboru
Koroška cesta 160, 2000 Maribor
simona.pulko@um.si; mira.krajnc@um.si; natalija.ulcnik@um.si

**Povzetek**

Prispevek predstavlja, kako so jezikovni viri in jezikovnotehnološki pristopi uporabljeni za razvoj učnega e-okolja Slovenščina na dlani. Jezikovnotehnološka podpora omogoča pomemben korak naprej v razvoju digitalnih učnih e-okolij, saj prinaša podporo prožnim oblikam poučevanja, poučevanju na daljavo, lajša učiteljevo delo, digitalno okolje pa omogoča tudi motiviranje učencev prek igričarskih elementov. V prispevku predstavljamo zasnovo e-okolja, podporne jezikovne vire in izvedbo vaj za štiri osrednje vsebinske sklope (pravopis, slovnico, frazeme in pregovore, besedila) ter razlage k vajam.

**Using Language Resources and Tools to Develop the E-Environment for Learning Slovene "Slovene in the Palm of Your Hand"**

This paper addresses the issue of using language resources and tools in the development of the e-environment for learning, "Slovene in the Palm of Your Hand". The use of language resources enables an important step forward in the development of digital e-environments. It supports flexible learning, learning at distance and facilitates teacher's work, while digital environments hold the potential to motivate learners with gaming elements. The paper represents the conceptual framework of the "Slovene in the Palm of Your Hand" environment, the developed language resources and the designed tasks for four major domains (orthography, grammar, phrasemes and proverbs, and texts) and accompanying explanations.

## 1. Uvod

Jezikovni viri, jezikovnotehnološka orodja in postopki so postali uveljavljeno področje raziskovanja jezika in njegove rabe; so nujen sestavni del številnih tehnoloških postopkov, zlasti na področjih velikih podatkov in umetne inteligence, ter temelj razvoja končnih jezikovnotehnoloških produktov, kot so prevajalnik, razpoznavalnik in sintetizator; nepogrešljivi so postali pri razvoju slovarjev in drugih jezikovnih priročnikov.

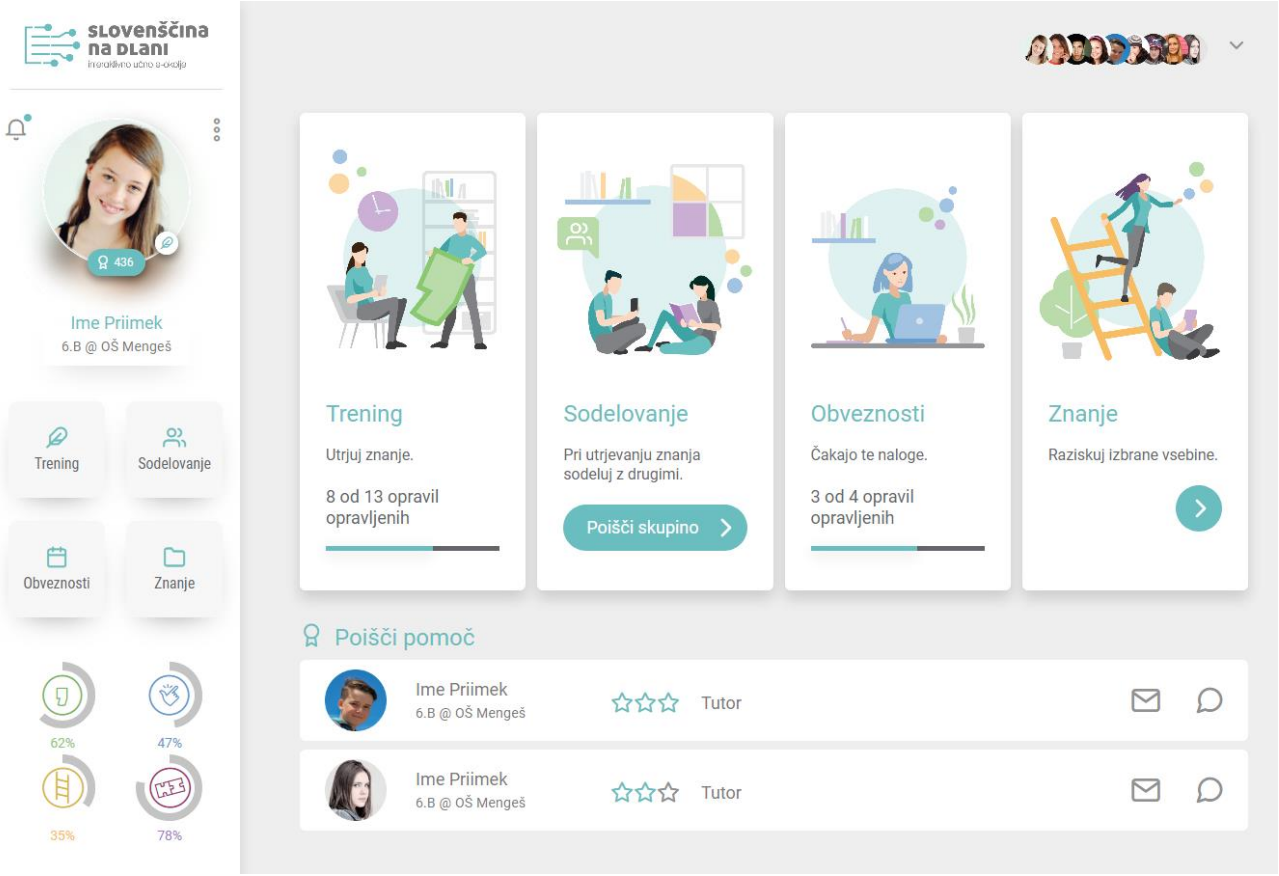
V projektu Slovenščina na dlani[1] smo raziskovali načine, kako jezikovne vire in jezikovnotehnološke pristope uporabiti pri razvoju didaktičnih pripomočkov za učenje jezika (Rozman et al., 2020). Cilj je bil razviti avtomatizirano, interaktivno, z jezikovnimi opisi podprto prosto dostopno učno e-okolje za slovenščino, v okviru katerega so *na dlani* različne dopolnjujoče se aktivnosti, s svojo dinamično zasnovo bistveno drugačne od obstoječih e-gradiv. Podpora z jezikovnimi viri in jezikovnotehnološkimi pristopi je bila pri tem ključen element, ki omogoča pomemben napredek v primerjavi s podobnimi obstoječimi okolji, zlasti v smislu upoštevanja prožnih in inovativnih oblik učenja, učne diferenciacije in individualizacije, tj. prilagajanja zmožnostim in potrebam vsakega posameznega uporabnika, ter formativnega spremljanja posameznikovega napredka kot učinkovitega vzvoda za prevzemanje odgovornosti za lastno učenje in napredovanje.

Pojem prožne oblike poučevanja (po definiciji Evropske komisije) vsebuje učenje in poučevanje, ki je odzivno na potrebe učečega se in njegove močne strani. Fleksibilno učenje ponuja učečemu se izbiro glede načina, okolja in časa učenja z namenom spodbujanja motivacije in vztrajnosti, lahko tudi v primerih, ko je prisotnost na lokaciji izvajanja učnega procesa otežena. Prožne oblike učenja in poučevanja tako omogočajo fleksibilno obliko dela, saj gre za sistem poučevanja in učenja, v katerem imajo učeči se možnost, da (lahko) del učenja opravijo tudi izven šolskega okolja, omenjene oblike učenja in poučevanja namreč omogočajo prilagajanje izvedbe učnih programov ob upoštevanju zahtev učnega programa ter pričakovanih učnih izidih, opredeljenih z učnimi načrti in katalogi znanj. Ob upoštevanju teh izhodišč smo pri pripravi e-okolja *Slovenščina na dlani* ustrezno prilagodili tudi metode in oblike poučevanja in učenja.

## 2. Zasnova učnega e-okolja Slovenščina na dlani

Učno e-okolje Slovenščina na dlani ponuja vaje in podporne strokovne vsebine s štirih področij slovenskega jezika: 1) pravopisa, 2) slovnice, 3) frazemov in pregovorov ter 4) besedil. Pri pravopisu in slovnici je poudarek na tistih temah in vsebinah, pri katerih se pomanjkanje znanja pogosto kaže skozi napake v pisanju učečih se. Na področju frazeologije se posvečamo podobi, pomenu in rabi frazemov ter pregovorov, pri čemer prav

[1] Projekt sofinancirata Republika Slovenija in Evropska unija iz Evropskega socialnega sklada..




Slika 1. Uvodna stran e-okolja za učenca

tako skušamo odpraviti primanjkljaje, ki se kažejo pri poznavanju in razumevanju teh leksikalnih enot, s čimer težimo k izboljšanju frazeološke kompetence med uporabniki e-okolja. Pri besedilih uporabniki e-okolja razvijajo (kritično) sporazumevalno jezikovno zmožnost z reševanjem vaj in nalog o zunajjezikovnem in jezikovnem kontekstu, in sicer z avtentičnimi besedili osmih tematskih sklopov. Vsak tematski sklop zajema vsaj tri sklope besedil, slikovito predstavljene s slogani. Vsak slogan pa predstavlja kombinacijo vsaj dveh pogostih in za večino govorcev slovenščine pomembnih besedilnih skupin.

Učno e-okolje loči dva glavna tipa uporabnikov: učitelje in učence. Učitelj v e-okolju upravlja sezname in aktivnosti svojih skupin in njihovih udeležencev; dodeljuje naloge; pregleduje uspešnost skupin in udeležencev; komunicira z njimi; si pripravlja teste in delovne liste; ali pa tudi dodaja in pregleduje nove vsebine. Učenec rešuje vaje, bodisi tiste, ki jih določi učitelj, bodisi jih samostojno izbira; seznani se lahko z znanji, potrebnimi za izvajanja posameznih vaj, ali pa prosto brska po bazi znanj; pregleduje svojo uspešnost; poišče pomoč učitelja ali tutorjev; pošilja sporočila in se povezuje z drugimi uporabniki; vnaša nove vsebine v e-okolje.

Lastnosti učnega e-okolja Slovenščina na dlani so avtomatiziranost, individualizacija, interaktivnost tako s sistemom kot med uporabniki in igričarski pristop.

Avtomatiziranost pomeni, da učno e-okolje učenca samo vodi skozi vaje in daje povratne informacije o učenčevi uspešnosti. Individualizacija se kaže v avtomatskem prilagajanju vsebine, količini in zahtevnosti vaj vsakemu posameznemu uporabniku, odvisno od rezultatov o njegovi uspešnosti pri posameznih vsebinah. Interaktivnost med uporabniki je omogočena skozi povezovanje uporabnikov med seboj, funkcionalnost tutorske pomoči ali pomoči učitelja, možnost združevanja v skupine, dodeljevanje nalog uporabnikom s strani učiteljev, pošiljanje sporočil ipd. Interakcija pa poteka tudi med sistemom in uporabnikom, s tem ko sistem uporabniku sproti signalizira uspešnost njegovega dela, mu pomaga z namigi, omogoča prebiranje razlag, po drugi strani pa lahko uporabnik tudi sam prispeva vsebine v učno e-okolje z vnašanjem razlag. Igričarski pristop je izveden tako skozi grafično oblikovanje kot tudi skozi samo zasnovo e-okolja, pri kateri uporabnik učenec zbira točke, osvaja medalje ter tekmuje s samim seboj ali z drugimi.

Oblikovanje uporabniškega vmesnika Slovenščina na dlani je bilo izvedeno v sodelovanju s profesionalnim oblikovalcem.[2] Slika 1 prikazuje uvodno stran e-okolja za učenca.

## 3. Vaje iz pravopisa in slovnice

Osnova za vaje iz pravopisa in slovnice je korpus besedil, ki zajema leposlovna, publicistična in spletna besedila, izbrana tako, da so vsebinsko in po zahtevnosti primerna in zanimiva za šolajočo se populacijo, hkrati pa tudi jezikovno pregledana.

[2] Samo Kramberger (https://samokramberger.com/).

### 3.1. Korpus MAKS

Korpus MAKS (akronim za MlAdinski KorpuS) obsega pribl. 10 mio. besed oz. pribl. 12 mio. pojavnic. Dobro polovico tega sestavljajo besedila iz mladinskega in drugega leposlovja ali priročnikov. Dobrih 40 % besedil je zajetih iz publicistike, manjši delež, dobrih 300.000 besed, pa s spleta. Omejevanje zajetih vsebin s spleta je bilo potrebno predvsem zaradi tega, ker je jezikovni pregled pri teh manj natančen, za korpus MAKS pa je bil to pomemben kriterij, saj je za izvajanje vaj in nalog iz slovnice in pravopisa treba priklicati zgolj slovnično in pravopisno pravilne primere. Besedilodajalci[3] so bili večinoma založbe, posamezni leposlovni avtorji, kar nekaj besedil pa je bilo prevzetih tudi iz korpusa Gigafida.

Vsa zajeta besedila so bila ročno pregledana; preverjeno je bilo, ali po vsebini (vsebina ni oglasna, ideološko zaznamovana, nasilna, spolna, zelo strokovna in težko razumljiva ipd.) in jezikovno (rabljen je knjižni jezik) ustrezajo specifičnim potrebam e-okolja Slovenščina na dlani. Pokazalo se je, da so vsebine, ki za učeče se niso primeren vir povedi za vaje iz slovnice in pravopisa, zelo pogoste: v leposlovju, zlasti nemladinskem, so bile pogoste vsebine, povezane z nasiljem, občasno pa smo morali vsebine izločati tudi zaradi slenga, vulgarizmov, narečnega ali starinskega jezika. V publicistiki so se po drugi strani pojavljale propagandne ali ideološko zaznamovane vsebine, občasno pa tudi jezikovno nezadostno pregledana besedila. Zahtevnim strokovnim vsebinam pa smo se skušali izogniti že pri izboru virov.

Ob pregledu smo besedilom pripisali vir, leto objave, naslov, avtorja, primernost za osnovno ali srednjo šolo ter teme, vsa besedila pa smo nato oblikoslovno, skladenjsko in imenskoentitetno označili.

Za oblikoslovno in skladenjsko označevanje smo uporabili Stanford Parser V3 (Qi et al., 2018). Pred označevanjem smo besedilo segmentirali na povedi in besede z orodjem Obeliks4J (Grčar et al., 2012). Na koncu smo z Stanford NER (Finkel et al., 2005) izvedli še označevanje imenskih entitet v besedilu.

Korpus MAKS je dostopen prek Clarinovega konkordančnika NoSke. Kot odprtodostopna zbirka ni dostopen zaradi omejitev pri izhodiščnih avtorskih pravicah.

### 3.2. Definiranje vaj

Vaje iz slovnice in pravopisa so zasnovane tako, da pokrivajo področja, ki se kažejo skozi napake pri pisanju, pri čemer smo si pomagali z analizo napak iz korpusa Šolar (Kosem et al., 2012, Rozman et al., 2020).

V sklopu pravopisa je največ nalog namenjenih pravilni rabi vejice, in sicer smo pripravili vaje za vejico pri podredju, priredju, pastavkih, pristavkih, vrivkih in pri zahtevnejših primerih, kot je na primer vejica pri večbesednih veznikih. Velik delež nalog predstavljajo tudi zahtevnejši primeri zapisa, pri katerih uporabniki lahko vadijo črkovanje, zapis prevzetih besed in druge zahtevnejše primere, zapis z veliko ali malo začetnico ter (ne)stičen zapis ali zapis z vezajem. V nekoliko manjšem obsegu so zastopane vaje za rabo drugih ločil.

Vsaki vaji smo določili zahtevnostno stopnjo na lestvici od 1 do 3 (osnovna, srednja, zahtevna), zato so vaje primerne za uporabnike z različnim predznanjem, torej za tiste, ki šele usvajajo pravopisna ali slovnična pravila, ali za tiste, ki želijo svoje znanje zgolj osvežiti. Zlasti za slednjo skupino so primerne zaključne naloge pravopisa (zahtevna raven), pri katerih uporabnik na igriv način rešuje pravopisne probleme vseh poglavij.

Pri slovnici je največ nalog namenjenih samostalnikom (naloge, povezane s sklanjanjem, zanikanim rodilnikom idr.), zaimkom (oziralni zaimki, ujemanje zaimka, zahtevnejši primeri idr.), glagolom (dvojina, opisni deležnik na *-l*, namenilnik, nedoločnik idr.) in predlogom, nekoliko manj nalog pa obravnava veznike in pridevnike ter besedni red.

Pravopisne in slovnične naloge temeljijo na skupno 16 osnovnih tipih nalog, s katerimi smo želeli zagotoviti ravnovesje med igrivostjo in avtomatizacijo nalog. Na osnovni stopnji uporabnik najprej prepoznava napake (npr. z izbiranjem, iskanjem, označevanjem (ne)pravilnega odgovora idr.), vendar je podoba nalog grafično različna, tako da uporabnik odgovor išče npr. v rešetu, ga kot sestavljanko dopolnjuje z danimi odgovori ali, tako kot v igri spomin, išče pare s podobnimi napakami. Pri zahtevnejših nalogah mora uporabnik napake prepoznati in jih tudi odpraviti (npr. s popravljanjem, premikanjem, vstavljanjem ipd.). Najzahtevnejše naloge pa od uporabnika zahtevajo tudi, da svojo odločitev utemelji. Da bi zagotovili čim bolj raznoliko aktivnost uporabnika, e-okolje poleg izbiranja nudi tudi naloge vpisovanja in premetank z odvečnimi črkami v različnih grafičnih oblikah.

### 3.3. Programiranje vaj

Programiranje vaj obsega več nalog, ki jih mora zagotoviti učno e-okolje. Prva naloga je najti primere povedi, ki ustrezajo področju, zahtevnosti naloge in tipu naloge, za katerega so uporabljeni. Navedeni primeri morajo zato vsebovati ustrezne gradnike, ki jih mora računalnik razpoznati, da lahko sestavi programsko predstavitev problema, ki ga nato ponudi uporabniku. Znotraj učnega e-okolja smo formalno definirali slabih 200 gradnikov, ki so lahko zelo preprosti (npr. vejica) ali pa zelo kompleksni (npr. vrstilni, ločilni in množilni števniki iz vsaj dveh sestavin).

Druga naloga je takšno programsko predstavitev problema prikazati uporabniku, mu hkrati glede na izbrano vajo prikazati navodila in omogočiti interaktivno reševanje omenjenega problema ter ustrezno hranjenje uporabnikovih odgovorov. Tretja naloga je vrednotenje odgovorov uporabnika, ki lahko poteka avtomatsko ali z zunanjim (ročnim) vrednotenjem.

Pri prvi nalogi, iskanju ustreznih primerov, smo z leksikalno analizo in formalnim zapisom pogojev za primere posameznih vaj dobili kandidate za primere posameznih nalog. Navedeni kandidati so bili nato ročno pregledani, da smo lahko izločili vsebinsko neprimerne povedi in tiste povedi, kjer je pri leksikalni analizi prišlo do napak. Končni nabor povedi tako predstavlja množico povedi, iz katerih se v učnem e-okolju tvorijo vaje. Pri vsaki posamezni vaji je izdelan tudi uporabniški vmesnik za urejanje navodil ter pregled in dodajanje primerov, kar omogoča, da je v učno e-okolje mogoče dodajati tako nove

[3] Navedeni so na spletni strani projekta http://projekt.slo-na-dlani.si/.

primere kot tudi nove vaje. Za posamezno vajo je poleg formalnega zapisa pogojev za iskanje primerov ustvarjen tudi formalni zapis pogojev za gradnike, ki predstavljajo interaktivne elemente (npr. število izbrisanih vejic, gradnik, v katerem smo spremenili velikost začetnice, itd.).

Prikaz vaje in njeno reševanje predstavlja osrednji del programiranja vaj. Programska predstavitev vaje je uporabniku prikazana v obliki spletnega vmesnika, katerega videz močno variira glede na tip naloge. Osnovni elementi interakcije, kot so vstavljanje preko besedilnih vnosnih elementov ali izbira posameznih elementov, ne predstavljajo večjih izzivov, bolj zahteven pa je prikaz interaktivnih elementov, ki jim uporabnik lahko spreminja pozicijo v besedilu (npr. ločila), ali čisto ločeni uporabniški vmesniki, kot so za rešeta ali sestavljanke. V vseh primerih je treba ustrezno hraniti uporabnikovo interakcijo, saj se poleg samega vnosa, kot je besedilo ali izbira, beležita tudi uporabnikovo napačno ravnanje ter čas, ki ga potrebuje za posamezno akcijo. Akcije se sproti beležijo in posredujejo v zaledje učnega sistema, kjer so ovrednotene, saj se na njihovi podlagi po logični shemi prožijo različne akcije. Logična shema se prav tako obnaša dinamično, saj je odvisna od vsebine same vaje, od vrste opravila, ki ga uporabnik opravlja, ter tudi od informacij, ki jih imamo o uporabniku. Glede na vse navedeno je bodisi ponujena pomoč ob neuspešnem reševanju bodisi obvestilo o pravilno ali napačno rešenemu primeru, hkrati pa ima navedeno tudi vpliv na naslednjo ponujeno vajo.

Samo vrednotenje uporabnikovih rešitev vaj poteka v zaledju sistema na dva načina. Prvi način je samodejno vrednotenje, kjer že pri tvorjenju vaje poznamo pravilno rešitev in jo lahko v času reševanja ovrednotimo. Samo vrednotenje je vezano na področje in zahtevnost vaje, je pa lahko v sklopu širšega nabora vaj tudi ustrezno uteženo. Samodejno vrednotenje daje uporabniku ob reševanju neposredno povratno informacijo, hkrati pa tvori njegov profil, v katerega beleži zgodovino uspešnosti reševanja. Drugi način vrednotenja je vezan na naloge, kjer pri generiranju vaje ni jasno določena pravilna rešitev. Takrat mora rešitev ovrednotiti drug uporabnik, običajno v vlogi učitelja ali tutorja. Slednje se dogaja predvsem pri nalogah, kjer mora uporabnik vnesti večjo količino besedila in kjer računalniški algoritmi še niso dovolj zanesljivi, da bi ponudili ustrezno vrednotenje.

# 4. Vaje s področja frazeologije

Na področju frazeologije smo predvideli vaje, ki so vezane na frazeme kot osnovne frazeološke enote, in na pregovore, ki jih umeščamo k frazeologiji v širšem smislu. Ciljni uporabniki e-okolja Slovenščina na dlani, tj. osnovno- in srednješolci, se z obojimi srečujejo pri jezikovnem pouku slovenščine ter jih z ozirom na obstoječa učbeniška gradiva sistematično obravnavajo, in sicer praviloma v 7. razredu osnovne šole ter v 2. letniku srednje šole. Učitelji slovenščine pri tem opažajo, da učečim se težave povzroča njihovo razumevanje, zaradi česar so negotovi tudi pri njihovi rabi, ob tem pa poseben problem predstavlja dejstvo, da v slovenskem prostoru ni na voljo priročnika, ki bi se posebej posvečal slovarski obravnavi idiomatskih enot in bi bil prilagojen šolajoči se populaciji. Izhajajoč iz teh ugotovitev smo naprej pripravili izbor sto frazemov in sto pregovorov, za katere smo v naslednji fazi pripravili vsestranske slovarske opise, na njihovi osnovi pa smo zasnovali tudi vaje s področja frazeologije. Pri tem smo težili k iskanju možnosti za njihovo avtomatizacijo.

## 4.1. Zbirka frazemov in pogovorov FRIDA

Zasnovali smo zbirko frazemov in pregovorov ter jo poimenovali z akronimom FRIDA (Frazemi in pRegovorI na DlAni).

Najprej smo izdelali ožji seznam enot. Izbrali smo sto frazemov in sto pregovorov (Ulčnik, 2019; Ulčnik in Meterc, 2019) ter pri tem izhajali iz naslednjih kriterijev: 1) aktualnost enot, 2) didaktična relevantnost enot, 3) pokrivanje različnih tematskih skupin. Aktualnost enot je bila določena na osnovi prisotnosti v učbeniških gradivih[4] ter frekventnosti pojavljanja v korpusih sodobnih pisnih besedil.[5] Za pregovore smo ob tem lahko upoštevali izdelan paremiološki optimum (Meterc, 2017), tj. seznam tristotih najbolj poznanih in uporabljanih enot, medtem ko za frazeme tovrsten seznam ne obstaja. Pri določanju didaktične relevantnosti smo izhajali iz predpostavke o oteženem razumevanju enot z izkazano večjo stopnjo idiomatičnosti oz. frazeološke pomenske neprozornosti, upoštevali pa smo tudi konkretne predloge učiteljev – ti so izpostavili enote, ki so med učečimi se slabše poznane ali pa je bilo pri pouku opaženo njihovo pomanjkljivo razumevanje (npr. *gordijski vozel, priti z dežja pod kap*). Pokrivanje različnih tematskih skupin je izhajalo iz osnovne zamisli o vključevanju frazemov in pregovorov z različnih tematskih skupin (npr. človek, odnosi, dejavnost in bivanje, predmetnost in pojavnost, čas in prostor, količina, mera, stopnja), s čimer smo želeli opozoriti na raznovrstnost njihove rabe in prisotnost na različnih semantičnih poljih. Vključili smo tudi enote, ki so značilne za slovenski govorni prostor in obenem v rabi izkazujejo zadostno frekventnost (npr. *vleči dreto*).

Izbrane enote smo v naslednji fazi celovito slovarsko opisali, pri čemer imajo opisi dvojno funkcijo – namenjeni so podrobnejšemu seznanjanju s konkretnimi frazemi, ki se pojavljajo tudi znotraj vaj, obenem pa služijo kot zgled za samostojno pripravo slovarskih opisov še neobdelanih enot s širšega seznama. Za vnašanje slovarskih opisov smo izdelali programski vmesnik in v njem predvideli večnivojsko podatkovno shemo. Na prvem nivoju smo predstavili oblikovne in tipološke lastnosti izbranih enot, na drugem nivoju pomensko-pragmatične lastnosti, na tretjem nivoju pa njihove slovnične lastnosti. Formulacije smo vseskozi prilagajali primarnim uporabnikom e-okolja, torej osnovno- in srednješolcem. Četrti nivo je bil namenjen navajanju dodatnih zgledov rabe, ki so uporabni pri pripravi vaj. Pri tem smo bili posebej pozorni na zglede z več frazeološkimi enotami, na pojasnjevanje izvora posameznih enot (npr. na mitološka pojasnila in etimološke razlage), na opažene dobesedne pomene pri enotah, ki omogočajo t. i. dvojno branje (Kržišnik, 2006: 260), na prisotnost uvajalnih sredstev ipd., obenem pa smo pazili tudi na zadostno navajanje besedilnega okolja, iz katerega

[4] Frazemi in pregovori so bili izpisani iz izbranih gradiv založb Mladinska knjiga in DZS (npr. iz beril od 6. do 9. razreda in od 1. do 4. letnika). Več o gradivu in poteku analize v Ulčnik, 2019; Ulčnik in Meterc, 2019.

[5] Izhodišče za preverjanje rabe je predstavljal korpus Gigafida v. 2.0 Dedup, upoštevali pa smo tudi KAS (akademska besedila), slWac (Slovene Web) in Janes (družbena omrežja).

je možno razpoznati pomen predstavljenih frazeoloških enot. Pripravljeni opisi, ki že v osnovi sledijo jasnemu, nazornemu in razumljivemu podajanju podatkov, so v e-okolju prikazani v prilagojeni zaslonski sliki, pri čemer sledimo uporabnikovemu interesu in omogočamo večnivojski prikaz z možnostjo selektivnega izbiranja podatkov v razponu med osnovnim in razširjenim oz. celostnim prikazom (prim. tudi Jesenšek in Ulčnik, 2014: 286). Za frazeme so tako sprva prikazani le izhodiščna enota, parafraza in zgled, šele v naslednji fazi pa pomenski opis, pragmatična pojasnila, morebitne so- in protipomenske enote ipd. Podatke, zbrane na vmesniku, smo skušali v čim večji meri izkoristiti tudi pri zasnovi in pripravi vaj.

### 4.2. Definiranje vaj

Pri definiranju vaj smo izhajali iz osnovne delitve na: 1) vaje, ki smo jih zasnovali z ozirom na podatke, predstavljene v slovarskih opisih, in 2) dodatne vaje, ki predpostavljajo uporabnikovo ustvarjalnost. Na podlagi spoznanj teoretične frazeologije in skladno z izsledki frazeodidaktike smo vaje iz prve skupine vezali na tri podsklope: 1) podoba, 2) pomen, 3) raba. Osnovni namen prvega podsklopa je doseči, da bodo uporabniki e-okolja frazeme in pregovore prepoznali v besedilu, da se bodo zavedali njihove oblikovne oz. sestavinske podobe (večbesednost, ustaljenost) in da bodo znali ločevati med frazemi ter pregovori. V drugem podsklopu preverjamo razumevanje pomena frazeoloških enot, (pre)poznavanje medfrazemskih razmerij (so- in protipomenskost), zmožnost njihovega pomenskega pojasnjevanja ter sposobnost njihovega nadomeščanja s slogovno nezaznamovanimi besedami. V tretjem podsklopu se osredinjamo na ustrezno frazeološko rabo ter sposobnost nadomeščanja nezaznamovanega izražanja s frazeološkim. Pri pregovorih smo posebej pozorni tudi na uveljavljeno rabo s t. i. uvajalnimi sredstvi (npr. *kot pravi znan slovenski pregovor*).

Pri vseh treh podsklopih je bilo v ospredju iskanje možnosti za čim večjo avtomatizacijo pri preverjanju odgovorov in posledično pri točkovanju opravljenih nalog, kar v končni fazi uporabniku zagotavlja takojšnjo povratno informacijo. Druga skupina vaj temu kriteriju ni mogla zadostiti, saj je vezana na kreativno rabo in izhaja iz problemskega pristopa, ki se nujno povezuje z doseganjem višjih taksonomskih ravni (analiza, sinteza, vrednotenje), te pa je zaradi neenoznačnih in nepredvidljivih odgovorov (skoraj) nemogoče avtomatično ocenjevati. Posledično smo za tovrstne vaje predvideli samovrednotenje oz. samorefleksijo na osnovi vnaprej pripravljenih možnih rešitev. Dva različna principa pri zasnovi vaj pa sta zahtevala tudi različne pristope z vidika njihovega programiranja.

### 4.3. Programiranje vaj

Podobno kot pri vajah iz pravopisa in slovnice je tudi pri frazeologiji programiranje izvedeno v več korakih. Prvi korak je izbrati frazeološke enote, ki ustrezajo zahtevnosti in tipu naloge. Naslednji korak je tvorjenje prikaza naloge in uporabniku omogočiti interaktivno reševanje. Na koncu je potrebno še vrednotenje pravilnosti uporabnikovega odgovora.

Pri izbiri ustrezne frazeološke enote je pomembno, da so v zbirki za izbrano enoto izpolnjeni vsi podatki, potrebni za tvorjenje naloge. Prav tako moramo upoštevati, da je izbrana frazeološka enota primerna za zahtevnostno stopnjo, ki jo določa naloga.

Tvorjenje prikaza naloge uporabniku vključuje izpis navodila za reševanje, izbor ustreznih namigov, ki so uporabniku v pomoč pri reševanju, ter izpis same naloge v spletnem uporabniškem vmesniku. Izpis je odvisen od tipa naloge ter od naprave, ki jo uporablja uporabnik. Nekateri tipi nalog potrebujejo še smiselne napačne odgovore. Pokazalo se je, da za določene naloge s področja frazeologije računalnik ne zna avtomatsko poiskati ustreznih alternativnih odgovorov, zato zahtevnost nalog ne bi bila primerna.[6] To težavo smo rešili tako, da smo zbirko frazeoloških enot ročno dopolnili s smiselnimi napačnimi odgovori, ki jih pri posamezni nalogi lahko izbere računalnik. Na tak način smo povečali zahtevnost tovrstnih nalog in njihovo kvaliteto.

Med reševanjem se uporabnikove akcije sproti beležijo v zalednem delu e-okolja. Pri vrednotenju končnega uporabnikovega odgovora se lahko upoštevajo še vmesni koraki, napačni poskusi in porabljen čas. Način vrednotenja je odvisen od naloge. Pri tem dopuščamo tudi nekatere nenamerne napake uporabnika, kot so npr. odvečni presledki pri vnosu besedila. E-okolje vsebuje tudi naloge, ki jih ni mogoče natančno avtomatsko ovrednotiti. Za take naloge uporabljamo računalniške algoritme, ki podajo približno oceno, lahko pa te naloge ovrednoti tudi drug uporabnik, običajno v vlogi učitelja ali tutorja.

## 5. Vaje iz besediloslovja

Delo z besedili v učno-vzgojnem procesu predstavlja svojevrsten izziv. Pogosto se zgodi, da sicer zanimivemu besedilu sledijo tri ali štiri naloge, ki od učečega se zahtevajo poznavanje kraja in časa nastanka besedila, udeležencev, njihovega družbenega razmerja, ki je običajno povsem zadostno opisano s terminoma (ne)formalno in (ne)javno, sledita še vprašanji o temi in sporočevalnem namenu. Za natančno razumevanje besedil pa zadostuje določitev prevladujočega slogovnega postopka in nekaj koreferenc k eni ali dvema referencama. Ostale naloge se nanašajo na slovnična in/ali leksikalna vprašanja. Temu načinu dela se želimo izogniti, in sicer tako, da smo v ta namen oblikovali zbirko avtentičnih besedil in da se vaje in naloge nanašajo izključno na besedilne značilnosti posameznega besedila, s čimer želimo uporabniku pomagati pri razvijanju bralne in funkcionalne pismenosti.

### 5.1. Zbirka besedil BERTA

Izhodišče za tvorjenje vaj in nalog področja besedil učnega e-okolja *Slovenščina na dlani* predstavlja zbirka BEsedil pRakTičnega sporazumevanjA BERTA. Zbirka ima dva podkorpusa, in sicer govornega (G-BERTA), ki zajema 59 besedil, in pisnega (P-BERTA), ki zajema 216 besedil, skupno 275 besedil. V zbirko vključena besedila so skrbno izbrana. Pred začetkom zbiranja smo namreč

[6] Npr. pri vaji, kjer je izločena ena sestavina frazema in mora uporabnik med štirimi ponujenimi odgovori izbrati ustreznega (primer: *narediti iz muhe …* – ob ustrezni sestavini *slona* z avtomatskim iskanjem težko dobimo relevantne alternativne odgovore, ki zagotavljajo tudi ustrezno zahtevnost naloge).

pregledali učne načrte in učbeniško gradivo ter popisali besedilne skupine, ki so del učne snovi od šestega razreda osnovne šole dalje. Med popisanimi besedilnimi skupinami smo izbrali npr. novico, vremensko napoved, telefonski pogovor, razgovor, govorni nastop, prijavnico, pritožbo, življenjepis, opis postopka, opis življenja osebe, pripoved o življenju osebe, poljudnoznanstveni prispevek. Izbiri besedilnih skupin je sledil izbor tematike oziroma tematskih sklopov. Izhajali smo iz sporazumevalnih tem za poučevanje slovenščine ali katerega koli drugega jezika kot drugega jezika (SEJO, 2010), npr. osebna identiteta, zdravje, odnosi z drugim, izobraževanje, storitve. Skladno s tem smo si zamislili slogane kot načine dela v e-okolju in zbirali avtentična besedila. Prednost smo dajali besedilom, ki so prototipični predstavniki izbranih besedilnih skupin in ki nastajajo v za večino govorcev slovenščine pomembnih ali priljubljenih, v vsakem primeru pa živih in aktualnih komunikacijskih položajih. Z vsebinskega vidika smo zajeli besedila tistih sporazumevalnih tem, ki so po načinu obravnave teme blizu šolajoči se mladini od 11. do 19. leta starosti.

Besedil torej nismo združevali le po besedilnih skupinah – nabor besedilnih skupin smo omejili na 26, kar ni veliko, a je za prvi tovrstni način dela dovolj –, temveč tudi po tematiki. Podobne tematike, npr. *agrumi, češnja, čokolada, sivka, paradižnik,* smo povezali v enotne sklope, naslovljene s slikovitimi slogani, npr. *Lačen kot Lakotnik*. Te pa smo združili po dveh merilih, in sicer glede na tematski sklop (*narava*) in glede na razred/letnik, za katerega predvidevamo, da je besedilo primerno za obravnavo. Združevanje besedil hkrati po kriteriju besedilne skupine in tematike učečemu se omogoča, da razlike med besedili iste tematike doživi in dojame kot posledice različnih kontekstov tvorjenja obravnavanega besedila. Učeči se namreč znotraj istega slogana rešuje vaje in naloge za isto tematiko, npr. tematika agrumi je predstavljena kot novica (*Veste, zakaj se pomaranče vedno prodajajo v rdečih mrežastih vrečkah?*) ali kot opis postopka (*Solata z agrumi*). To izhodišče je zbiranje gradiva precej upočasnilo, saj je bilo treba tako rekoč za vsako besedilo pridobiti soglasje avtorja. Če soglasja nismo uspeli pridobiti, je to pomenilo spremembo koncepta včasih tudi celotnega slogana, npr. glede na zastavljena merila smo našli ustrezno besedilo k nekemu sloganu, vendar za to besedilo nismo uspeli pridobiti soglasja avtorja ali več avtorjev znotraj sklopa. V takih primerih smo bili primorani preoblikovati slogan, kdaj neuporabiti tudi besedila, za katera smo sicer soglasja pridobili.

Da bi bila obravnava besedila kar najbolj podobna interpretaciji besedil, torej spontanemu razbiranju bistvenih in izločanju manj pomembnih podatkov, da lahko iz besedila dojamemo njegov smisel in namen, je bilo treba besedila prikazati v realnem kontekstu, torej ustrezno večpredstavnostno (ali multimodalno). To omogoča in zagotavlja le celotni prikaz besedila (npr. pdf-posnetek besedila), pri čemer so vidni tudi neverbalni elementi besedila (oblikovanje, grafični elementi, video/avdio itd.). Poleg primernosti in zanimivosti tematike ter besedilne skupine za določeno starostno skupino uporabnikov e-okolja je bilo treba upoštevati tudi dolžino besedila. Ta so obsegovno zelo različna: od nekajvrstičnih šal in tudi do petnajststranskih poljudnoznanstvenih prispevkov. Tako obsežna besedila pri vnašanju nismo delili na odlomke, saj uporabnik sicer ne bi dobil informacij o obsegu besedila kot eni od lastnosti besedilne skupine.

Pri definiranju natančnejšega seznama besedilnih skupin je bilo treba najprej opredeliti različne termine za poimenovanje in pojmovanje množice besedil. V pregledanem osnovno- in srednješolskem gradivu se namreč za poimenovanje množice besedil uporablja termin besedilna vrsta. Besedilne vrste, kot jih razumemo tukaj, se, upoštevaje večrazsežnostni model opredelitve besedila, »konstituirajo prek prototipičnih ena z drugo povezanih kontekstualnih in strukturnih značilnosti. Oblikujejo okvir za prototipične na osnovi konvencij jezikovnih uporabnikov temelječih jezikovnih/govornih vzorcev s karakterističnimi funkcijskimi, medijsko-položajnimi in tematskimi značilnostmi kot tudi glede na te značilnosti ustrezno obliko strukture besedila« (Gansel, Jürgens, 2007). Besedilne vrste so produkti konvencionalnih jezikovnih dejanj znotraj posameznega komunikacijskega področja (Heinemann, 2000). Tovrstna opredelitev besedilne vrste je pokazala, da v učnih procesih učeči se pravzaprav ne spoznavajo značilnosti besedilnih vrst ali le redko, temveč v veliki večini primerov spoznavajo značilnosti besedilnih tipov, npr. prošnja, zahvala, voščilo. Besedilni tip je skupina besedil s skupnimi ali podobnimi le jezikovnimi značilnostmi in je vir za oblikovanje besedilnih vrst. Da bi se izognili nejasnostim in zmedi, smo bili prisiljeni vpeljati termin, ki vključuje tako besedilne tipe kot vrste, to je termin besedilna skupina.

## 5.2. Definiranje vaj

Ker niti dve besedili nista enaki, četudi obravnavata isto tematiko, je skladno s tem treba oblikovati tudi vaje in naloge. Zato smo najprej popisali tipe nalog, ki so primerni za učno e-okolje in omogočajo tvorjenje več različnih nalog s pomočjo enega vnosa.

Prvi sklop nalog je vezan na zunajjezikovne kontekstualne spremenljivke, to so podatki o udeležencih, njihovem medsebojnem družbenem razmerju, kraju in času tvorjenja besedila, naslovu, mestu objave, dostopnosti besedila za širši krog naslovnikov in kanalu, ki so skupaj s tematskim sklopom, besedilno skupino, temo in jezikovno zvrstjo bili že popisani pri vnašanju besedila v vmesnik BERTA.

Temu sledi vnos dodatnih informacij, ki se uporabijo pri tvorjenju nalog, vezanih na vsebino vsakega posameznega besedila, kamor sodijo tudi naloge o sporočevalnem namenu in slogovnem postopku, in nalog, vezanih na značilno leksiko ali druge besedilotvorne značilnosti besedila.

Poseben sklop nalog je vezan na značilnosti besedila kot predstavnika besedilne skupine. Večina navedenih nalog bo vsaj deloma avtomatizirana.[7] Za avtomatsko generiranje besediloslovnih nalog besedilom pripisujemo že zgoraj navedene informacije, npr. informacije o udeležencih, temi, sporočevalnem namenu, slogovnem postopku, primernosti besedila za uporabnika določenega razreda oz. letnika, in druge informacije, npr. pravilne in napačne trditve o vsebini konkretnega posameznega besedila.

[7] Priprava vaj iz besediloslovja je aktivnost, ki je v teku in se zaključi leta 2021.

Še en prav tako poseben sklop pa predstavljajo naloge tvorjenja besedil, pri katerih želimo, da uporabnik povzame, obnovi izhodiščno besedilo ali tvori besedilo iste ali podobne besedilne skupine.

### 5.3. Programiranje vaj

Tudi pri vajah iz besediloslovja sistem izvaja tri naloge: izbor ustreznih besedil, tvorjenje naloge ter vrednotenje uporabnikovega odgovora.

Pri izboru ustreznega besedila upoštevamo primernost besedila glede na starost uporabnika, ustreznost besedila za izbrano nalogo ter vsebnost vseh potrebnih informacij, ki jih potrebujemo za tvorjenje naloge.

Vaje iz besediloslovja se pogosto nanašajo na pomen obravnavanega besedila. Računalniški algoritmi imajo težave z razumevanjem pomena besedila, zato takšnih nalog ne moremo povsem avtomatsko tvoriti. Zbirko besedil smo zato opremili z dodatnimi informacijami, ki smo jih vnesli ob pregledu besedil. Računalnik na podlagi teh informacij lahko tvori različne tipe nalog. Če npr. pri vsakem besedilu zapišemo nekaj vprašanj in pravilnih odgovorov, lahko potem računalnik tvori naloge tipa 'Odgovori na vprašanje', 'Izberi pravilen odgovor' ali celo uporabi več vprašanj in tvori nalogo 'Poveži vprašanja s pripadajočimi odgovori'. S takim polavtomatskim pristopom smo prihranili čas, ki bi ga sicer potrebovali za ročno tvorjenje nalog in hkrati povečali nabor vaj, ki se nanašajo na pomen zapisanega ali govorjenega besedila.

Pri izpisu nalog v spletnem uporabniškem vmesniku smo poskrbeli za preglednost naloge in obravnavanega besedila. Uporabniku je v vsakem trenutku reševanja omogočen vpogled v izhodiščno besedilo. Hkrati smo naloge opremili z namigi in s sklici na bazo znanja, ki so uporabniku v pomoč pri reševanju.

Vrednotenje pravilnih odgovorov poteka enako kot pri ostalih sklopih vaj. Pri nalogah, ki predvidevajo daljši vnos besedila, upoštevamo dva scenarija. Če smo za nalogo vnesli primer pravilne rešitve, lahko z računalniškimi algoritmi izračunamo približno oceno uporabnikovega odgovora. Poleg tega mu lahko ponudimo možnost, da se sam ovrednoti na način, da mu izpišemo predviden pravilni odgovor. Če predvidenega pravilnega odgovora nismo vnesli, lahko uporabnikov odgovor ovrednoti drug uporabnik, običajno v vlogi učitelja ali tutorja. Izbiro načina ovrednotenja v takšnih primerih prepuščamo uporabniku, to je običajno učitelj, ki dodeli naloge v reševanje učencem.

## 6. Vključevanje razlag

Ob izbranih vajah vseh štirih vsebinskih sklopov bodo uporabnikom na voljo krajše sprotne razlage in dostop do baze znanja, kjer bodo obširneje predstavljene izbrane jezikovne teme.[8]

Osnovni cilj vključevanja razlag je takojšnja podpora pri reševanju vaj in nalog ter s tem povezano doseganje nemotenega uporabnikovega napredovanja. Sprotne razlage ob reševanju nalog bodo namreč uporabnikom omogočale, da se zaradi težav pri reševanju naloge ali nepoznavanja konkretnega pravopisnega, slovničnega, frazeološko-paremiološkega ali besedilnega problema seznanijo z vsebino, ki bo nazorno in razumljivo pojasnila izpostavljene jezikovne teme. Če bo uporabnik želel ponujene vsebine še dodatno raziskati, bo lahko po njih brskal tudi prek posebnega zavihka v e-okolju, ki bo opravljal funkcijo jezikovnega priročnika.

Razlage[9] bodo prilagojene zmožnostim oz. predznanju učečih se in bodo smiselno vključene k pripravljenim vajam in nalogam. Da bi se približali celotni ciljni publiki, tj. učencem med 11. in 19. letom, bodo razlage razumljive, preproste in opremljene z verodostojnimi in nazornimi primeri. Uporabnik bo lahko priklical razlago kadarkoli med reševanjem nalog s klikom na ustrezno ikono. Sprotna razlaga ob reševanju nalog in vaj se bo prikazala najprej kot povzetek oziroma kratka razlaga v povedi ali dveh. To bo uporabniku pomoč, na kaj mora biti pri nalogi najbolj pazljiv ali da si na hitro razjasni največjo zagato pri reševanju. Če mu to ne bo zadostovalo za uspešno reševanje naloge, se mu bo s klikom na gumb *Preberi več* odprla daljša razlaga, ki bo podrobnejša, obširnejša, po potrebi opremljena s slikovnimi elementi in tabelami ter primeri. Ta razlaga bo zadostovala za uspešno rešeno nalogo in seznanitev z najpomembnejšimi jezikovnimi vsebinami pri nalogi. Če bo uporabnik o ponujeni temi želel izvedeti še več ali ga bodo zanimale povezane jezikovne vsebine, bo s klikom na gumb *Preberi še* preusmerjen k poglavjem, kjer bodo povezane izbrane vsebine predstavljene še bolj celostno, ali s klikom na dodane povezave k drugim relevantnim spletnim jezikovnim priročnikom in virom.

Za vnos razlag smo pripravili poseben vmesnik, v katerega bomo vnesli več kot 200 razlag, ločenih v sprotne razlage za pomoč pri izvajanju vaj ter združene v poglavja baze znanj Orehek. Poleg razlag, ki jih bomo pripravili avtorji e-okolja Slovenščina na dlani, pa bodo razlage lahko vnašali tudi uporabniki sami. Teme teh razlag bodo definirane vnaprej in povezane s posameznimi vajami in nalogami, njihovo vsebino pa bo izpolnil uporabnik učenec in pregledal in popravil učitelj in/ali član projektne skupine Slovenščina na dlani, preden bo vidna ostalim uporabnikom.

## 7. Zaključek

V projektu Slovenščina na dlani s pomočjo jezikovnih virov in avtomatiziranih postopkov izdelujemo interaktivno učno e-okolje za podporo učenju slovenščine v osnovnih šolah od 6. razreda naprej in v srednjih šolah. Za potrebe tega e-okolja smo razvili manjši korpus besedil, primernih za mladino (MAKS), večmodalno zbirko besedil praktičnega sporazumevanja (BERTA), slovar frazemov in pregovorov (FRIDA) in bazo razlag (Orehek). Izdelani viri bodo na voljo tudi kot jezikovni vir pod licencami CC BY, razen v primerih, kjer to ni mogoče zaradi omejitev pri izvornih avtorskih pravicah besedil oz. vsebin. Dostopni so oz. bodo prek repozitorija CLARIN.SI oz. prek Clarinovih konkordančnikov (https://www.clarin.si/noske/).

Do oktobra 2020 je pripravljenih okrog 1000 različnih vaj s področja pravopisa, slovnice in frazeologije ter

[8] V slovenskem prostoru je podoben princip že uporabljen v Pedagoškem slovničnem portalu, ki je nastal v okviru projekta Sporazumevanje v slovenskem jeziku (prim. http://slovnica.slovenscina.eu/).

[9] Aktivnost priprave razlag še ni zaključena.

besedil, izdelane pa bodo tudi vsebinske razlage za več kot 200 različnih tematik. Za vsako vajo je na voljo večje število različnih primerov, pri pravopisu in slovnici tudi do 500 za eno vajo.

V šolskem letu 2020/21 nadaljujemo z zadnjo fazo izgradnje vaj, tj. s priklicem vaj iz zalednih baz v spletni vmesnik e-okolja, dodajanjem besedilnih nalog in pisanjem razlag, izdelavo različnih funkcionalnosti, kot so prikazovanje doseženih rezultatov, prikazovanje razlag ob vajah ipd., ter začenjamo testiranje e-okolja in analizo vedenja uporabnikov. S šolskim letom 2021/22 bo e-okolje odprto dostopno zainteresirani javnosti.

Avtomatizacija s podporo jezikovnih tehnologij in digitalno okolje imata nekaj prednosti, ki jih papirni medij ne omogoča: veliko količino primerov in vaj, prilagajanje zahtevnosti vaj znanju uporabnika, avtomatsko ocenjevanje in usmerjanje med vajami, enostavno prikliljivo pomoč v obliki razlag, prilagojenih vsaki posamezni nalogi, podporo pri reševanju z namigi ali sprotno komunikacijo z drugimi uporabniki sistema, motiviranje z grafičnimi elementi, načini točkovanja ali sodelovanjem v skupinah. Ob raziskovanju možnosti za motivacijo in spodbujanje ustvarjalnosti pa smo ugotavljali tudi omejitve v primerjavi z osebnim stikom pri učenju. Ob tem da je učno e-okolje visoko avtomatizirano in individualizirano, namreč omogoča veliko manj ustvarjalnih in interaktivnih nalog. Avtorji e-okolja Slovenščina na dlani zato že od same zasnove okolja naprej sledimo načelu, da je digitalno učno e-okolje koristno dopolnilo in popestritev pouka, nikakor pa ne nadomestilo za tradicionalne oblike poučevanja.

## 8. Literatura

Christina Gansel in Frank Jürgens. 2007. *Textlinguistik und Textgrammatik. Eine Einführung.* 2. Auflage. Vandenhoeck & Ruprecht, Göttingen.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: *Zbornik Osme konference Jezikovne tehnologije*, str. 82–87. (http://nl.ijs.si/isjt12/JezikovneTehnologije2012.pdf)

Wolfgang Heinemann. 2000. Textsorte – Textmuster – Texttyp. V: K. Brinker, G. Antos, W. Heinemann, S. F. Sager, ur., *Text- und Gesprächslinguistik: ein internationales Handbuch zeitgenössischer Forschung*, str. 507–523, Handbücher zur Sprach- und Kommunikationswissenschaft, Zv. 16. Walter de Gruyter, Berlin, New York.

Vida Jesenšek in Natalija Ulčnik. 2014. Spletni frazeološko-paremiološki portal: redakcijska vprašanja ob slovenskem jezikovnem gradivu. V: V. Jesenšek, S. Babič, ur., *Več glav več ve: Frazeologija in paremiologija v slovarju in vsakdanji* rabi, str. 276–292, Oddelek za germanistiko, Filozofska Fakulteta Univerze v Mariboru in ZRC SAZU Ljubljana, Inštitut za slovensko narodopisje.

Iztok Kosem, Mojca Stritar, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt, Tadeja Rozman. 2012. *Analiza jezikovnih težav učencev: korpusni pristop*. Trojina, zavod za uporabno slovenistiko, Ljubljana.

Mira Krajnc Ivič. 2020. Obravnava besedil: merila za razlikovanje med besedilno vrsto in besedilnim tipom. *Slavistična revija*, 68(1):55–71. (https://srl.si/ojs/srl/article/view/2020-1-1-4)

Erika Kržišnik. 2006. Izraba semantične potence frazemov. *Slavistična revija*, 56(1):259–279.

Matej Meterc. 2017. *Paremiološki optimum: Najbolj poznani in pogosti pregovori ter sorodne paremije v slovenščini*. Založba ZRC, ZRC SAZU, Ljubljana.

*Pedagoški slovnični portal.* Sporazumevanje v slovenskem jeziku (http://slovnica.slovenscina.eu/).

Jenny Rose Finkel, Trond Grenager in Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. V: *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, str. 363–370. (http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf)

Tadeja Rozman, Irena Krapš Vodopivec, Mojca Stritar, Iztok Kosem. 2020. *Empirični pogled na pouk slovenskega jezika.* Znanstvena založba FF UL, Ljubljana. (https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/227/327/5303-1)

*SEJO = Skupni evropski jezikovni okvir: učenje, poučevanje, ocenjevanje.* Irena Kovačič (pr). El. knjiga. Ljubljana: Ministrstvo RS za šolstvo in šport, Urad za razvoj šolstva, 2011. (https://centerslo.si/wp-content/uploads/2015/10/SEJO-komplet-za-splet.pdf).

Peng Qi, Timothy Dozat, Yuhao Zhang in Christopher D. Manning. 2018. Universal Dependency Parsing from Scratch. V: *Proceedings of the CoNLL 2018 UD Shared Task*, str. 160–170. (https://www.aclweb.org/anthology/K18-2000.pdf)

Natalija Ulčnik. 2019. Izbor frazemov za bazo FRIDA. V: N. Ulčnik, ur., *Slovenščina na dlani 2*, str. 37–45, Univerzitetna založba Univerze v Mariboru. (https://press.um.si/index.php/ump/catalog/book/447)

Natalija Ulčnik in Matej Meterc. 2019. Izbor pregovorov za bazo FRIDA. V: N. Ulčnik, ur., *Slovenščina na dlani 2*, str. 47–55, Univerzitetna založba Univerze v Mariboru. (https://press.um.si/index.php/ump/catalog/book/447)

# Selected methodological challenges of sign language lexicography

**Lucia Vlášková*,†, Hana Strachoňová*,†**

* Support Centre for Students with Special Needs (Teiresiás), Masaryk University
Komenského náměstí 2, 602 00 Brno, Czech Republic
† Department of Linguistics and Baltic Languages, Faculty of Arts, Masaryk University
Arna Nováka 1, 602 00 Brno, Czech Republic
l.vlaskova@gmail.com
hana.strachonova@gmail.com

**Abstract**

Sign language lexicography, as a growing field of study within sign language linguistics, faces many challenges that have already been answered for the audio-oral language material. In this paper, we present some of these challenges together with methods developed to help navigate the often problematic field of lexical classification. The described methods and strategies are implemented in the first Czech sign language online dictionary, a part of the platform *Dictio*. Among the covered topics are questions of lemmatisation and how to decide what constitutes a lexeme in sign language; what is the place of classifier constructions in a dictionary, given their peculiar semantic status; what is the proper treatment of mouthings and mouth gestures concerning citation forms and derivation; why is it difficult in sign language to distinguish synonyms from variants and how can our proposed phonological criteria help; and, finally, what kinds of utterances are considered good examples of use.

## 1. Introduction

*Dictio* is a multilingual online dictionary that includes multiple languages, both sign and spoken. This ongoing project is being realised at the Masaryk University in Brno, Czech Republic. At the moment, it includes entries for the following languages (ordered by the approximate number of entries): Czech (120 thousand), Czech Sign Language (13 thousand), Slovak Sign Language (7 thousand), Slovak (7 thousand), English (5 thousand), Austrian German (800), Austrian Sign Language (650), International Sign (170), and American Sign Language (120). Only a part of the entries is published, the rest is still a subject of the editing work of multiple working groups, including international teams of Deaf university employees. At the moment (August 2020), the numbers of the published entries for the sign languages are as follows: Czech Sign Language – 3670, Slovak Sign Language – 130, International Sign – 100, American Sign Language – 80, and Austrian Sign Language – 0.

The field of sign language lexicography is growing rapidly. Considering Stokoe's (1960/2005) description of ASL lexical units as the pioneering work, which respect the established linguistics principles, sixty years later, we make use of systematised databases for a whole range of sign languages in the form of printed books or off-line and on line databases (see the overview in McKee and Vale, 2017 or Fenlon et al., 2015). Since the seminal work of Johnston and Schembri (1999) on lemmatisation of the AUSLAN corpus (and closely connected AUSLAN lexical database), several researchers have published their experiences in the form of applicable universal guidelines for the lexicographic work on any sign language. At the moment, many topics concerning mainly the electronic lexical databases are addressed in the literature: e.g., history and options of the sign description and search (Zwitserlood, 2010, focusing on dictionaries of NGT), lexicographic specifics of sign languages compared to spoken languages (Kristoffersen and Troelsgård, 2012, with particular focus on the DTS lexical database), phonological and morphological variation in the process of lemmatisation (Fenlon et al., 2015, on the material of BSL), and others. At the beginning (around 2009), our project was inspired mainly by the work of Johnston and Schembri (1999) and on-line public dictionaries that included semantic definitions in sign language (e-LIS – the dictionary of LIS; and Elix – the dictionary of LFS).

In the absence of a representative Czech SL corpus, the linguistic material for the Czech SL part comes from two primary sources: published dictionaries and Czech SL informants. *Dictio* has the ambition to collect all the published Czech SL dictionaries and make them available in one database. That covers printed books (mainly Potměšil, 2002; 2004a; 2004b), CDs (Langer, 2005a; 2005b; 2008, a.o.), and other individual projects (e.g., diploma theses focusing on specific semantic fields, teaching materials for Czech SL commercial or university courses). The collection of previously published material is being edited, annotated and completed by the team of native signers of Czech SL. The substantial part of the team's work is to discuss synonyms and variants for the published items. This way, plenty of new material is being elicited for the *Dictio* database.

In this paper, we introduce selected topics from sign language lexicography. The idea is to describe some linguistic issues we have encountered while working on the material for the Czech SL part of the dictionary and propose a guideline that could be applicable in the field of sign language lexicography in general. Czech SL was the first language introduced into the dictionary. Creating the linguistic methodology has been especially challenging since the original vision of the whole project was to build the first monolingual dictionary, in this case, a dictionary of Czech SL, where the meaning and the use of the signs are explained and illustrated solely in Czech SL. Such entries, then, include the links to the parts containing other languages (translations). However, this paper is focused on the internal classification of Czech SL material and does not cover translations between spoken and sign languages.

## 2. Lemmatisation and types of dictionary entries

The most fundamental question to be raised when composing a sign language dictionary is what kind of signs to include, i.e., what constitutes an entry in a dictionary. To answer this question, the following strategy has been developed: first, we take all the possible kinds of signs occurring in natural speech (lexeme, deixis, description, compound, collocation, set phrase) and divide them into two groups according to their complexity: the ones that do not consist of individual semantic units (lexeme, deixis) and the ones that do (description, collocation, compound, set phrase). Among the first group, we set aside the expression, the meaning of which changes according to the referent (deixis) and select the expression with a conventionally established meaning (lexeme). From the second group, we single out the expressions with a non-compositional meaning, i.e., the set phrase and the compound. Thus, this strategy leaves us with only three candidates for a dictionary entry: a traditional lexeme, a compound, and a set phrase, with each of their meanings conventionally established.

In our dictionary, however, we classify signs into four types of entries: simple signs, compounds, set phrases, and derivatives.

Simple signs are lexemes with one root, one handshape, and one place of articulation. Thus, the Selected Finger Constraint and Place Constraint hold (Sandler, 2006).[1]

Compounds are morphologically complex signs that originated by merging two independent signs, i.e., two free morphemes. From the semantic point of view, compounds are not bound to introduce a new meaning, as seen in the Czech SL example of SUNGLASSES = SUN^GLASSES. Nevertheless, it is possible, e.g., FLOWER^SPRING = MAY (Mladová, 2009). It is often difficult to distinguish compounds from set phrases, the third type of entries in our dictionary. Set phrases also consist of two (or more) free morphemes, but their meaning is not compositional, e.g., in Czech SL sign UNIVERSITY, which consists of the signs HIGH and SCHOOL. However, in the case of compounds it is not the semantic shift that classifies them as such but the phonological reduction/assimilation, as defined by Zeshan (2004): the first sign is shortened and loses stress, any repetitions and internal movements are deleted, handshape and location can be assimilated, and the passive hand can function as a place of articulation.[2] On the other hand, no such modification can be found in set phrases, where all constituting signs are fully realised.

The last type represented in our dictionary are the derivatives, defined as forms that have been derived from their respective motivating signs by means of adding or changing a non-manual component, specifically mouthing or mouth gesture. Typically, this process occurs while deriving a technical or more specific term from a general vocabulary sign. Take an example from Czech SL where SACCHARIDE is derived from SUGAR. These two signs have the same manual component but differ in mouthing. SUGAR is standardly articulated without mouthing, and SACCHARIDE contains the mouthing of the Czech word *sacharid*, proving mouthing to be of a significant lexical role (Sandler, 2006).[3] We will cover mouthing and mouth gestures in more detail in Section 4.

Another important question is the choice of a citation form of each entry. Following Johnston and Schembri (1999), only the unmodified signs in their basic forms are present in the lexicon (and, therefore, the dictionary), inflection and modification is part of the grammar. Modification can take several forms, as defined in Zeshan (2003): (i) modified movement expresses the change in aspect, number, degree or directionality (verbal inflection encoding the subject and/or the object of the given verb); (ii) modified handshape signals classifier constructions and numeral incorporation; (iii) modified facial expressions distinguish between indicative, interrogative, negative and other clause types (Zeshan, 2004). In our dictionary, the information whether a sign can incorporate numerals, subject and/or object, and other modifiers is given in the grammatical part of the dictionary entry for the corresponding lemma in its basic form, i.e. singular, non-modified and non-intensified sign.[4] The basic form for signs that incorporate a numeral is the one with incorporated ONE. For directional signs, it is the form directed from the speaker to the addressee.

However, there are exceptional cases when the dictionary also covers other than basic forms of signs. Such instances include deixis with fixed initial position, e.g., the pronouns I and MY that are always signed facing the speaker, and, correspondingly, YOU and YOUR, always facing the addressee. Furthermore, lexicalised forms of different types have their place in the dictionary, e.g., lexicalised deixis. Take the Czech SL verb HEAR, which is realised by pointing to the speaker's ear with a crooked index finger. Normally, this would be a deictic sign meaning *ear*, which does not belong in the dictionary, but slight modifications (different facial components – raised eyebrows, added mouthing) indicate that it is lexicalised. A semantic shift occurred where the meaning is no longer the object at which the finger is pointing but the activity of hearing. Other forms of lexicalisation include lexicalised classifier constructions, which we will discuss in the following section, or lexicalised fingerspelling, as is the case with the sign for *engineer* – ING, fingerspelled with the letters of the Czech SL alphabet.

## 3. Classifiers, specifiers and lexicalised constructions

Classifiers have repeatedly proven to be an exciting research topic among sign linguists. This section will focus on different types of classifiers, a closely related group of specifiers, and the ways of properly incorporating them into a dictionary.

Sign language classifiers are considered a special kind of morphemes representing nominals, the meaning of which is not precisely specified. They denote relevant

[1] Consequently, items that, synchronically, do not consist of two lexical roots but have two different places of articulation are classified as compounds. See, e.g., the sign POST_OFFICE in the appendix, with the movement realised from the head towards the non-dominant hand. Note that most of the following examples will also be provided in the appendix.

[2] At least one of the reduction/assimilation patterns must be present to classify the item as a compound.

[3] More precisely, the sign for SUGAR may be accompanied by the mouthing of the Czech word *cukr* (sugar), but the sign for SACCHARIDE must be articulated with the mouthing of *sacharid*.

[4] See, e.g., the entry for HOUR, unfortunately, with the commentary in Czech only.

properties of entities via different configurations of the manual articulator (Zwitserlood, 2012), specify shapes and dimensions of objects, and denote spatial relations and motion events (Sandler and Lillo-Martin, 2006b). Such entities are then categorised according to the properties into groups, e.g., flat objects, long and thin objects, two-legged beings, etc. Classifiers have been attested in all known sign languages (Sandler and Lillo-Martin, 2006b), thus constituting a stable class with common general attributes, although the inventory of the particular classifiers differs from one language to another (Zwitserlood, 2012).

The categorisation of different types of classifiers has been a subject of much discussion. Earlier literature (Supalla, 1986, a.o.) had divided them into multiple classes based on various characteristics (e.g., semantics, shape, function, animacy) before currently stabilizing on two main ones: whole entity classifiers and handling classifiers, based more on their function in grammar rather than their semantic properties (Zwitserlood, 2012).

Whole entity classifiers denote their referents in their entirety. They are more abstract and "refer to general semantic classes rather than to visually perceived physical properties" (Sandler and Lillo-Martin, 2006b:77). However, various classifiers can be used to denote a single entity, each highlighting a different relevant aspect (Zwitserlood, 2012).

On the other hand, handling classifiers utilize iconicity on a larger scale; they indicate the shape of an entity as it is being held or manipulated with, while the manual articulator represents itself – a hand holding the entity. This strategy gives the speaker much more room to choose among different classifiers according to the situation in the actual world (Zwitserlood, 2012).

These types of classifiers, considered to be bound morphemes, must occur jointly with other expressions within so-called classifier constructions, within which they are incorporated mostly into classifier verbs, i.e. verbs denoting movement, position or existence of a referent in space or some kind of manipulation (Zwitserlood, 2012). Classifier constructions represent a very productive strategy in sign languages, and this unstable semantic and morphological status prevents them from being documented in a dictionary.

However, classifiers outside of classifier constructions (so-called classifier handshapes) can be documented. In our dictionary, classifier handshapes are registered in individual lexical entries if there is a (relatively neutral) stabilised representative form with (at least roughly) delimited meaning (e.g., via extensional definition by listing possible referents, see Section 5).

An example of such a classifier handshape from Czech SL is a basic, unmarked handshape – an open palm with all fingers stretched out. In the grammar part of this entry, the sign is categorised into its classifier group, whole entity classifiers, below which two meanings are listed: a denotation of either flat objects or four tired vehicles. Consequently, there are definitions and examples of use listed for each meaning separately; in this case a sentence where the classifier denotes a book in the former, and a car in the latter meaning.

Furthermore, there is another type of expression formerly thought of as classifiers: tracing Size and Shape Specifiers (SASSes), "which have a movement of the hand(s) that outlines an entity's size/shape" (Supalla, 1986). Although some linguists still include them into the group of classifiers, the standard conclusion is to keep them separate, based on their multiple differing traits (presence of movement, not just a handshape; no need for antecedent, syntactically can function as nominals, adjectives or adverbs) (Zwitserlood, 2012). Specifiers are not incorporated into verbal predicates; instead, they compound with nominals and usually serve a modifying function (Sandler and Lillo-Martin, 2006b:77).

For a specifier to be registered as a separate entry in our dictionary, the same criteria apply as those for classifiers; a stabilised representative form with a roughly delimited meaning has to be attested.[5]

However, when classifiers and specifiers are part of lexicalised classifier constructions, they are automatically included in a dictionary and are treated as lexemes. In these structures, the otherwise productive forms become "frozen", their individual features (handshape, movement, location) no longer contribute morphological content to the given expression but bear only phonological status (Sandler and Lillo-Martin, 2006a). Among such constructions in Czech SL are, e.g., signs BOW (≈ ARCHERY) and TREE, which originated by lexicalising a classifier, or YOGURT and OMELETTE, in which a former specifier is lexicalised.

## 4. Mouth patterns accompanying signs

Non-manual components of signs defined as "all linguistically significant elements that are not expressed by the hands" (Pfau and Quer, 2010) are equally important for speech comprehension and production as the manual articulators. These components can take the form of head and body movements, facial expressions, or mouth patterns. In this section, we will focus on the last type and assess which mouth patterns should and should not be documented in a dictionary.

Mouth patterns are commonly divided into mouth gestures and mouthings, differing in their relationship to the surrounding spoken language.

Mouthings (or spoken components) are either influenced or directly derived from the corresponding word in the spoken language; they are silent articulations of the whole word or a part of it, usually its first syllable (Pfau and Quer, 2010). Mouthings are understood as cross-modal borrowings (Sandler and Lillo-Martin, 2006a; Mareš, 2011), where it is possible to observe a gradual change and adaptation to the "host" language, a process typical for borrowings observed among spoken languages.

Mouth gestures (or oral components), on the other hand, are defined as "all motions/positions of the mouth that are not derived from a spoken language and contribute to the speech structure" (e.g., Mareš, 2011:8) and are therefore considered a native component of sign language.

In order for mouth patterns to be included in our dictionary, they need to satisfy two conditions: (i) they are obligatory for the given sign and (ii) they do not introduce additional meaning, they do not modify the sign in terms of intensification, adjectival or adverbial modification, nor do they express speaker's attitude (Mareš, 2011:24; Pfau and Quer, 2010:385). Condition (ii) does not apply to the mouth patterns (whether mouth gestures or mouthings) that have a disambiguating function between signs with identical or similar manual components, e.g., in the case of (a) two general signs: SALT vs. PEPPER, each with different mouthing of the corresponding Czech word *sůl* and *pepř*, respectively; (b) a general sign and a specific term: SUGAR vs. SACCHARIDE covered above in Section 2; and

[5] See the appendix for an example.

(c) a noun and a verb: CAR vs. DRIVE, where CAR is articulated with the mouthing of the corresponding Czech word *auto*, while DRIVE is performed with a mouth gesture.

In the case of a single sign (conveying a single meaning) with variable mouth patterns available, the citation form is accompanied by the most neutral one, usually a mouth gesture. The other options are classified as variants of that sign. Any obligatory mouth patterns are given in the grammatical description for each meaning of the lexical entry (corresponding Czech word for mouthings and specialised symbols for different mouth gestures), and information about the association with another sign in a noun-verb pair.

## 5. Strategies of semantic definitions

So far, we have discussed what kinds of lexemes are eligible to be listed in a dictionary, but let us now turn to each lexical entry structure with a particular focus on their definitions.

The definition of a lexical entry is a crucial part of any monolingual dictionary. Thus it is of great importance to develop a firmly established method before beginning any lexicographic work, and to adhere to it throughout the process of compiling a dictionary. It can be especially challenging in the case of sign language dictionaries, where there is very little prior work to build on, and one may encounter several unprecedented issues.

In *Dictio*, we face these challenges with the help of precisely determining processes for forming each definition.

There are two main strategies for defining the meaning – intensional and extensional definition (Filipec, 1995). To define a lexeme intensionally means to specify necessary and sufficient conditions for the usage of a given lexeme. Such intensional definition has the following structure: first, the closest general term, a hypernym, is posited to categorise the lexeme into a broader semantic class; the next step is to list necessary distinguishing properties in order to differentiate the lexeme from other elements of the same semantic class. This way, we delimit all potential occurrences while ruling out other cases.[6] A nice example of the application of this general lexicographic strategy is the definition of the sign CD-ROM, which is given here in glosses and can be seen under the link: CD-ROM THAT CLF_ROUND_OBJECT SPC_THIN_OBJECT THAT SAVE DATA HOW CLF:DRAW_CIRCLES APPEAR SPC_LITTLE_HILLS 0 1 0 1.

Extensional definitions employ a different strategy, namely that they specify an extension of a given lexeme, e.g., by naming a typical representative or several objects that are members of a specific set, requiring the reader to extract the properties common to all listed examples and compile the meaning of the lexeme from them. Such a definition can be accompanied by qualitative or circumstantial properties of a concept, e.g., size, color, or application. An example is the semantic definition of the sign BLACK, which is given here in glosses and can be seen under the link: COLOR THAT LOOK_LIKE SUN GO_DOWN GET_DARK THAT.

Between the two strategies, it is always preferred in our dictionary to use the intensional definition. However, in sporadic cases, the meaning can be determined extensionally or by combining the two, i.e., by specifying a superordinate concept followed by several examples of referents.

## 6. Multiple meanings and semantic relations

In each lexical entry, the field of semantic relations includes both the intra-language relations (synonyms, antonyms), and the inter-language relations (translations). We will comment in detail on the first type, leaving the latter for another occasion. However, let us first consider the cases of polysemy.

In our dictionary, we follow the traditional practice of listing every meaning of a polysemous word under one lexical entry. These individual meanings differ, and therefore separate definitions, examples (and translations) are needed for them.

In principle, we have encountered three types of situations: (i) a general term with multiple meanings (e.g., GERMAN, which may stand for the country or a citizen of the country); (ii) a technical term with different meanings for their respective semantic fields of use (e.g., the sign BASIS with three different meanings – for the field of informatics, mathematics, and chemistry); and (iii) a sign with general and technical use. If the two forms are entirely identical – including the non-manual component – two meanings can be defined with the general one listed as first. However, more often, new mouthing is added during the creation of the technical term. In this case, we understand the non-manual component as a phoneme, and we register each sign under a separate entry.[7]

In *Dictio*, we register synonyms (expressions with identical or nearly identical meaning) and antonyms (expressions with the opposite meaning). A question closely tied to synonyms is how to distinguish them from variants (because, after all, both pairs share the same meaning) and classify synonyms and variants according to their semantic relationship to a given lexical entry.

For audio-oral languages, a dictionary entry standardly contains the citation form of a lexeme and all the variants (Čermák, 1995), e.g., the gender variants in Czech: *brambor* (potato-masculine) vs. *brambor-a* (potato-feminine). However, two (or more) expressions of a different word-forming nature are not considered variants but synonyms (Filipec, 1995), e.g., the Czech pair: *jazykověda* (linguistics; Czech origin) vs. *lingvistika* (linguistics; foreign origin).

What seems like a simple task for spoken languages (basically, common root signals variants, different roots – synonyms) becomes a challenge for sign languages while the discussion about the definition of morphemes and lexical roots is still open-ended (Zwitserlood, 2012). The lexicographic processing of the variants in sign languages has been addressed in the canonical work of Johnston and Schembri (1999) for AUSLAN, although the topic of synonyms is not elaborated.

[6] Since the key to intensional definition is to capture the internal hierarchy of a given semantic area, the work of Půlpánová (2017) on Czech SL becomes useful. In her thesis, she investigated the functional signs used for categorisation in Czech SL. Under *categorisation*, she understands the expression of hyper-hyponymic relations in the lexicon. Such functional signs are, e.g., TYPE and GROUP in her elicited Czech SL expression ANIMAL + TYPE + GROUP + HOME (in the meaning of *pet*).

[7] See Section 4 above, namely examples SUGAR and SACCHARIDE.

In *Dictio*, a method has been developed (and is now being applied) to distinguish variants from synonyms in Czech SL (with possible extension to other sign languages).

We propose to classify a pair of lexemes as variants in case their (possibly multiple) differing phonological features fall within only one of the three main categories (as defined in Sandler, 2006): handshape, place of articulation or movement. Variants altering within the place of articulation (location) are exemplified on a pair of lexemes COFFEE#1 and COFFEE#2, whereas PRAGUE#1 and PRAGUE#2 demonstrate variants with different handshape. On the other hand, the two signs for *kitchen* (see examples KITCHEN#1 and KITCHEN#2) present an apparent synonymy; the lexemes differ in all three main categories, and there is no doubt they do not share a morphological root. However, examples similar to MAY#1 and MAY#2 (which represent two forms from several variants and synonyms for *May*) are challenging since they present two morphologically related forms. Nonetheless, given that they differ in two of three main categories, namely handshape and movement, we conclude that they should be classified as synonyms. More complicated cases show that we are working with a scale rather than a binary distinction. However, the proposed solution appears to be extremely useful for the lexicographic task at hand.[8]

From a formal point of view, variants are presented in the grammatical or stylistic part of an entry, according to their status. A grammatical variant is a lexeme that is freely interchangeable with the citation form which does not add any extra information about the speaker. On the other hand, a stylistic variant adds such information about, e.g., social status, regional categorisation or a generation. However, synonyms are linked to individual meanings and not to the whole lexical entry because they may differ according to their semantic content.

## 7. Examples of use

The final section discusses examples, namely what kinds of expressions are appropriate for an example and what guidelines need to be followed when adding an example to an entry.

In the absence of Czech SL representative corpus, the examples of use are not elicited but created by the team of native signers, forming a small corpus by itself.

It is desirable to include at least one, but ideally, several examples are listed in each lexical entry, demonstrating the use of a given lemma in different communicative situations. An example could be an expression (two or more signs), a sentence, or an utterance (several sentences) illustrating the use of the lemma or its variants.

The fundamental idea of examples is to portray how lexemes are used in natural language. Therefore, it is not unusual to exemplify modification where possible, such as numeral and classifier incorporation, inflection of directional verbs, aspectual modification, and plural and negated forms.

As an illustration of the strategy described above, consider two examples for MONTH. The first example contains the simple citation form, the second one – a pluralized form with an incorporated numeral: (i) TOMORROW MONTH MAY (video under the link), (ii) SUMMER IN_THAT YEAR PERIOD HAVE FOUR MONTH_THREE-redupl THAT SEGMENT IN_THAT JUNE 21 UNTIL SEPTEMBER 22 (video under the link).

## 8. Conclusion

*Dictio* is a work in progress, similar to any other dictionary trying to capture and describe natural language. However, even now, in its developing stages, it already serves multiple functions. *Dictio* has been used in Czech SL and linguistic education, providing valuable examples of signs and their categorisation. Moreover, it represents the biggest Czech SL material collection to date, making it the closest equivalent to Czech SL corpus available. Finally, *Dictio* poses many lexicographic challenges, and solving them brings us closer to understanding the nature of Czech sign language (among others) and its phenomena.

## 9. References

František Čermák. 1995. Paradigmatika a syntagmatika slovníku: možnosti a výhledy. In: F. Čermák in R. Blatná, eds., *Manuál lexikografie*, Vol. 1, pages 14–49. H&H, Jinočany.

Dictio: Multilingual Online Dictionary [online], Masaryk University, Brno [cit. 2020-03-27]. Available at: www.dictio.info.

Jordan Fenlon et al. 2015. The Lemma Dilemma Revisited. *International Journal of Lexicography*, 28(2):169–206.

Josef Filipec. 1995. Teorie a praxe jednojazyčného slovníku výkladového. In: F. Čermák in R. Blatná, eds., *Manuál lexikografie*, Vol. 1, pages 14–49. H&H, Jinočany.

Trevor Johnston in Adam Schembri. 1999. On Defining Lexeme in a Signed Language. *Sign Language & Linguistics*. 2(2):115–185.

Jette Hedegaard Kristoffersen in Thomas Troelsgård. 2012. The Electronic Lexicographical Treatment of Sign Languages: The Danish Sign Language Dictionary. In: S. Granger in M. Paquot, eds., *Electronic Lexicography*, pages 293–318. Oxford University Press, Oxford.

Jiří Langer et al. 2005a. *Znaková zásoba českého znakového jazyka k rozšiřujícímu studiu surdopedie se zaměřením na znakový jazyk (I, II).* Palacký University, Olomouc.

Jiří Langer et al. 2005b. *Znaková zásoba českého znakového jazyka k rozšiřujícímu studiu surdopedie se zaměřením na znakový jazyk (III, IV)*. Palacký University, Olomouc.

Jiří Langer et al. 2008. *Slovník vybraných pojmů znakového jazyka pro oblast biologie člověka a zdravovědy*. Fortuna, Praha.

Jindřich Mareš. 2001. *Orální komponenty v českém znakovém jazyce*. Bc. thesis, Charles University.

Rachel McKee in Mireille Vale. 2017. Sign language lexicography. In: P. Hanks in G. M. de Schryver, eds., *International Handbook of Modern Lexis and Lexicography*. Springer, Berlin/Heidelberg.

Pavlína Mladová. 2009. *Kompozita v českém znakovém jazyce*. Bc. thesis, Charles University.

Roland Pfau in Josep Quer. 2010. Nonmanuals: their grammatical and prosodic roles. In: D. Brentari, ed., *Sign Languages*, pages 381–402. Cambridge University Press, New York.

[8] Our approach builds on the phonological Hand-Tier model of Sandler (2006) and contributes a set of clear criteria for distinguishing variants from synonyms. However, a more detailed discussion and formal analysis of the Hand-Tier model and its modification are beyond this paper's scope.

Miloň Potměšil et al. 2002. *Všeobecný slovník českého znakového jazyka, A-N*. Fortuna, Praha.

Miloň Potměšil et al. 2004a. *Všeobecný slovník českého znakového jazyka, O-Ž.* Fortuna, Praha.

Miloň Potměšil et al. 2004b. *Všeobecný slovník českého znakového jazyka, O-Ž - doplněk.* Fortuna, Praha.

Lucie Půlpánová. 2007. *Kategorizace v českém znakovém jazyce*. Mgr. thesis, Charles University.

Wendy Sandler. 2006. Phonology. In: W. Sandler in D. Lillo-Martin, eds., *Sign Language and Linguistic Universals*, Vol. 1, pages 111–271. Cambridge University Press, New York.

Wendy Sandler in Diane Lillo-Martin. 2006a. Entering the lexicon: lexicalization, backformation and cross-modal borrowing. In: W. Sandler in D. Lillo-Martin, eds., *Sign Language and Linguistic Universals*, Vol. 1, pages 111–271. Cambridge University Press, New York.

Wendy Sandler in Diane Lillo-Martin. 2006b. Classifier constructions. In: W. Sandler in D. Lillo-Martin, eds., *Sign Language and Linguistic Universals*, Vol. 1, pages 111–271. Cambridge University Press, New York.

William C. Stokoe. 1960/2005. Sign language structure: an outline of the visual communication systems of the American deaf. Reprinted in *Journal of Deaf Studies and Deaf Education*, 10:3–37.

Ted Supalla. 1986. The classifier system in American sign language. *Noun classes and categorization*, 7:181–214.

Ulrike Zeshan. 2003. Towards a notion of ‘word’ in sign languages. In: R. Dixon in A. Aikhenvald, eds., *Word: A Cross-linguistic Typology*, pages 153–179. Cambridge University Press, Cambridge.

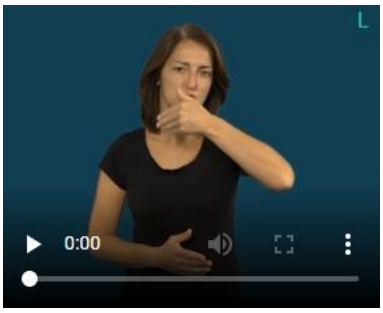
Ulrike Zeshan. 2004. Interrogative and Negative Construction in Sign Languages. *Language*, 80(1):7–39.

Inge Zwitserlood. 2010. Sign Language Lexicography in the Early 21st Century and a Recently Published Dictionary of Sign Language of the Netherlands. *International Journal of Lexicography*, 23(4):443–476.

Inge Zwitserlood. 2012. Classifiers. In: R. Pfau, M. Steinbach in B. Woll, eds., *Sign Language. An International Handbook*, pages 158–186. De Gruyter Mouton, Berlin/Boston.

# 10. Appendix

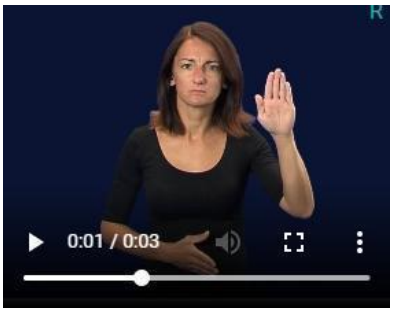
BASIS




BLACK

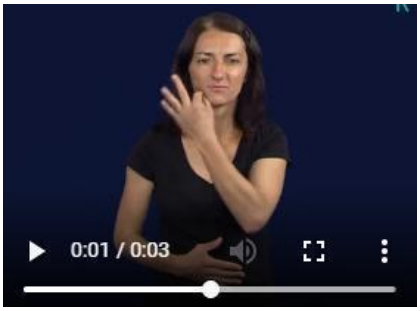
BOW / ARCHERY

CD-ROM

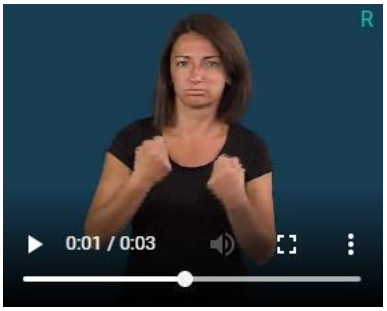



classifier

COFFEE#1

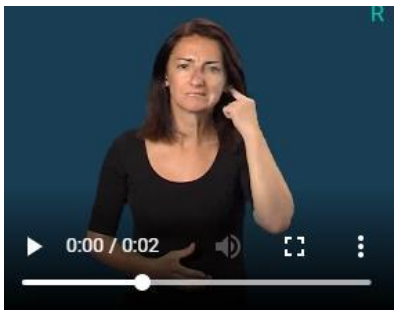



COFFEE#2




DRIVE / CAR

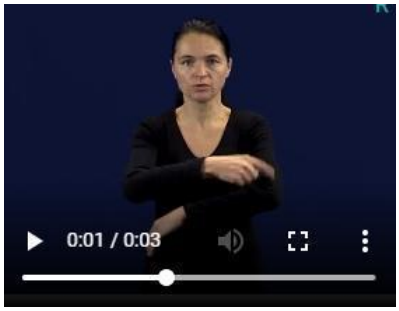
GERMAN / GERMANY




HEAR

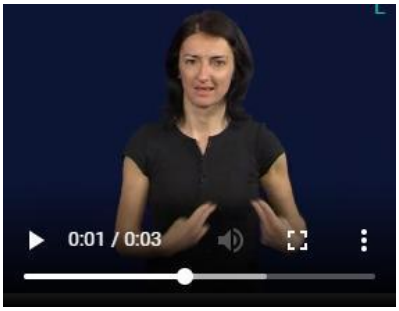
HOUR

ING




KITCHEN#1

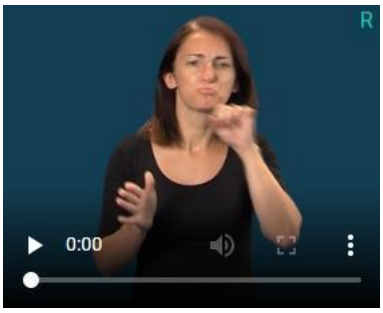
KITCHEN#2




MAY#1

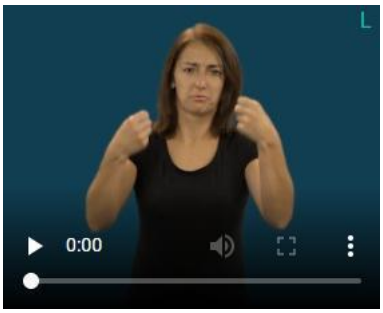
MAY#2

OMELETTE




POST_OFFICE (initial position)

POST_OFFICE (final position)




PRAGUE#1

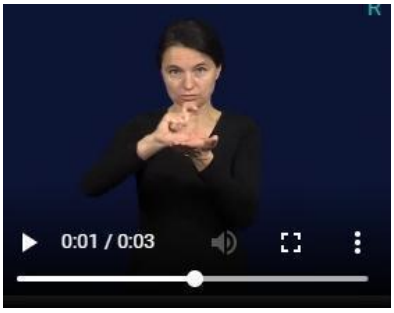
PRAGUE#2

specifier

SPICE / SALT / PEPPER

SUGAR → SACCHARIDE

TREE




YOGHURT

# Razpoznavalnik tekočega govora UMB Broadcast News 2020: prehod na arhitekturo z nevronskimi mrežami

**Andrej Žgank, Mirjam Sepesy Maučec, Gregor Donaj**

Laboratorij za digitalno procesiranje signalov, Fakulteta za elektrotehniko, računalništvo in informatiko,
Univerza v Mariboru
Koroška cesta 46, 2000 Maribor
andrej.zgank@um.si, mirjam.sepesy@um.si, gregor.donaj@um.si

**Povzetek**

V članku bomo predstavili novo verzijo slovenskega avtomatskega razpoznavalnika govora za domeno televizijskih oddaj UMB Broadcast News 2020. Prehod na arhitekturo globokih nevronskih mrež ob upoštevanju razpoložljivih govornih virov je osrednja točka nove verzije sistema. Za učenje razpoznavalnika govora smo uporabili bazi UMB BNSI Broadcast News in IETK-TV. Skupni obseg govornih posnetkov je znašal 66 ur. Vzporedno z globokimi nevronskimi mrežami smo povečali slovar razpoznavanja govora, ki je tako znašal 250k besed. Na ta način smo znižali delež besed izven slovarja na 1,33 %. Z razpoznavanjem govora na testni množici smo dosegli najboljši WER 17,19 %, kar predstavlja 9,62 % izboljšanje glede na predhodno verzijo sistema. Med procesom vrednotenja smo izvedli tudi podrobnejšo analizo napak razpoznavanja govora na osnovi lem in F-razredov.

**UMB Broadcast News 2020 continuous speech recognition system: a new approach based on neural networks**

This paper presents a new version of the Slovenian UMB Broadcast News 2020 automatic speech recognition system for the TV domain. The main contribution of the new version is the transition to deep neural networks architecture, considering available speech resources. We used the UMB BNSI Broadcast News and IETK-TV speech databases to train the speech recognizer. The total amount of recordings was 66 hours. In parallel with the deep neural networks, the vocabulary size was increased from 64k to 250k words. In this way, we reduced the out of vocabulary rate to 1.33%. Speech recognition on the test set achieved the best WER of 17.19%, which represents a 9.62% improvement over our previous version of the system. During the evaluation process, we also performed a more detailed analysis of speech recognition errors based on lemmas and F-conditions.

## 1. Uvod

Napredna pametna okolja v veliki meri temeljijo na zajemanju širokega nabora informacij, ki imajo lahko ali uporabniški ali medijski izvor. Razvoj tehnologij je v zadnjem desetletju privedel do skokovitega povečanja količine podatkov, ki lahko služijo kot vir informacij. Hkrati prihaja v ospredje tudi naravna interakcija z napravami s pomočjo govora, kar uporabniku olajša rokovanje z napravami in zmanjšuje digitalno ločnico. Učinkovit preplet teh razvojnih trendov predstavlja inteligentno okolje interneta stvari.

Ena izmed jedrnih tehnologij, ki omogočajo ustrezno podporo za zajemanje informacij, tako iz uporabniškega ali medijskega toka, kot tudi iz uporabniškega vmesnika, je avtomatsko razpoznavanje govora (ASR). Deluje lahko v zelo različnih scenarijih, od preprostega ukaznega krmiljenja do zahtevnih sistemov za razpoznavanje spontanega govora več govorcev. S kompleksnostjo scenarija je praviloma obratno sorazmerna uspešnost razpoznavanja govora, pomemben vidik pa predstavlja tudi računska zahtevnost, ki lahko pogosto trči ob vprašanja zagotavljanja zasebnosti govora, kadar je v uporabi procesiranje v oblaku.

Področje avtomatskega razpoznavanja govora je neločljivo povezano z razpoložljivostjo govornih virov za posamezni jezik. Tukaj nastopi težava pri jezikih, za katere obstaja manjši (komercialni) interes za implementacijo ASR, kar se lahko še dodatno potencira s posebnostmi določenih jezikov, ki otežijo razpoznavanje govora. V kategorijo za procesiranje zahtevnih jezikov sodi tudi slovenščina, za katero je značilna visoka pregibnost besed in relativno prost vrstni red besed v stavku.

Razvoj prvih sistemov govornih tehnologij za slovenščino se je začel že pred 30 leti, vendar finančno in časovno zahteven razvoj govornih virov v zadnjem desetletju ni uspel slediti svetovnim trendom. Postopki globokega učenja razpoznavalnikov govora namreč za učinkovito delovanje potrebujejo govorne baze v obsegu več 100 oz. 1000 ur transkribiranih posnetkov. Za področje slovenskega jezika pričakujemo razpoložljivost tako obsežnih govornih virov kot enega od rezultatov projekta RSDO – Razvoj slovenščine v digitalnem okolju, ki bo potekal do leta 2022.

Cilj pričujočega raziskovalnega dela[1] je predstaviti nadgradnjo našega sistema za avtomatsko razpoznavanje govora, ki deluje za domeno televizijskih oddaj. Podati želimo oceno, kakšen primanjkljaj pri prehodu na nove arhitekture predstavljajo omejene govorne baze za slovenski jezik. V tem okviru bomo izvedli tudi analizo napak razpoznavanja govora glede na značilnosti jezika, ki so pomembne za rezultate razpoznavanja govora. Delo smo zasnovali na slovenski bazi televizijskih dnevno-informativnih oddaj UMB BNSI Broadcast News (Žgank et al., 2004) in IETK-TV, saj ti dve govorni bazi trenutno še vedno predstavljata najprimernejši vir za takšno analizo, hkrati pa omogočata tudi primerljivost rezultatov s starejšimi sistemi avtomatskega razpoznavanja govora (Žgank et al., 2014).

V nadaljevanju članka bomo najprej predstavili trenutno stanje na področju govornih virov za slovenski jezik. V tretjem poglavju bo sledila kratka predstavitev

[1] Raziskovalno delo je bilo delno sofinancirano s strani ARRS po pogodbi št. P2-0069 in s strani Ministrstva za kulturo RS v okviru projekta RSDO – Razvoj slovenščine v digitalnem okolju.

teoretičnega ozadja pristopov, ki se danes uporabljajo pri gradnji avtomatskih razpoznavalnikov govora. V četrtem poglavju bomo predstavili govorne in jezikovne vire, ki smo jih uporabili pri raziskavi. Postopek izdelave akustičnih in jezikovnih modelov eksperimentalnega sistema bomo opisali v petem poglavju. Rezultate in analizo vrednotenja razpoznavanja govora bomo predstavili v šestem poglavju. V zadnjem poglavju bomo podali zaključne misli.

## 2. Pregled govornih virov za slovenski jezik

Že v uvodu smo zapisali, da predstavljajo govorni viri ključno komponento za razvoj avtomatskega razpoznavalnika govora. Pomembno je, da pri tem s svojimi značilnostmi in obsegom materiala vplivajo tudi na to, katero arhitekturo nevronskih mrež, ki so danes najbolj aktualna tehnologija pri razvoju razpoznavalnikov, bo možno uspešno naučiti.

Dosedanji razvoj govornih virov za slovenski jezik lahko razdelimo na dve obdobji. V prvem obdobju, ki se je začelo v devetdesetih letih prejšnjega stoletja, je bil poudarek na razvoju govornih baz za omejene scenarije izoliranih ali vezanih besed. Snemalni kanal je bil ali studio ali telefon, obseg govornega material pa praviloma med 10 in 15 ur. V to skupino lahko uvrstimo govorne baze: FDB 1000 Slovenian SpeechDat(II) (Kaiser in Kačič, 1997), Polidat (Žgank et al., 2002), Gopolis (Dobrišek et al., 1998), VNTV/VNRAD (Žibert et al., 2003) in SNABI. Delni sklopi naštetih baz že vsebujejo tudi tekoči govor, vendar je zaradi omejene količine govornega materiala praktičen razvoj splošnega razpoznavalnika govora še nemogoč.

V drugem obdobju razvoja govornih baz za slovenski jezik, ki se je začelo okoli leta 2004, se aktivnosti osredotočijo na tekoči govor. Bistveno se razširi domena vključenega materiala, kot snemalni kanal pa se dodatno pojavi televizija oziroma druge oblike javnega govora, kot so npr. predavanja. Obseg govornih baz se poveča na nekaj 10 ur posnetkov. Sem lahko prištejemo sledeče televizijske baze: UMB BNSI Broadcast News (36 ur) (Žgank et al., 2004), SiBN Broadcast News (36 ur) (Žibert in Mihelič, 2004), IETK-TV (30 ur) in GOS javni podkorpus (42 ur) (Verdonik et al., 2013). Predavanja najdemo v bazi SI TEDx-UM (54 ur, avtomatske transkripcije) (Žgank et al., 2016) in bazi GOS-Videolectures (22 ur) (Verdonik, 2018). Baza SloParl (Žgank et al., 2006) vsebuje 100 ur posnetkov in magnetogramov parlamentarnih razprav iz DZ RS, baza SOFES (Dobrišek et al., 2017) pa 10 ur posnetkov s poizvedbami po letalskih informacijah.

Dostopnost predstavljenih govornih baz pokriva skoraj celotni spekter možnosti. Nekatere so prosto dostopne preko iniciative Clarin oz. na spletnih straneh avtorjev. Druge baze so dostopne proti plačilu preko organizacije ELRA. Del baz pa je namenjen izključno interni uporabi in tako nedostopen širši raziskovalni skupnosti.

Skupna dolžina transkribiranih posnetkov v predstavljenih govornih bazah je približno 250 ur. Dodatnih 150 ur posnetkov je transkribiranih samo avtomatsko ali v obliki magnetogramov. Tudi če bi kljub različnim omejitvam v dostopnosti uspeli združiti vse govorne baze, prihaja med njimi v zasnovi do tako velikih razlik, da bi bilo učenje razpoznavalnika govora na takšen način neizvedljivo. Ob upoštevanju kriterija sorodnosti in dostopnosti govornih baz je trenutno praktično možno za učenje slovenskega razpoznavalnika govora uporabiti med 50 in 100 urami posnetkov. Takšen obseg učnega materiala je premajhen za uporabo naprednejših arhitektur globokega učenja.

To dejstvo lepo kaže na nujno potrebo po tretjem obdobju v razvoju govornih baz za slovenski jezik, kjer je cilj pridobiti nekaj 100 do 1.000 ur posnetkov, ki so prosto dostopni in omogočajo potencialno kombiniranje virov v prihodnosti. V to kategorijo bo sodila govorna baza, ki nastaja v okviru projekta RSDO.

## 3. Arhitekture za avtomatsko razpoznavanja govora

Na področju arhitekture avtomatskih razpoznavalnikov govora obstajata dve glavni skupini. Prvo predstavljajo sistemi s prikritimi modeli Markova, ki so bili glavni gradnik akustičnega modeliranja v preteklosti. Drugo skupino, ki je danes standardna, pa predstavljajo sistemi na osnovi nevronskih mrež.

### 3.1. Prikriti modeli Markova

Prikriti modeli Markova uporabljajo pristop statističnega modeliranja, kjer na osnovi vhodnih vektorjev značilk ocenjujejo verjetnost hipoteze izgovorjenega besedila. Običajno se uporabljajo večstanjski levo-desni prikriti modeli, kjer je porazdelitvena funkcija gostote verjetnosti modelirana s skupino uteženih multivariantnih Gaussovih porazdelitvenih funkcij. Z vidika računske kompleksnosti in količine zahtevanega učnega materiala gre praviloma za manj zahtevne sisteme v primerjavi z globokimi nevronskimi mrežami.

### 3.2. Globoke nevronske mreže

Nevronske mreže predstavljajo pristop na področju strojnega učenja, ki deloma posnema dogajanje v nevronskem sistemu. Mreže so sestavljene iz nevronov, ki so razporejeni v plasti – vhodno plast, notranje plasti in izhodno plast.

Vsak nevron izvaja matematično operacijo, kjer najprej izračuna uteženo vsoto vrednosti na njegovih vhodih, nato pa to vsoto uporabi v aktivacijski funkciji, da dobi izhodno vrednost nevrona. Izhodi nevronov so potem povezani na vhode drugih nevronov.

V zadnjih letih so nevronske mreže postale popularne na raznih področjih strojnega učenja, tudi pri razpoznavanju govora. Ker pa gre tukaj za razpoznavanje časovne vrste, niso vse arhitekture nevronskih mrež primerne.

Med korakom učenja se nevronska mreža prilagaja na učne podatke tako, da spreminja uteži. Pri uporabi pa nato dajemo nove podatke na vhodno plast omrežja ter opazujemo rezultate na izhodni plasti.

## 4. Govorni in jezikovni viri

Osrednji vir podatkov za modeliranje je predstavljala govorna baza UMB BNSI Broadcast News (Žgank et al., 2004), ki jo distribuira organizacija ELRA (2020). Govorna baza vsebuje posnetke dnevno-informativnih televizijskih oddaj RTV Slovenija v obsegu 36 ur. Od tega je 30 ur namenjenih učenju akustičnih modelov. Oddaje so nastale v letih 1999–2003, tako da je bila z vidika naprav uporabljenih v produkciji tehnologija delno drugačna, kot

jo srečamo danes (npr.: snemalne naprave z izgubnimi kodeki, povezave VoIP, spletne komunikacijske platforme). V bazi je skupaj 1.565 govorcev, od tega 1.069 moških in 477 žensk. Za 19 govorcev spola ni bilo možno nedvoumno določiti.

Posnetki so bili ročno segmentirani in transkribirani. Hkrati je bilo označeno tudi akustično ozadje in negovorni akustično dogodki. To je posledica produkcije oddaj, saj je pogosto v ozadje zvočnega posnetka glavnega govorca montiran zvočni posnetek iz videa ali pa drugo zvočno ozadje, kot je na primer glasba. Pri avtomatskem razpoznavanju govora je pomemben vidik tudi, ali gre za bran, načrtovan ali spontan govor, saj ta značilnost pomembno vpliva na dosežene rezultate.

V predhodnem odstavku naštete parametre v domeni razpoznavanja govora televizijskih oddaj karakterizirajo F-razredi (Schwartz et al., 1997). Ti so definirani na sledeči način:

- F0: bran govor v studijskem okolju,
- F1: spontan govor v studijskem okolju,
- F2: bran/spontan govor preko telefona,
- F3: bran/spontan govor z glasbo v ozadju,
- F4: bran/spontan govor z drugim zvočnim ozadjem,
- F5: govorci, katerih materni jezik ni slovenščina,
- FX: preostalo.

Predstavljene F-razrede bomo uporabili pri podrobnejši analizi rezultatov v šestem poglavju, saj bodo služili za oceno težavnosti testnega scenarija. F-razredi so v govorni bazi zastopani v različnih deležih. Ker predstavlja testni nabor v dolžini 3 ur manj kot eno desetino baze, se to odraža tudi v zastopanosti F-razredov. Tako v testni množici v celoti manjka razred F5 z govorci katerih materni jezik ni slovenščina. Po obsegu pa je najmanjši razred F2 s telefonskim pogovorom. Vsebuje samo 8 segmentov 3 govorcev, ki skupaj izgovorijo nekaj več kot 100 besed.

Nabor učne množice za akustično modeliranje avtomatskega razpoznavalnika govora smo razširili z interno bazo IETK-TV. Ta baza predstavlja nadgradnjo baze UMB BNSI Broadcast News in je nastala na osnovi istih specifikacij. Obsega 29 ur transkribiranih posnetkov 784 govorcev, ki so v celoti namenjeni akustičnemu modeliranju. Nabor različnih televizijskih oddaje je v bazi IETK-TV razširjen, saj so vključeni tudi intervjuji in okrogle mize. Posledično je delež spontanega govora v bazi IETK-TV več kot enkrat večji kot v bazi UMB BNSI Broadcast News.

Za gradnjo novega jezikovnega modela učnega korpusa nismo razširjali. Uporabili smo obstoječe korpuse BNSI-Speech (573k besed), BNSI-Text (11M besed) in FidaPLUS (621M besed) (Arhar in Gorjanc, 2007). Korpus Večer smo iz učenja izločili, saj so njegovi članki vsebovani v korpusu FidaPLUS.

# 5. Eksperimentalni sistem

## 5.1. Akustično modeliranje

Za izgradnjo avtomatskega razpoznavalnika govora smo uporabili odprtokodno orodje Kaldi (Povey et al., 2011), ki omogoča izgradnjo sistema s pristopi globokega učenja.

Izvorni signal, ki je del uporabljene govorne baze, smo najprej oknili, nato pa tvorili značilke v obliki mel-frekvenčnih kepstralnih koeficientov (MFCC). Posamezni vektor značilk je imel 13 elementov, ki smo jim dodali še prvi in drugi odvod. Sledil je postopek akustičnega modeliranja. V primeru orodja Kaldi gre za hibridni pristop, ki v prvem koraku uči prikrite modele Markova, v drugem koraku pa globoko nevronsko mrežo. Kot osnovno enoto za akustično modeliranje smo ponovno uporabili slovenske grafeme, saj smo tako dosegli primerljivost s predhodno objavljenimi rezultati.

Prikriti modeli Markova, uporabljeni v akustičnem modeliranju, imajo tristanjsko levo-desno topologijo. Izgradnja akustičnih modelov poteka postopoma, kjer se koraki učenja parametrov modela z Baum-Welchevo re-estimacijo izmenjujejo s koraki prisilne poravnave učnih transkripcij. Za monofonske akustične modele smo uporabili 40 iteracij, za kontekstno odvisne trifonske modele pa 35 iteracij.

Sledilo je učenje globokih nevronskih mrež. Pri tem smo kot arhitekturo uporabili navadno usmerjeno globoko nevronsko mrežo s p-norm aktivacijsko funkcijo (Zhang et al., 2014). Vrednost parametra *p* smo nastavili na 2, saj smo tako dobili primerjalno najboljše rezultate. Podoben eksperiment smo ponovili v povezavi s spreminjanjem konfiguracije nevronske mreže, kjer se je pokazalo, da je smiselno uporabiti tri skrite plasti. Učenje nevronske mreže je potekalo v 10 regularnih epohah in 5 dodatnih, kar je skupaj predstavljajo 345 učnih iteracij. Predstavljeni parametri so v veliki meri odvisni tako od količine učnega materiala kot tudi od njegove raznolikosti. Posledično jih je potrebno ustrezno prilagoditi za vsak govorni vir. Cilj je doseči dobre rezultate razpoznavanja govora, hkrati pa ohraniti zmožnost generalizacije na nove testne vzorce. V nasprotnem primeru dosežemo prekomerno prileganje globoke nevronske mreže. Omejena količina razpoložljivega učnega govornega materiala je bila tudi razlog, da nismo uporabili kompleksnejših pristopov globokega učenja, kot so na primer »end-to-end« globoke nevronske mreže.

## 5.2. Jezikovno modeliranje

V eksperimentih smo uporabili 2 slovarja, prvi je vseboval 64.000 besed, drugi pa 250.000. Pripadajoča slovarja izgovorjav smo tvorili na osnovi grafemskih akustičnih enot, katerim smo dodali model tišine in pa ločen model različnih negovornih zvokov, ki jih je tvoril govorec. Prvi slovar je identičen slovarju iz prejšnjih eksperimentov (Žgank in Sepesy Maučec, 2010; Žgank et al., 2014), drugi pa vsebuje vse besede korpusov BNSI-Speech in BNSI-Text. Do velikosti 250.000 smo ga dopolnili z najpogostejšimi besedami iz korpusa FidaPLUS. Delež besed izven slovarja (OOV) prvega slovarja je 4,22 %, drugega pa 1,33 %. Ker oba slovarja vsebujeta besede iz korpusa BNSI-Speech, so med običajnimi besedami tudi različna mašila in onomatopeje, ki smo jih modelirali kar na osnovi njihove zvočne pojave in ne kot posebne, ločene, akustične modele.

Z orodjem SRI Language Modeling Toolkit (Stolcke, 2002) smo zgradili trigramske modele. Trigramski model s prvim slovarjem je identičen kot v prejšnjih eksperimentih (Žgank in Sepesy Maučec, 2010, Žgank et al., 2014). Tudi z drugim slovarjem smo zgradili interpoliran trigramski model. V vseh treh komponentah smo uporabili Good-

Turingovo glajenje in sestopanje po Katz-u. V komponenti BNSI-text smo izločili trigrame s frekvenco 1, v komponenti FidaPLUS pa bigrame s frekvenco 1 in trigrame s frekvencama 1 in 2. Na ta način smo dobili trigramski model, ki je bil primerljive velikosti kot trigramski model s prvim slovarjem. Perpleksnost modela na testni množici je bila 284.

## 6. Rezultati razpoznavanja govora

Vrednotenje različnih sistemov avtomatskega razpoznavanja govora smo izvedli na testni množici baze UMB BNSI Broadcast News (BNSI-eval), ki vsebuje 4 televizijske oddaje v obsegu 3 ur. Za metriko vrednotenja uspešnosti razpoznavanja govora smo uporabil delež napačno razpoznanih besed (Word Error Rate – WER), ki je definiran kot:

$$WER(\%) = \frac{(I+D+S)}{N} \cdot 100\,, \qquad (1)$$

kjer je $I$ število vrinjenih besed, $D$ število izbrisanih besed in $S$ število zamenjanih besed. $N$ predstavlja število vseh besed v testni množici. V delu analize rezultatov smo uporabili kot metriko tudi delež napačno razpoznanih lem (Lemma Error Rate – LER), ki je definiran kot:

$$LER(\%) = \frac{(i+d+s)}{n} \cdot 100\,, \qquad (2)$$

kjer je $i$ število vrinjenih lem, $d$ število izbrisanih lem in $s$ število zamenjanih lem. $n$ je skupno število vseh lem v testni množici in je enako številu besed $N$.

V prvem koraku evalvacije smo izvedli primerjavo med avtomatskim razpoznavalnikom govora s prikritimi modeli Markova in globokimi nevronskimi mrežami. Pri tem je sistem s prikritimi modeli Markova služil za primerjavo z rezultati sistema iz leta 2014, ki je takrat dosegel najboljši WER 26,81 % (Žgank et al., 2014). Rezultati napake razpoznavanja besed s trigramskim jezikovnim modelom in slovarjem besed z velikostjo 64k so predstavljeni v tabeli 1.

| Sistem | WER [%] |
|---|---|
| HMM | 26,42 |
| DNN | 20,85 |

Tabela 1: Rezultati razpoznavanja govora s trigramskim 64k jezikovnim modelom.

Izhodišča primerjava HMM sistema med letoma 2014 in 2020 kaže, da je prehod na novo ogrodje za avtomatsko razpoznavanje govora potekal brez težav, saj smo dosegli zelo primerljiv WER (iz 26,81 % na 26,42 %). Trenutne HMM akustične modele je sicer možno dodatno nadgraditi s pristopoma od govorca neodvisne transformacije značilk z uporabo LDA (angl. Linear Discriminant Analysis) in MLLT (angl. Maximum Likelihood Linear Transform) (Gales, 1999), kar izboljša rezultat na 24,48 %. Vendar je to izboljšanje relativno omejeno v primerjavi z možnostmi, ki jih v ustreznih pogojih omogoča globoko učenje. Prehod na globoke nevronske mreže za akustično modeliranje izboljša napako razpoznavanja besed na 20,85 %, kar predstavlja pomembno razliko. Pri tem je potrebno posebej izpostaviti, da je količina govornega učnega materiala relativno omejena za pristope globokega učenja.

V drugem koraku smo izvedli vrednotenje, kako vpliva na rezultate izboljšani jezikovni model z bistveno večjim slovarjem besed. Prehod iz 64k besed na 250k besed namreč izdatno zniža delež besed izven slovarja, in ga približa deležu, ki ga najdemo v tipičnih jezikovnih modelih za angleški jezik pri velikosti slovarja 64k. Se pa poveča perpleksnost takšnega jezikovnega modela. Rezultati razpoznavanja govora z akustičnimi modeli DNN in obema trigramskima jezikovnima modeloma so predstavljeni v tabeli 2.

| Jezikovni model | WER [%] |
|---|---|
| 64k-3g | 20,85 |
| 250k-3g | 17,19 |

Tabela 2: Rezultati razpoznavanja govora z akustičnimi modeli DNN in z različnima trigramskima jezikovnima modeloma.

Tudi v tem scenariju razpoznavanja govora je prišlo do znatnega zmanjšanja napake razpoznavanja besed, saj je WER znašal 17,19 %. S povečanjem slovarja razpoznavalnika govora smo tako izboljšali delovanje za 3,66 %, kar je primerljivo z zmanjšanjem deleža OOV. Pri tem smo ohranili kompleksnost sistema na primerljivi ravni, za kar smo poskrbeli med procesom izdelave jezikovnega modela. Razpoznavanje slovenščine z nevronskimi mrežami so predstavili tudi Ulčar et al., 2019. Dosegli so WER 27,16 % na bazi GOS VideoLectures 2.0. Zaradi uporabe različnih govornih in jezikovnih virov rezultati niso medsebojno primerljivi.

V nadaljevanju poglavja bomo podrobneje predstavili analizo doseženih rezultatov razpoznavanja govora. Odgovoriti poskušamo na vprašanje, kako različni faktorji vplivajo na WER. V to skupino sodijo delež besed izven slovarja, pregibna oblika besed, akustični ozadje in način govora.

Referenčne transkripcije in rezultate razpoznavanja smo oblikoslovno označili ter lematizirali z označevalnikom slovenskega jezika Obeliks (Grčar et al., 2012). Oznake besedne vrste in leme so nam koristile pri podrobnejši analizi rezultatov.

S primerjavo lematizirane referenčne transkripcije ter lematiziranih rezultatov razpoznavanja govora smo določili delež napačno razpoznanih lem ter izluščili napake, kjer je lema pravilno razpoznana, besedna oblika pa ne. Na takšen način smo lahko delno analizirali vpliv pregibnosti slovenskega jezika na rezultate razpoznavanja govora. S pomočjo oblikoslovnih oznak pa smo nato še napake v besedni obliki razdelili po besednih vrstah.

V tabeli 3 so predstavljeni podrobnejši rezultati. Razdeljeni rezultati po F-razredih in po spolu kažejo večinoma podobna izboljšanja pri prehodih med sistemi. Opazna je razlika med rezultati za moške in ženske govorce, ki znaša 4,56 %. To razliko bo potrebno še podrobneje analizirati v prihodnosti. Večja izboljšanja vidimo v razredih F1, F3 in FX pri prehodu iz sistema HMM na DNN ter pri razredu F2 pri prehodu na večji slovar, ki pa predstavlja le zelo majhen del testne množice. Med tem ko za bran studijski govor dosegamo WER 9,14 %, sprememba na spontani govor ali dodajanje akustičnega ozadja poslabša rezultate v rangu 10 do 15 %. Pri tem je pričakovano poslabšanje večje, v kolikor je v ozadju dodana glasba.

| Sistem | HMM 64k-3g | DNN 64k-3g | DNN 250k-3g |
|---|---|---|---|
| WER [%] | 26,42 | 20,85 | 17,19 |
| WER - F0 [%] | 14,89 | 12,35 | 9,14 |
| WER - F1 [%] | 35,37 | 26,70 | 23,72 |
| WER - F2 [%] | 53,38 | 51,69 | 33,89 |
| WER - F3 [%] | 38,53 | 28,60 | 24,13 |
| WER - F4 [%] | 29,70 | 23,71 | 19,52 |
| WER - FX [%] | 34,84 | 26,54 | 23,83 |
| WER - Moški [%] | 28,92 | 22,57 | 19,21 |
| WER - Ženske [%] | 23,26 | 18,57 | 14,65 |
| LER [%] | 23,66 | 18,07 | 14,85 |
| WER – LER | 2,76 | 2,78 | 2,34 |

Tabela 3: Podrobnejša predstavitev rezultatov razpoznavanja po F-razredih in spolu ter rezultati pravilnosti razpoznave lem.

Rezultati deleža napačno razpoznanih lem LER so po pričakovanjih nižji od rezultatov WER. Te razlike nakazujejo napake v razpoznavanju, kjer je sistem napačno razpoznal besedno obliko, vendar imata tako razpoznana kot pravilna beseda enako lemo. Vidimo, da je razlika manjša pri sistemu z večjim slovarjem, kar nakazuje, da je za del napačno razpoznanih besednih oblik odgovoren omejen slovar.

Treba je dodati, da je v nekaterih primerih bila razpoznana pravilna besedna oblika, vendar je lematizator označil različni lemi med hipotezo in referenco. Ti primeri so se šteli kot napake v vrednotenju LER. To so dogaja predvsem pri primerih, kjer se zaradi drugih napak (izbrisanih ali vrinjenih kratkih besed) spremeni kontekst besede. Na primer, besedna oblika *ukrepa* je lahko označena z lemo *ukrep* (samostalnik) ali pa *ukrepati* (glagol). Ocenjujemo pa, da je delež teh primerov le majhen. Iz tega sklepamo, da je delež napak, ki so posledica pregibnosti jezika, nekoliko višji kot pa razlika med WER in LER, namreč med 2,5 in 3 %.

V nadaljevanju smo pregledali napake v besedni obliki pri isti lemi glede na besedno vrsto. Rezultati so podani v tabeli 4. Podali smo le pregibne besedne vrste (brez zaimkov). Primerjamo sistem HMM 64k-3g in DNN 250k-3g. Vidimo, da je le relativno izboljšanje pri napačno razpoznanih oblikah števnikov primerljivo z relativnim izboljšanjem skupnega rezultata, ki je 34,9 %. Najmanjše relativno izboljšanje pa vidimo pri glagolih. Skupno relativno izboljšanje napak v besedni obliki pa je približno dvakrat manjše od relativnega izboljšanja skupnega rezultata.

Rezultati kažejo na to, da sistem s povečanim slovarjem in uporabo nevronskih mrež pomembno zmanjša skupni delež napak razpoznavanja. Vidimo pa, da je relativno zmanjšanje napak zaradi pregibnosti besed manjše glede na skupno zmanjšanje. V sistemu DNN 250k-3g je tako delež napak zaradi pregibnosti 13,3 %, kar je več kot pri sistemu HMM, kjer je ta delež 10,5 %.

Pregled posameznih najpogostejših parov zamenjav ne kaže zanimivih rezultatov glede pregibnih besed. Večinoma se v pogostih parih zamenjav pojavljajo kratke besede (npr. zamenjave so – se, na – no ipd.). Najpogostejši par zamenjave, kjer je prišlo do napake v besedni obliki polnopomenske pregibne besede je par stališče–stališča, ki se pojavi štirikrat v sistemu DNN 250k-3g.

## 7. Zaključek

V članku smo predstavili sistem za avtomatsko razpoznavanje govora v domeni televizijskih oddaj, ki je dosegel najboljši delež napake razpoznavanja besed 17,19 %. Izboljšanje je v pretežni meri rezultat uporabe akustičnih modelov z globokimi nevronskimi mrežami in vpliva zmanjšanja deleža besed izven slovarja. Z večanjem slovarja smo uspešno zmanjšali vpliv pregibnosti slovenskega jezika.

Podrobnejša analiza po F-razredih in lemah je pokazala, da je nadaljnje izboljšanje rezultatov možno doseči predvsem na račun izboljšanja akustičnega modeliranja v primeru kratkih besed in govora v zahtevnejših pogojih.

## Zahvala

Zahvaljujemo se avtorjem besedilnega korpusa FidaPLUS, ki so nam omogočili njegovo uporabo za jezikovno modeliranje avtomatskega razpoznavalnika govora.

| Besedna vrsta | Št. napak v HMM 64k-3g | Št. napak v DNN 250k-3g | Relativna izboljšava [%] |
|---|---|---|---|
| Samostalnik | 309 | 265 | 14,2 |
| Pridevnik | 155 | 112 | 27,7 |
| Glagol | 148 | 135 | 8,8 |
| Števnik | 16 | 10 | 37,5 |
| Prislov | 0 | 1 | - |
| SKUPAJ | 628 | 522 | 16,9 |

Tabela 4: Napačno razpoznane besedne oblike glede na besedno vrsto.

## 8. Literatura

Špela Arhar in Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2., 95—110.

Simon Dobrišek, Jerneja Gros, France Mihelič in Nikola Pavešić. 1998. Recording and labelling of the GOPOLIS Slovenian speech database. V: *First International Conference on language resources & evaluation*: Granada, Spain, 28-30 May 1998 (str. 1089—1096). European Language Resources Association.

Simon Dobrišek, Jerneja Žganec Gros, Janez Žibert, France Mihelič in Nikola Pavešić. 2017. Speech Database of Spoken Flight Information Enquiries SOFES 1.0.

ELRA. 2020. BNSI Catalog Reference : S0275: www.elra.info.

Mark J. Gales. 1999. Semi-tied covariance matrices for hidden Markov models. *IEEE transactions on speech and audio processing*, 7(3), 272—281.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: *Zbornik Osme konference Jezikovne tehnologije*, Ljubljana, Slovenija.

Janez Kaiser in Zdravko Kačič. 1997. SpeechDat (II) Slovenian Database for the Fixed Telephone Network. Maribor, Slovenia: University of Maribor.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel … in Jan Silovsky. 2011. The Kaldi speech recognition toolkit. V: *IEEE ASRU 2011 Workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Richard Schwartz, Hubert Jin, Francis Kubala in Spyros Matsoukas. 1997. Modeling those F-Conditions - or not. *Proc. DARPA Speech Recognition Workshop*, Chantilly, ZDA.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. *International Conference on Speech and Language Processing*, II: 901—904.

Matej Ulčar, Simon Dobrišek in Marko Robnik-Šikonja. 2019. Razpoznavanje slovenskega govora z metodami globokih nevronskih mrež. *Uporabna informatika. 27, 3*.

Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek in Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language resources and evaluation*, 47(4), 1031—1048.

Darinka Verdonik. 2018. Korpus in baza Gos Videolectures. *Zbornik 11. konference Jezikovne tehnologije*, Informacijska družba-IS, Ljubljana.

Xiaohui Zhang, Jan Trmal, Daniel Povey in Sanjeev Khudanpur. 2014. Improving deep neural network acoustic models using generalized maxout networks. V: *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (str. 215—219). IEEE.

Andrej Žgank, Zdravko Kačič in Bogomir Horvat. 2002. Preliminary evaluation of Slovenian mobile database PoliDat. V: *Proceedings of the Third International Conference on Language Resources and Evaluation* (LREC'02).

Andrej Žgank, Tomaž Rotovnik, Mirjam Sepesy Maučec, Darinka Verdonik, Jani Kitak, Damjan Vlaj, Vladimir Hozjan, Zdravko Kačič in Bogomir Horvat. 2004. Acquisition and annotation of Slovenian broadcast news database. *Fourth international conference on language resources and evaluation*, LREC 2004, Lizbona, Portugalska.

Andrej Žgank, Tomaž Rotovnik, Matej Grašič, Marko Kos, Damjan Vlaj, in Zdravko Kačič. 2006. Sloparl-Slovenian parliamentary speech and text corpus for large vocabulary continuous speech recognition. V: *Ninth International Conference on Spoken Language Processing.*

Andrej Žgank in Mirjam Sepesy Maučec. 2010. Razpoznavalnik tekočega govora UMB Broadcast News 2010: nadgradnja akustičnih in jezikovnih modelov. *Jezikovne tehnologije 2010*, Ljubljana, Slovenija.

Andrej Žgank, Gregor Donaj in Mirjam Sepesy Maučec. 2014. Razpoznavalnik tekočega govora UMB Broadcast News 2014: kakšno vlogo igra velikost učnih virov. *Zbornik 9. konference Jezikovne tehnologije, Informacijska družba-IS* (str. 147—150).

Andrej Žgank, Mirjam Sepesy Maučec in Darinka Verdonik. 2016. The SI TEDx-UM speech database: A new Slovenian spoken language resource. V: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (str. 4670—4673).

Janez Žibert, Sanda Martinčić-Ipšić, Ivo Ipšić in France Mihelič. 2003. Bilingual speech recognition of Slovenian and Croatian weather forecasts. V: Proceedings EC-VIP-MC 2003. *4th EURASIP Conference focused on Video/Image Processing and*

*Multimedia Communications* (IEEE Cat. No. 03EX667) (Vol. 2, str. 637—642). IEEE.

Janez Žibert in France Mihelič. 2004. Development of Slovenian broadcast news speech database. V: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04).*

# Challenges in Building MedCorpInn - a Corpus of Unstructured German Radiology Reports

**Karoline Irschara*, Claudia Posch*, Birgit Waldner*†, Stephanie Mangesius†, Leonhard Gruber†, Anna-Lena Huber†, Gerhard Rampl*, Astrid Grams†, Bernhard Glodny†**

*University of Innsbruck, Department of Linguistics
Innrain 52d, 6020 Innsbruck
karoline.irschara@uibk.ac.at[1], claudia.posch@uibk.ac.at[1], gerhard.rampl@uibk.ac.at
†Medical University of Innsbruck, Department of Radiology
Anichstraße 35, 6020 Innsbruck
Birgit.Waldner@i-med.ac.at, leonhard.gruber@i-med.ac.at, anna.huber@i-med.ac.at, stephanie.mangesius@tirol-kliniken.at, astrid.grams@i-med.ac.at, bernhard.glodny@i-med.ac.at

## 1. Introduction

This extended abstract reports on the ongoing project *MedCorpInn - Retrospective Intersectional Corpuslinguistic Analysis of Radiology Reports of Innsbruck Medical University*[2] and outlines the construction of the *MedCorpInn*-corpus, a large linguistically tagged corpus of medical reports. *MedCorpInn* contains 5,002,933 written reports in German (2007-2019) from the Clinic of Radiology and Neuroradiology at Medical University of Innsbruck.[3] First, we give a general description of our data and the current project status. We focus on the particular challenges the radiology reports pose for pipeline building and we suggest possible solutions for the difficulties encountered. Lastly, we give a roadmap on research intended with the corpus once completed.

## 2. The MedCorpInn data

Some preliminary work had been previously carried out in the smaller pilot project *KARBUN* (Irschara, Posch and Glodny, 2017), which serves as a best practice model for building the more extensive corpus of radiology reports. These texts play an essential role in the communication of physicians and serve as a legal record documenting the imaging procedures (Kahn et al., 2009). They contain information and interpretation regarding the different imaging procedures and types of examinations, e.g. computer tomography, ultrasound, magnetic resonance imaging, angiography, X-ray, fluoroscopy etc. Furthermore, they include demographic information (e.g. age, gender, nationality, type of insurance, occupational status) as well as medical metadata (e.g. mode of examination, time frame, medical indication, referral diagnosis etc.).

Concerning corpus building, there are issues at all levels of pre-processing: A first considerable problem arises when structuring the unstructured data. While the metadata occur in a rather uniform way, the report texts are different varieties of free text entries. For example, reports would often include individual headings or subheadings in a random order, depending on the date of the report, the department or/and the individual doctors. The usage of headings is however not consistent among the records. Another important task is the implementation of a de-identification strategy. The data management first of all complies with data confidentiality (§6 GDPR, current version) as declared in the approval of the ethical review committee of Medical University of Innsbruck. Most of the sensitive information (namely patient and/or doctor names and IDs) were already removed from the metadata during the process of data extraction from the clinical information system (e.g. the patient name was unticked in the extraction form and thus never appears in the metadata in the first place). Some types of metadata are generalized (e.g. date of birth, occupation). However, a limited amount of sensitive information might still appear in the free texts and must be detected and masked. For example, occasionally doctors' names appear and they are removed by using RegEx codes. This is feasible because they always are preceded by either one or more of their academic degrees and/or positions within the clinic or another form of appellation like "Frau" ('Ms.') or "Herr" ('Mr.'). The specific language used in radiology reports makes it particularly challenging to implement a text processing pipeline: Not only do the reports contain many abbreviations, short forms and ad-hoc forms, but also Latin terms or germanized

[1] corresponding authors
[2] The project *MedCorpInn* is funded by the *go!digital* 2.0 call of the Austrian Academy of Science.
[3] Radiology reports were chosen as a starting point because they were readily accessible; the use of further text types as well as data from other clinics in future projects is intended.

(pseudo-)Latin terms. In addition, typos, missing spaces and the lack of paragraph structure make it difficult to perform paragraph and sentence splitting as well as tokenization using available NLP packages. Hence, conventional German POS-taggers have problems with the correct recognition of medical terms since they are usually oriented towards standard language texts (cf. Hellrich, Matthies, Faessler and Hahn, 2015). Nonetheless, we consider POS-tagging important for NER and lemmatization. It is also useful for word-sense disambiguation and for finding lexical or grammatical patterns in the data.

Additionally, medical term and abbreviation recognition are crucial for further corpus development, such as sentence boundary disambiguation and tokenization, which form the basis of our corpus pipeline. Regarding medical term recognition and abbreviation detection we are in the process of constructing relevant thesauri from the corpus and also partly work with the German translation of the extensive radiological ontology RadLex (RNSA, 2017).

Furthermore, the team is currently testing the possibility of transferring some of the solutions on noisiness of clinical data that exist for English (cf. Cai et al., 2016; Chapman et al., 2011; Šuster et al., 2017) onto German, even though this will not be possible for all the language specific problems.

## 3. MedCorpInn as a source for language and medical research

We suggest viewing the reports as in *MedCorpInn* as linguistic events, which may be affected by linguistic and social aspects, which can be analysed and researched with the corpus *MedCorpInn*. Since the 1990s, research on healthcare communication has predominantly focused on qualitative methods for investigating, for example, doctor-patient interaction (Maynard and Heritage, 2005; Atkins and Harvey, 2010; Menz, 2011). Subsequently, the focus was expanded to medical communication in a broader sense, e.g. to medical discourses or internal clinical communication (Crawford, Brown and Harvey, 2014). Recently, the fields of Corpus Linguistics and Natural Language Processing have opened new approaches to medical data which allow qualitative and quantitative methods to be combined (Taylor and Marchi, 2018; Wiegand and Mahlberg, 2019; Demjén, 2020).

For the German language, there are only few studies which apply mixed methods on clinical data: Most research of the investigations in this area focuses on developing specific NLP applications (e.g. information retrieval, de-identification etc.), but there is no substantial linguistic research on the data itself (Crawford, Brown and Harvey, 2014). Also, there is little research concerning corpus linguistic and especially discourse linguistic investigations of clinical text corpora (Demjén, 2020).

*MedCorpInn* therefore serves as a unique source for studies both in the fields of linguistics as well as medicine. The language used in the MedCorpInn data provides insights into the everyday communicative practices between health professionals within the fields of radiology and neuroradiology, which have only been marginally investigated from a linguistic perspective so far (Reiner, 2012).

The research questions we intend to investigate with this corpus concern (discourse) linguistic questions, which can be linked to gender medicine issues. For example, we want to analyse whether salient linguistic patterns such as keywords, n-grams or specific collocations (and collocation types) in the data are somehow connected to social categories in the metadata (e.g. age, gender, origin, type of insurance etc.) and how. Could such a connection for example indicate bias, e.g. by certain usage patterns of diminutives and amplifiers? We aim to find out how patients/groups/people are talked about, which information is made explicit on the linguistic surface and how (e.g. an examiner constantly labelling a patient as ‘asylum seeker’). As the texts are often highly standardized and schematic, variability may also be an interesting subject for study.

From a gender medicine perspective, the corpus can be queried for proposed and described medical procedures (e.g. screenings, preventive examinations) and investigate if such propositions are connected to the social and economic categories in the metadata. We also perform specific information extraction tasks on the data, e.g. the extraction of measurements of tumour diameters and the type of tumour described in the corpus. There is a number of different motivations behind researching the differences in measurement accuracies, for example to learn about mean sizes of tumours at the time of diagnosis, to look for potential gender differences or to learn about margins of resolution used in the reports in different organs, and with different modalities, such as MRI or CT and many more. We will use the results of this research to determine if there are differences regarding the accuracy of the measurements (e.g. numbers with or without decimal places) in connection with social categories such as gender or age.

As mentioned above, *MedCorpInn* is a corpus project still in progress. It is the first large corpus of its kind for the German language and will be used to research (discourse) linguistic as well as gender medicine research questions. Additionally, the specific type of data will also help furthering NLP methods for German.

## 4. References

Sarah Atkins and Kevin Harvey. 2010. How to use corpus linguistics in the study of health communication. In: Anne O'Keeffe and Michael McCarthy, ed., *The Routledge Handbook of Corpus Linguistics*. Routledge, New York.

Tianrun Cai, Andreas A. Giannopoulos, Sheng Yu, Tatiana Kelil, Beth Ripley, Kanako K. Kumamaru, Kanako K. Kumamaru, Frank J. Rybicki, and Dimitrios Mitsouras. 2016. Natural Language Processing Technologies in Radiology Research and Clinical Applications. In: *Radiographics: a review publication of the Radiological Society of North America*, Inc 36 (1), S. 176–191. DOI: 10.1148/rg.2016150080.

Paul Crawford, Brian Brown, and Kevin Harvey. 2014. Corpus linguistics and evidence-based health communication. In: Chou Wen-ying, Sylvia Hamilton, Heidi Ehernberger, ed., *The Routledge Handbook of Language and Health Communication*, pages 75–90. Routledge, London, New York.

Wendy W. Chapman, Prakash M. Nadkarni, Lynette Hirschman, Leonard W. D'Avolio, Guergana K. Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. In: *Journal of the American Medical Informatics Association*: JAMIA 18 (5): S. 540–543. DOI: 10.1136/amiajnl-2011-000465.

Zsófia Demjén. 2020. *Applying linguistics in illness and healthcare contexts*. Bloomsbury Academic, London, New York.

Johannes Hellrich, Franz Matthies, Erik Faessler, and Udo Hahn. 2015. Sharing models and tools for processing German clinical texts. In: *Studies in health technology and informatics*, 210: 734–738.

Karoline Irschara, Claudia Posch, and Bernhard Glodny. 2017. *KARBUN. Korpus radiologischer Befunde der Universitätskliniken für Neuroradiologie und Radiologie.* Leopold-Franzens-Universität Innsbruck.

Charles E. Kahn, Curtis P. Langlotz, Elizabeth S. Burnside, John A. Carrino, David S. Channin, David M. Hovsepian, and Daniel L. Rubin. 2009. Toward best practices in radiology reporting. In: *Radiology* 252 (3): S. 852–856. DOI: 10.1148/radiol.2523081992.

Douglas Maynard and John Heritage. 2005. Conversation analysis, doctor-patient interaction and medical communication. In: *Medical education* 39 (4): 428–435.

Florian Menz. 2011. Doctor-Patient-Communication. In: Ruth Wodak, Barbara Johnstone, Paul Kerswill, ed., *The Sage Handbook of Sociolinguistics*, pages 330–344. Sage, Los Angeles, Calif.

Bruce I. Reiner. 2012. Using Analysis of Speech and Linguistics to Characterize Uncertainty in Radiology Reporting. In: *J Digit Imaging*, 25: 703–707. https://doi.org/10.1007/s10278-012-9535-x.

RNSA. 2017. RadLex Term Browser, available from http://radlex.org/ (cited 2020 Sept 7).

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A Short Review of Ethical Challenges in Clinical Natural Language Processing. In: *arXiv preprint* arXiv:1703.10090.

Charlotte Taylor and Anna Marchi, ed. 2018. *Corpus Approaches to Discourse*. Routledge, New York.

TEI Consortium. 2020. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Release 4.0.0.

Viola Wiegand and Michaela Mahlberg. 2019. *Corpus Linguistics, Context and Culture*. De Gruyter, Berlin, Boston.

# XML-Encoding of a spoken Serbian corpus targeting forms of address

**Dolores Lemmenmeier-Batinić[1], Nikola Ljubešić[2], Tanja Samardžić[3]**

[1] Slavisches Seminar, University of Zurich
Plattenstrasse 43, 8032 Zurich
dolores.lemmenmeier@uzh.ch
[2] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
nikola.ljubesic@ijs.si
[3] URPP Language and Space, University of Zurich
Freiestrasse 16, 8032 Zurich
tanja.samardzic@uzh.ch

## 1. Introduction

Serbian has long been an under-resourced language. However, following the recent global trend to represent languages with corpora, several projects have attempted to amend this lack of publicly available resources regarding the written register (Erjavec, 2012; Ljubešić and Klubička, 2014; Miličević and Ljubešić, 2016). Resources for spoken standard Serbian in interaction are still lacking, despite there being some work on Serbian dialectal corpora (Vuković et al., 2019) and repositories for speech recognition (Suzić et al., 2014). Meanwhile, tools and resources of spoken registers that are being created for similar South Slavic languages such as Croatian (Kuvač Kraljević and Hržica, 2016: hrAL) and Slovenian (Verdonik et al., 2013: GOS) are gaining in popularity.

Since collecting data of spoken interaction requires not only field research, but also intensive manual work, it is fruitful to start addressing this lack of resources by gathering existing material and arranging it in a way to make it useful for a larger audience. In this paper, we present the process of compiling a corpus of spoken Serbian starting from an existing collection of transcripts that have been gathered for investigating the use of forms of address in spoken Serbian (Ulrich, 2018). Given its relatively substantial size of 172,345 tokens and 10,061 types (in comparison, hrAL contains 250,000 tokens), a fine-grained transcription, and a complete documentation of metadata, the collection of interviews gathered by Ulrich (2018) presents a valuable linguistic resource for spoken Serbian. The current version of the corpus can be accessed at a SWITCHdrive public link.[1] For the long-term deposit, we plan to use CLARIN.Sl.

After describing the data, we show the procedure of detecting and correcting transcription inconsistencies, enriching the corpus with linguistic information, and converting this resource into a standard XML format. Our goal is to add value to this resource, underlining at the same time the need for data reuse and the need for sharing data from the start in similar future projects.

## 2. Data source

### 2.1. Recordings and metadata

The corpus consists of transcripts of audio-recorded biographical interviews with 19 participants (9 female, and 10 male) recorded in 2008 and 2009, lasting 63 minutes on average. The interviewees are asked about the forms of address they use in colloquial and in formal settings, and about attitudes and evaluations concerning particular forms of address. There is a set of questions asked in each interview, but the conversations are casual and often contain short anecdotal information. The majority of the participants at the time of recording resided in Niš and in Belgrade, and have a university degree. They predominantly speak in standard spoken Serbian, but sometimes also use regional varieties of particular forms.

### 2.2. Transcription

The records were transcribed according to the GAT transcription system (Selting et al., 1998; 2009), which differentiates between three levels of granularity in transcribing talk-in-interaction: minimal, basic and fine

[1] https://drive.switch.ch/index.php/s/oNpLQcsiRDojzuG (09.09.2020).

transcripts. Features belonging to all tree levels of granularity were used in Ulrich's (2018) transcriptions, although no convention has been adopted entirely. An excerpt of one of the transcripts is given in Example 1.

Example 1: Excerpt of an original transcript (transcript id: F2)

S: tako (-) .h kako osloviš članove (-) tvoje porodice
M: <<lachend> oslovljavaš> (--) to je pra= (e) perfektivni glagol
S: e (-) da
M: <<langsam> oslovljavaš>
S: dobro to (.) to sam još mislila (-) da li je možda (-) bolje (schreibt) .h (-) vljavaš (-) e (.) dobro (-) <<leise> (xxx)>
M: članove (-) ↑a (.) i možda bolje (--) čla= (-)
S: aha (.) SVOje porodice (-) ups (-) da
M: (-) ili tvoje (---) kako oslovljavaš [članove svoje po]
S: [pa ti: (--) svoje] dadada
M: svoje (-) (svoje) (-) mislim možeš i da kažeš tvoje nije tako strašno
S: <<p> dobr↑o>

Transcribing inconsistencies, mixing standards and typing errors were inevitable, since the transcripts were made without using transcription software that could control the syntax of GAT conventions.[2] Same information was occasionally annotated with different types of parentheses, and annotations for uncertain segments or for transcript gaps were sometimes used for marking comments, and vice versa. For instance, in Example 1, the same parenthesis type was used to annotate uncertain words "(svoje)" [English: "your own"] as well as non-verbal events "(schreibt)" [English: "writes"]. In rare cases, annotations that are not mentioned in GAT were used (for instance: * - <), and some GAT conventions were used for annotating other things than those described in the manual (for example, transcribing "Brankice=e" instead of "brankice:" for marking that the last syllable is long). Metalinguistic information was mostly given in German, but sometimes also in Serbian (for instance: "smeje se" [Serbian] and "lacht" [German] for annotating laughter).

## 3 Corpus compilation

### 3.1. Preprocessing

In order to convert the raw text files to an XML format, we first normalised the white space inconsistencies, and deleted the symbols that are irrelevant for further processing steps, such as the various codes for writing quotations marks (‘ ' ’ , , „ ” “). As shown in section 2.2., the inconsistencies were unsystematic, and correcting them automatically would have led to more errors in further processing steps. We wanted to ensure the consistency of the annotation in order to be able to process and categorise it in the next steps. Therefore, we used regular expressions to extract all the unique occurrences of annotations of a particular type, and saved them in separate files. We corrected and categorised the annotations in order to convert them more easily to TEI[3] conventions in the following steps (see Table 1).

| Original annotation | English translation | Changes (intermediate step) |
|---|---|---|
| {Auslassung 14:58-15:53} | *omission 14:58-15:53* | ((gap:extent: 55s)) |
| ((Exkurs über Mathe-Lehrerin nicht transkribiert)) | *digression about the math teacher was not transcribed* | ((gap:reason: deo o učiteljici matematike nije transkribovan)) |
| {Telefon klingelt} | *the phone is ringing* | ((incident: zvoni telefon)) |
| ((klopft auf den Tisch)) | *knocking on the table* | ((incident: kuca o sto)) |

Table 1: Categorising comments in the preprocessing phase (excerpt)

Once each annotation has been checked, we replaced the original data with the corrections. Totally, 693 annotations were checked, out of which 604 have been changed. The most common corrections regarded annotations for metalinguistic comments and the use of the equals sign "=".

[2] See FOLKER transcription software: https://exmaralda.org/de/folker-de/ (09.09.2020).
[3] https://tei-c.org/release//doc/tei-p5-doc/en/html/TS.html (09.09.2020).

Since the interviews were meticulously transcribed, we aimed to keep as many annotations as possible for further processing. However, we decided not to consider the conventions of the "fine transcript" (Selting et al., 2009) in the first release of the corpus because they were either sporadically used (annotation of pitch peaks and accentuation) or required extensive manual corrections (annotations of loudness and speed). As shown in Table 1, in order to make sure that comments and tokens are written in the same language, we also translated transcriber's comments from German to Serbian.

### 3.2. XML conversion

We converted the preprocessed files into XML format following the TEI conventions for transcriptions of speech. The transcripts were segmented for each turn, and each word in a turn was segmented as well. In addition to lemmatised and normalised forms (@lemma, @norm), we provided MULTEXT-East morphosyntactic specifications (@msd) and universal pos tags (@pos), as shown in Example 2. We marked unclear segments, deletions, gaps, incidents, vocal elements, and pauses.

Example 2: XML version of a part of the excerpt shown in Example 1

```
<u who="#F2" xml:id="F2-u6">
   <w lemma="oslovljavati" "  pos="VERB" msd="Vmr2s" xml:id="F2-u6-w1">oslovljavaš</w>
   <pause length="middle" xml:id="F2-u6-p2"/>
   <w lemma="taj" pos="DET"  msd="Pd-nsn"  xml:id="F2-u6-w3">to</w>
   <w lemma="biti" pos="AUX"  msd="Var3s"  xml:id="F2-u6-w4">je</w>
   <del type="truncation" xml:id="F2-u6-w5">pra</del>
   <unclear>
      <w lemma="e" pos="INTJ"  msd="I"  xml:id="F2-u6-w6">e</w>
   </unclear>
   <w lemma="perfektivan" pos="ADJ"  msd="Agpmsny" xml:id="F2-u6-w7">perfektivni</w>
   <w lemma="glagol" pos="NOUN"  msd="Ncmsn" xml:id="F2-u6-w8">glagol</w>
</u>
```

### 3.3. Normalisation

In order to normalise the transcribed data, we detected and automatically replaced all the tokens in which standard forms were reduced (for instance, "išo" instead of "išao" [English: "went"]; "kolko" instead of "koliko" [English: "how much"], etc.) by comparing the transcribed tokens with the tokens in Serbian lexicon srLex (Ljubešić et al., 2016). We stored the normalised tokens in the @norm attribute. Then, we extracted a list of all the tokens that did not occur in srLex and checked them manually. Out of 387 types, 119 were correct, although they were not present in our lexicons (mostly uncommon words, proper names, or slang expressions). The remaining 268 types were either (lowercase) proper names or reduced standard forms, which we corrected and marked as @norm, or they were due to orthographic and typing errors like "osnačavaju" instead of "označavaju" [English: "they mark"], which we marked as original transcriptions (@orig). In total, 3,824 tokens (2.2%) and 962 types (9.6%) were affected by the normalisation.

### 3.4. Tagging the corpus

For automatic annotation of the (normalised) corpus with morphosyntactic and lemma information we used the state-of-the-art tagger for Serbian and other South-Slavic languages CLASSLA-StanfordNLP (Ljubešić and Dobrovoljc, 2019)[4] - a fork of the StanfordNLP tagger with a series of improvements, especially on the lemmatisation level. The estimate of the accuracy on standard data of this tagger for Serbian is 95.23 F1 for morphosyntax and 97.89 F1 for lemmatisation.[5] However, given that this corpus consists of spoken data transcriptions, we annotated the corpus with a novel model trained on a union of all available training data for Serbian and Croatian, namely the SETimes.SR corpus of newspaper texts (Batanović et al., 2018), the hr500k Croatian reference training corpus (Ljubešić et al., 2016), the ReLDI-NormTagNER corpus of Serbian and Croatian tweets (Miličević and Ljubešić, 2016), and the RAPUT corpus of Croatian non-professional writing (Štefanec et al., 2016).

[4] https://github.com/clarinsi/classla-stanfordnlp (09.09.2020).
[5] https://github.com/clarinsi/babushka-bench (09.09.2020).

## 4. Conclusion and future work

We presented a number of processing steps needed for turning a collection of raw transcripts into a standardised spoken language corpus accessible to a wider community. These steps consisted in a) resolving the inconsistencies in the original transcripts, b) converting the transcripts into XML, c) enriching the data with normalisation and part-of-speech annotation, and d) publishing the corpus. The corpus can be used for investigating peculiarities of spoken Serbian in talk-in-interaction, for studying disfluencies in spontaneous speech, as well as for studying Serbian spoken as native and foreign language. We are currently working on the alignment of corpus turns with the respective audio-segments. The original transcripts contain some additional mark-up that has been left out in the current XML version, such as the annotation of loudness and speed. The integration of these items, as well as the correction of automatic annotations, is left for future work.

## 5. References

Vuk Batanović, Nikola Ljubešić and Tanja Samardžić. 2018. SETimes.SR – A Reference Training Corpus of Serbian. In: *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)*, pages 11-17, Ljubljana, Slovenia.

Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Lang Resources & Evaluation*, 46(1):131–142.

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, Gothenburg, Sweeden.

Nikola Ljubešić, Filip Klubička, Željko Agić and Ivo-Pavao Jazbec. 2016. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4264-4270, Portorož, Slovenia.

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.

Jelena Kuvač Kraljević and Gordana Hržica. 2016. Croatian Adult Spoken Language Corpus (HrAL). *Fluminensia: Journal for philological research*, 28(2):87-102.

Maja Miličević and Nikola Ljubešić. 2016. Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research,* 4(2):156-188.

Margret Selting, Peter Auer, Birgit Barden, Jörg Bergmann, Elizabeth Couper-Kuhlen, Susanne Günthner, Uta Quasthoff, Christoph Meier, Peter Schlobinski and Susanne Uhmann. 1998. Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte* 173, 91-122.

Margret Selting, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzlufft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte, Anja Stukenbrock and Susanne Uhmann. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion,* (10):353-402.

Siniša Suzić, Stevan Ostrogonac, Edvin Pakoci and Milana Bojanić. 2014. Building a Speech Repository for a Serbian LVCSR System. *Telfor Journal,* 6(2):109-114.

Vanja Štefanec, Nikola Ljubešić and Jelena Kuvač Kraljević. 2016. Croatian Error-Annotated Corpus of Non-Professional Written Language. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* pages 3220-3226, Portorož, Slovenia.

Darinka Verdonik, Iztok Kosem, Anna Zwitter Viter, Simon Krek and Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation,* 47(4):1031-1048.

Teodora Vuković, Nora Muheim, Olivier Winistörfer, Ivan Šimko, Anastasia Makarova and Sanja Bradjan. 2019. Corpora and Processing Tools for Non-Standard Contemporary and Diachronic Balkan Slavic. In: *Proceedings of the Student Research Workshop (RANLPStud 2019)*, pages 62-68, Varna, Bulgaria.

Sonja Ulrich. 2018. *Anredeformen im Serbischen*. Slavistische Beiträge (508), Wiesbaden.

# Modul Leto plus – prvi korak do korpusa slovenščine kot tujega jezika

**Mojca Stritar Kučuk**

Oddelek za slovenistiko, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
mojca.stritarkucuk@ff.uni-lj.si

## 1. Uvod

Korpusi usvajanja tujega jezika so plodno raziskovalno in aplikativno področje.[1] Največkrat je sicer vezano na angleščino kot tuji jezik, obstajajo pa tudi korpusi za druge jezike, kot so hrvaščina (Mikelić Preradović et al., 2015), češčina (Hana et al., 2010), ruščina (Kutuzov in Kunilovskaya, 2014), nemščina (Reznicek et al., 2012), švedščina (Hammarberg, 2010) ali arabščina (Alfaifi et al., 2014). O korpusih usvajanja slovenščine kot tujega jezika je bilo že precej govora, vendar je bilo v glavnem omejeno na teoretične razmisleke in pilotne projekte (Stritar, 2012). Do konkretne izdelave korpusa usvajanja slovenščine kot tujega oz. drugega jezika pa še ni prišlo, predvsem zaradi finančnih oz. človeških razlogov. Izpeljava takih, manjših projektov, ki so zanimivi za ožjo skupino uporabnikov, je namreč nemalokrat odvisna od prizadevnosti posameznikov.

V zadnjem letu se je projekt izdelave korpusa usvajanja slovenščine kot drugega oz. tujega jezika končno začel. Čeprav ne čisto terminološko ustrezno – ne nazadnje gre pri večjem delu vključenih besedil za slovenščino kot drugi in ne tuji jezik –, smo korpusu zaradi ekonomičnosti nadeli poimenovanje KOST (= korpus slovenščine kot tujega jezika). V tem prispevku bodo predstavljeni prvi koraki na poti do pravega, splošneje uporabnega korpusa.

## 2. Zametek projekta: modul Leto plus

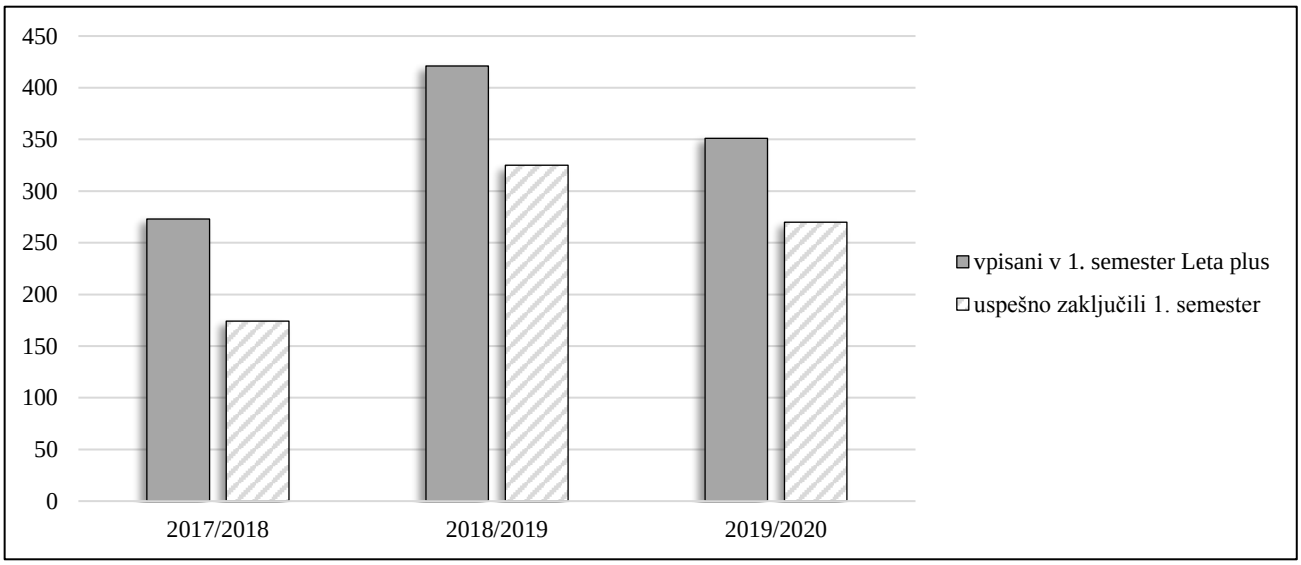
Za zagon izdelave KOST-a je ključna odločitev Univerze v Ljubljani, da je v okviru projekta *Izboljšanje procesov internacionalizacije slovenskega visokega šolstva* v študijskem letu 2016/17 pilotno izvedla Leto plus. Gre za poseben modul, namenjen tujim študentom, redno vpisanim v študijske programe Univerze v Ljubljani. V nadaljnjih študijskih letih se Leto plus[2] izvaja kot eden od ukrepov internacionalizacije univerze, financirano pa je iz razvojnega stebra univerzitetnega financiranja.

Najpomembnejša stvar, ki jo Leto plus ponuja redno vpisanim tujim študentom, je, da jim hkrati z rednim študijem omogoča brezplačno učenje slovenščine na dveh lektoratih v obsegu treh študijskih ur na teden. To pomeni okoli 90 kontaktnih ur v predavalnici, pridružuje pa se jim še okoli 30 ur dodatnih projektov, ki so izpeljani izven predavalnice ali pa jih študenti izvedejo sami.

Modul Leto plus je za zagon KOST-a idealen, saj imamo v njem zagotovljen dostop do večjega števila govorcev slovenščine kot tujega jezika (slika 1).

[1] Za rastoči seznam obstoječih korpusov prim. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

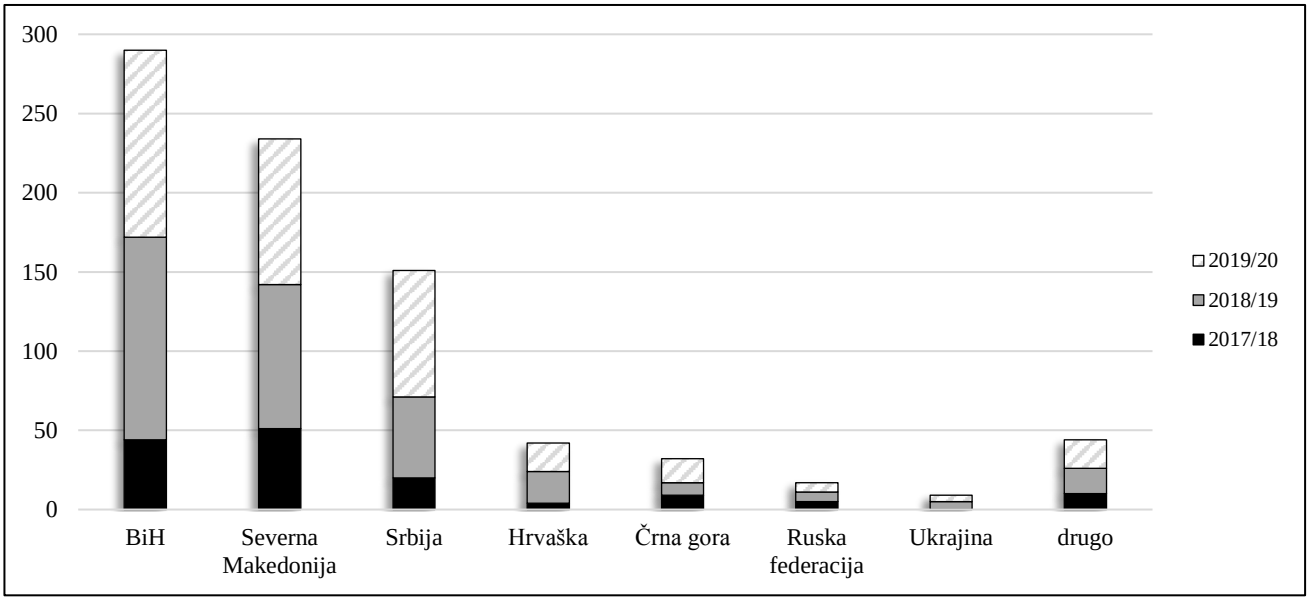
[2] https://www.uni-lj.si/studij/leto-plus.




Slika 1. Število tujih študentov v modulu Leto plus.[3]

Gre za govorce s precej homogenim jezikovnim ozadjem. Kot kaže slika 2, jih največ prihaja iz Bosne in Hercegovine, Srbije in Makedonije. Tudi njihova jezikovna zmožnost v slovenščini je precej primerljiva: v prvem, zimskem semestru večinoma začnejo brez predhodnega znanja slovenščine, zaradi skupne slovanske osnove pa hitro napredujejo in so ob koncu letnega semestra sposobni jezikovno (pre)živeti v slovenskem okolju, komunicirati o vsakdanjih temah in do neke mere v slovenščini opravljati redne študijske obveznosti. Dodatna prednost Leta plus je, da s temi govorci slovenščine kot tujega jezika nimamo samo stika, pač pa od njih redno tudi dobivamo besedila v slovenščini. Za pristop k izpitu morajo namreč oddajati pisne domače naloge.

Ne nazadnje je za KOST pomembno tudi, da smo pedagoške delavke na Letu plus večinoma redno zaposlene lektorice, tako da izdelava korpusa sodi med naše redne delovne obveznosti.




Slika 2. Udeleženci Leta plus po državah. Podatki so za zimski semester.

[3] Podatki na grafu so varljivi: največ študentov je bilo sicer v študijskem letu 2018/19, vendar le zaradi tega, ker vpis v modul ni bil omejen. Interes za Leto plus dejansko narašča; v študijskem letu 2019/20 je bilo zavrnjenih več kot sto študentov, ker zanje ni bilo mogoče zagotoviti mest na lektoratu.

# 3. Zbiranje besedil v okviru Leta plus

Čeprav dobra praksa uči, da je treba zasnovo jezikovnih korpusov načrtovati pred začetkom same gradnje, smo se izdelave KOST-a lotili oportunistično. Leto plus je namreč popoln vir za pridobivanje korpusnih besedil, ki bi ga bilo škoda ne izkoristiti.

Z zbiranjem gradiva sem pričela v študijskem letu 2018/19, v študijskem letu 2019/20 pa smo pri zbiranju že sodelovale vse redno zaposlene lektorice na Letu plus. Za uskladitev zbirke besedil in podatkov o tvorcih skrbim avtorica tega prispevka. Do sedaj je bilo zbranih skoraj 1500 pretežno krajših besedil, ki jih je napisalo skoraj 290 tvorcev. Pri enotnem zbiranju in shranjevanju besedil smo sprejeli nekaj odločitev, podrobneje pojasnjenih v nadaljevanju.

## 3.1. Dovoljenje za uporabo besedil in osebni podatki

Ker sta ureditev pravic za uporabo podatkov in varovanje osebnih podatkov ključnega pomena, so vsi študenti dobili v podpis izjavo, s katero dovoljujejo vključitev besedil, ki jih pišejo pri predmetih v okviru modula Leto plus v tekočem študijskem letu, v KOST. K izjavi sodi zbiranje osebnih podatkov, ki so nujni za analizo korpusnega gradiva: spol, starost, fakulteta, letnik in stopnja študija, izobrazba, prvi jezik in ostali jeziki, ki jih znajo govorci, ter podatki o morebitnem predhodnem učenju slovenščine ali bivanju v Sloveniji. Ti podatki bodo vključeni v glave besedil v KOST-u. V korpusu bodo tvorci anonimni, osebni podatki, ki se pojavljajo v besedilih, pa bodo odstranjeni oz. prekriti s kodami.

Izjavo, ki so jo pravno preverili na Oddelku za upravljanje s tveganji in varstvo osebnih podatkov na Univerzi v Ljubljani, smo študentom na Letu plus v podpis ponudile njihove lektorice. Pred podpisom smo jim natančno razložile projekt in pogoje sodelovanja. Razveseljivo je, da so v preteklih dveh študijskih letih izjavo podpisali vsi, ki jim je bila ponujena.

Vse podpisane izjave so digitalizirane in shranjene v obeh oblikah, papirni in digitalni.

## 3.2. Podatki o besedilih

Kar se tiče prvega jezika tvorcev besedil, oportunistično zberemo vsa besedila, ki jih dobimo. V skladu s študentsko populacijo Leta plus so zaenkrat prvi jeziki skoraj treh četrtin tvorcev srbščina, bosanščina ali hrvaščina. Pri označevanju podatka o prvem jeziku tvorca upoštevamo tisto, kar sami napišejo v prej omenjeni izjavi. Šele analiza zbranega korpusnega gradiva bo pokazala, ali prihaja med govorci teh treh jezikov do pomembnejših razlik pri usvajanju slovenščine. Prvi jezik za veliko skupino, to je kar četrtino govorcev, je makedonščina. Sicer pa so med zbranimi besedili še besedila govorcev s prvimi jeziki češčino, hebrejščino, italijanščino, madžarščino, nemščino, ruščino, slovaščino in ukrajinščino.

Študenti na lektoratu slovenščine tvorijo različnejše vrste besedil. Največ je esejev oz. spisov na različne teme (npr. o družini, prehrani, zdravju), nekaj pa je tudi praktičnega pisanja (npr. življenjepis, prošnja za delo). Ker je v praktičnih besedilih veliko osebnih podatkov, ki bi jih bilo treba zakrivati, in le malo dejanske tvorbe besedil, smo se odločili, da jih v KOST vključujemo v manjši meri.

Tri četrtine do sedaj zbranih besedil so študenti napisali doma kot domačo nalogo. Pri tem je nemogoče nadzirati, ali jih dejansko napišejo sami in koliko si pomagajo z jezikovnimi viri in orodji. V KOST bodo vključena vsa, če pa tisti, ki pripravlja vnos, sumi, da gre za uporabo spletnega prevajalnika ali pomoč rojenega govorca slovenščine, to zabeleži kot posebno opombo. Preostanek predstavljajo besedila, napisana na izpitu, torej pod bolj nadzorovanimi pogoji tvorjenja.

Pomembna je tudi jezikovna zmožnost tvorcev besedil. Večina študentov začne brez predznanja slovenščine, ker so z južnoslovanskega govornega področja, pa je njihov napredek hiter. V korpus zaenkrat vključujemo njihova besedila, ki jih napišejo na izpitu ob koncu prvega semestra in dobijo pozitivno oceno, ter besedila iz drugega semestra učenja slovenščine. Če je to njihovo prvo leto učenja, stopnjo jezikovne zmožnosti označimo kot »Južni Slovan začetnik«. Tako zbranih besedil je trenutno dobrih 70 %, preostalo pa so besedila nadaljevalcev in izpopolnjevalcev, ki bi jih v korpusu sicer želeli imeti več. V nadaljnjih fazah gradnje bomo v KOST predvidoma vključili tudi besedila začetnikov iz drugih govornih področij.

### 3.3. Način zbiranja besedil

Velika večina besedil, ki jih oddajajo študenti na Letu plus, je napisana na računalnik. Študenti jih oddajajo po elektronski pošti ali prek spletne učilnice na Filozofski fakulteti. Digitalna oblika zbiranje besedil seveda zelo olajša. Letni semester 2019/20 je bil glede dostopa do takih besedil zelo produktiven, saj se je zaradi pandemije celotno poučevanje preselilo v digitalno okolje.

Besedila, ki nastanejo na izpitu ali med lektoratom v razredu, pa so napisana na roko in jih je treba pretipkati. V skladu s svojimi časovnimi zmožnostmi to opravljamo lektorice Leta plus, saj predvidevamo, da bi se med besedili, ki nastanejo kot domače naloge, in tistimi, ki so napisana v bolj stresnih izpitnih okoliščinah, lahko pokazale tudi jezikovne razlike. Zaenkrat je med zbranimi besedili takih, ki so bila napisana na roko, slaba tretjina. Če bodo prve raziskave na gradivu pokazale, da razlik ni, in če se bo izkazalo, da nam za to delo primanjkuje časa, pa se bomo omejili samo na digitalno oddano gradivo.

### 3.4. Priprava besedil za vključitev v KOST

Besedila bodo v KOST-u točno taka, kot jih napišejo študenti, brez kakršnih koli jezikovnih popravkov. Ob tem si dovoljujemo izjemo. Ker pravopis ni v ospredju raziskav pri slovenščini kot tujem jeziku in ker bi radi s tem zvišali uspešnost morebitnega avtomatskega označevanja korpusnega gradiva v prihodnosti, v besedilih popravimo stičnost ločil ter odstranimo dvojne presledke. Zaradi lažje berljivosti smo se odločili tudi, da pri pretipkavanju besedil, prvotno napisanih samo z velikimi tiskanimi črkami, uporabimo male tiskane črke in pri tem upoštevamo slovenska pravopisna pravila, pa čeprav jih tvorec besedila morda ne bi.

Že v fazi zbiranja besedil smo se povezali s Centrom za jezikovne vire in tehnologije, kjer so nam pomagali pri nekaterih tehničnih vidikih. Vsako besedilo, vključeno v KOST, je shranjeno v samostojno tekstovno datoteko. V Excelovi tabeli so zbrani metajezikovni podatki o tvorcih in besedilih. Na podlagi te tabele in tekstovnih datotek z besedili bo zgeneriran korpus, primeren za analizo in objavo na spletu.

## 4. Nadaljnji koraki

Z začetkom študijskega leta 2020/21 smo zbiranje gradiva za KOST razširili še na druge institucije, neposredno povezane s slovenščino kot tujim jezikom. Najprej je to seveda Center za slovenščino kot drugi in tuji jezik (CSDTJ) Filozofske fakultete Univerze v Ljubljani,[4] na katerem imajo stalen dotok besedil tujejezičnih govorcev slovenščine zagotovljen v programih Tečaji slovenščine, Izpitni center, Seminar slovenskega jezika, literature in kulture, Slovenščina na tujih univerzah ter Slovenščina za otroke in mladostnike. Ker so programi CSDTJ za razliko od Leta plus manj vezani na govorce iz držav nekdanje Jugoslavije, računamo, da bo z vključitvijo teh besedil KOST zajemal bistveno več prvih jezikov. Če bo mogoče, bomo njihove deleže uravnotežili, tako da KOST ne bo samo korpus južnoslovanskih govorcev slovenščine kot tujega jezika.

KOST bo najprej dostopen v surovi, neoznačeni obliki. Vendar je največja dodana vrednost korpusov usvajanja jezika v označenosti napak. Jezikoslovne dileme ob tem so bile do neke mere rešene že v nekaterih predhodnih raziskavah in v obstoječih jezikovnih virih (Kosem et al., 2012, korpus lektorskih popravkov Lektor),[5] za tehnološki vidik pa še iščemo najustreznejšo rešitev, ki bi omogočala učinkovito dodajanje oznak napak v korpusna besedila. Spomladi 2020 je bil KOST vključen v uspešno prijavo na razpis Razvoj slovenščine v digitalnem okolju.[6] S tem so bila pridobljena sredstva za razvoj orodja za označevanje jezikovnih napak oz. njihovih popravkov v besedilih ter za razvoj novega konkordančnika, ki bo omogočal napredno iskanje po jezikovnih napakah in popravkih, njihovo vizualizacijo in izvažanje korpusnih podatkov. Šele ko bo poskrbljeno za to, se bomo lahko lotili zamudnega označevanja besedil in spotoma ocenili še, kolikšen delež korpusnih besedil je smiselno in izvedljivo označevati. Predvidoma bo delno označeni KOST javno dostopen na spletu leta 2022.

[4] https://centerslo.si/.
[5] http://korpus-lektor.net.
[6] https://www.gov.si/zbirke/javne-objave/javni-razpis-razvoj-slovenscine-v-digitalnem-okolju-jezikovni-viri-in-tehnologije/.

Zaradi stalnega dotoka besedil je KOST lahko zbirka, ki raste in se dopolnjuje, posebno dodano vrednost pa vidimo tudi v možnosti longitudinalnih raziskav. Kar nekaj govorcev slovenščine kot tujega jezika namreč prehaja med različnimi programi na Letu plus in CSDTJ, vse od začetne stopnje učenja do najvišjega, izpopolnjevalnega nivoja. Kako zagotoviti sledljivost in hkrati tvorcev tovrstnih besedil, ostaja vprašanje, ki ga bomo morali še rešiti.

## 5. Literatura

Abdullah Alfaifi, Eric Atwell in Ibraheem Hedaya. 2014. Arabic Learner Corpus (ALC) v2: A New Written and Spoken Corpus of Arabic Learners. V: *Proceedings of the Learner Corpus Studies in Asia and the World*, Kobe. https://www.dropbox.com/s/646ysp9q9h7353v/ALFAIFI%20LCSAW2014.pdf?dl=0.

Björn Hammarberg. 2010. *Introduction to the ASU Corpus: a longitudinal oral and written text corpus of adult learner Swedish with a corresponding part from native Swedes. Version 2010-11-16*. Stockholm University, Stockholm. http://su.diva-portal.org/smash/get/diva2:778204/FULLTEXT01.pdf.

Jirka Hana, Alexandr Rosen, Svatava Škodová in Barbora Štindlová. 2010. Error-tagged Learner Corpus of Czech. V: *Proceedings of the Fourth Linguistic Annotation Workshop*, str. 11–19, ACL, Uppsala.

Iztok Kosem, Mojca Stritar, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012. *Analiza jezikovnih težav učencev: korpusni pristop*. Trojina, zavod za uporabno slovenistiko, Ljubljana.

Andrey Kutuzov in Maria Kunilovskaya. 2014. Russian Learner Translator Corpus: Design, Potential and Application. V: *Text, Speech and Dialogue*, str. 315–323. Springer International Publishing.

Nives Mikelić Preradović, Monika Berać in Damir Boras. 2015. Learner Corpus of Croatian as a Second and Foreign Language. Multidisciplinary Approaches to Multilingualism. Ur. Kristina Cergol Kovačević, Sanda Lucija Udier. Peter Lang, Frankfurt am Main. 107–126.

Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann in Torsten Andreas. 2012. *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01*. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2/.

Mojca Stritar. 2012. *Korpusi usvajanja tujega jezika*. Zveza društev Slavistično društvo Slovenije, Ljubljana.

# Projekt Vojaške žrtve prve svetovne vojne na Slovenskem

**Andrej Pančur,* Neja Blaj Hribar,* Mihael Ojsteršek,* Mojca Šorn***

* Inštitut za novejšo zgodovino
Privoz 11, 1000 Ljubljana
andrej.pancur@inz.si, neja.blaj-hribar@inz.si, mihael.ojstersek@inz.si, mojca.sorn@inz.si

## 1 Uvod

Različni dogodki, ki so bili povezani z obeleževanjem stote obletnice prve svetovne vojne, so v zadnjih letih še dodatno povečale zanimanje za zgodovinske vire in raziskave o prvi svetovni vojni (1914-1918). Nekateri uspešni projekti so se odvijali na mednarodni ravni, npr. Europeana Collections 1914-1918,[1] 1914-1918-online[2] in CENDARI,[3] še več takšnih projektov je bilo na nacionalni ravni, med drugim tudi v Sloveniji.[4] Večina teh projektov se posveča omogočanju dostopa do digitalizirane kulturne dediščine o prvi svetovni vojni, ki jo hranijo ustanove s področja varstva kulturne dediščine. Nekatere ustanove pa so objavile tudi baze podatkov o (padlih) vojakih prve svetovne vojne [WW1]. (Clavert, 2018a, 2018b) V prispevku bomo predstavili digitalno bazo podatkov (vojaških) žrtev prve svetovne vojne iz Slovenije,[5] ki je nastala na podlagi sodelovanja med različnimi raziskovalnimi in kulturnimi ustanovami ter nekaterimi posamezniki. V konzorciju je sodelovalo 10 muzejev,[6] 2 arhiva[7] in 2 raziskovalni ustanovi[8] in posameznik.[9] Projekt je koordiniral Inštitut za novejšo zgodovino, ki ima dobro razvito raziskovalno skupino za digitalno humanistiko in je tudi nacionalna koordinacijska ustanova za Digital Research Infrastructure for the Arts and Humanities (DARIAH) v Sloveniji.

V nadaljevanju bomo najprej predstavili zasnovo projekta zbiranja podatkov o vojaških žrtvah prve svetovne vojne iz Slovenije ter njegove prednosti in pomanjkljivosti. V naslednjem poglavju bomo opisali postopke, orodja in tehnologije, ki smo jih uporabili pri odpravljanju teh pomanjkljivosti projekta. Na koncu bomo na kratko predstavili glavni rezultat projekta - bazo vojaških žrtev ter naše načrte za bodočnost.

## 2 Projekt zbiranja podatkov

Projekt zbiranja podatkov o žrtvah prve svetovne vojne iz Slovenije (ZV1) je dober primer uspešnega institucionalnega povezovanja med lokalnimi ustanovami s področja varstva kulturne dediščine, raziskovalnimi ustanovami in digitalno raziskovalno infrastrukturo. Ker je projekt nastal na pobudo arhivov, muzejev, knjižnic in galerij (GLAM ustanov), je tudi dober primer uspešnega reševanja izzivov (Seillier, 2018), s katerimi se soočajo pobudniki listine o ponovni uporabi podatkov kulturne dediščine: DARIAH, APEF, CLARIN, Europeana, E-RIHS, Iperion-CH, PARTHENOS.

Delovanje konzorcijskih partnerjev je temeljilo na sledečih načelih:

- **Ločeno financiranje:** Projekt smo financirali iz obstoječih dejavnostih posameznih partnerjev.
- **Uporaba obstoječih baz podatkov:** Partnerji so prispevali podatke iz svojih že obstoječih baz podatkov.
- **Ločeno zbiranje podatkov:** Partnerji so zbirali podatke prvenstveno za namene dejavnosti matičnih ustanov.
- **Odprti dostop:** Zbrane podatke so partnerji dali na voljo pod pogoji odprtih licenc Creative Commons.
- **Enotno skrbništvo podatkov:** Partnerji so zbrane podatke shranili v verificiranem repozitoriju portala Zgodovina Slovenije – SIstory.
- **Enotna baza podatkov:** Zbrani podatki vseh partnerjev so bili poenoteni in uvoženi v relacijsko MySQL bazo podatkov.

---

[1] http://www.europeana-collections-1914-1918.eu/.
[2] http://www.1914-1918-online.net/.
[3] http://www.cendari.eu/.
[4] http://www.rtvslo.si/1sv, http://www.100letprve.si.
[5] http://zv1.sistory.si/.
[6] Dolenjski muzej Novo mesto, Gorenjski muzej Kranj, Mestni muzej Idrija, Muzej novejše zgodovine Slovenije, Muzej Velenje, Muzejsko društvo Vrhnika, Posavski muzej Brežice, Tolminski muzej, Ustanova "Fundacija poti miru v Posočju", Vojaški muzej Slovenske vojske.
[7] Arhiv Republike Slovenije, Nadškofijski arhiv Ljubljana.
[8] Inštitut za novejšo zgodovino; ZRC SAZU, Zgodovinski Inštitut Milka Kosa.
[9] Renato Podbersič.

Za takšen način delovanja smo se odločili iz sledečih razlogov:

- Za namene zbiranja podatkov o vseh žrtvah prve svetovne vojne na ozemlju sedanje Slovenije ni bilo mogoče pridobiti ustreznega projektnega financiranja.
- Čeprav je bilo v okviru projekta opravljeno veliko prostovoljnega dela, so glede na svoje želje in pričakovanja intenzivnost delovanja v projektu določali partnerji sami.
- Konzorcij je združeval raziskovalce iz raznolikih raziskovalnih ustanov, z različnimi stopnjami tehničnega znanja in izkušenj. Zato so projektni partnerji sami izbrali ustrezna orodja in programe. Večina je uporabljala programe za preglednice, v prvi vrsti Microsoft Excel.

Zato pa je tako zasnovani projekt imel več pomanjkljivosti:

- **Manjkajoči podatki:** Ker so projektni partnerji večinoma zbirali podatke za samo določeno regijo, konzorciju pa se niso pridružile vse ustrezne ustanove iz Slovenije, se za nekatere regije podatkov sploh ni zbiralo.
- **Dvojniki:** Ker so podatke zbirali različni partnerji, vsak za svoje namene, imajo lahko v skupni bazi iste osebe več zapisov.
- **Neenotni podatki:** Ker so projektni parterji podatke zbirali iz različnih razlogov, spremenljivke zbranih podatkov med seboj niso bile vedno primerljive.

# 3 Odpravljanje pomankljivosti projekta

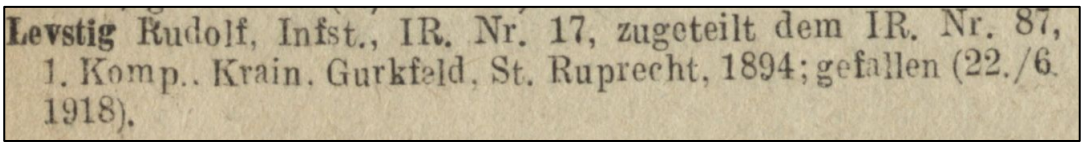
Infrastrukturna skupina iz Inštituta za novejšo zgodovino si je zadala za nalogo, da s pomočjo digitalno humanističnih metod in orodij čim bolj učinkovito razreši pomanjkljivosti projekta zbiranja podatkov o vojaških žrtvah prve svetovne vojne, ki so izhajale iz same zasnove projekta:

## 3.1 Manjkajoči podatki

Po vzoru na nemške[10] in avstrijske[11] geneološke projekte je uporabila digitalizirane Sezname izgub (Verlustliste) za celotno Avstro-Ogrsko.[12] Ti tiskani seznami, ki so redno izhajali med prvo svetovno vojno in v mesecih neposredno po njej (12. 8. 1914 – 29. 1. 1919) vsebujejo osnovne podatke o ranjenih, zajetih, mrtvih in pogrešanih avstro-ogrskih vojakih (predvidoma skoraj 3 milijone zapisov). Podatki o žrtvi so zapisani na sistematičen način (vrstni red, vejice, podpičja, oklepaji), kar nam je omogočilo polavtomatsko označevanje teh zapisov:

- ime in priimek
- čin
- vojaški oddelek in pododdelek
- dežela, okraj in kraj
- leto rojstva
- kdaj in kaj se je vojaku zgodilo: ranjen, umrl, vojni ujetnik.

Ta del projekta je potekal v treh delovnih fazah. Najprej smo na spletni strani ANNO (Historische Österreichiche Zeitungen and Zeitschriften) iz besedilnih datotek (format txt, Unicode) izluščili OCR vseh dostopnih seznamov izgub (515 datotek, 4.356.826 vrstic). Zaradi relativno slabega tiska, ki se je zaradi vojnih razmer in vedno slabšega papirja z leti samo še slabšal, je temu primerno tudi OCR vseboval napake. Na Sliki 1 je prikazan primer izseka digitalizirane slike zapisa podatkov o umrlem vojaku, na Sliki 2 pa originalni OCR tega zapisa.


Levstig Rudolf, Infst., IR. Nr. 17, zugeteilt dem IR. Nr. 87, 1. Komp., Krain, Gurkfeld, St. Ruprecht, 1894; gefallen (22./6. 1918).

Slika 1: Izsek skena.

[10] http://wiki-de.genealogy.net/Verlustlisten_Erster_Weltkrieg/Projekt.
[11] http://wiki-de.genealogy.net/Verlustlisten_%C3%96sterreich-Ungarns_1914-1919.
[12] http://anno.onb.ac.at/cgi-content/anno?aid=vll.

```
tzkan Peter, Schütze, k. k. SchR. Nr. 22, Bukowina, Radautz
1888; gestorben (3./1. 1918).
Levstig Rudolf, Infst., IR. Nr. 17, ziigcteilt dem IR. Nr. 87,
1. Komp.. Krain. Gurkfeld. St. Ruprecht, 1894;gef.illen (22./6.
1918).

Lewene^ Emilian, Lstlnfst., IR. Nr. 9, 7. Komp., Gabi
Zydaczöw, Holeszöw, 1898; gefallen (2./ i. 1918). ^
```

Slika 2: OCR zapis skena zgornje slike 1.

V drugem koraku smo besedilne datoteke pretvorili v datoteke XML. Pri tem smo se naslonili na smernice konzorcija TEI (Text Encoding Initiative). Prvotne datoteke TEI so v <teiHeader> imele zapisane osnovne metapodatke, vsaka vrstica prvotnega OCR besedila pa je bila označena kot odstavek <p>. S pomočjo regularnih izrazov smo najprej označili dele besedil z navedbo habsburških dežel (Kranjska, Štajerska, Koroška, Primorska: Goriška, Istra), katerih deli ozemlja so danes sestavni del slovenskega državnega ozemlja. Potem so sledili popravki OCR napak.

V tretjem koraku smo na podlagi niza pretvorb XSLT avtomatsko označili strukturo zapisa o žrtvah v skladu s TEI modulom Nemes, Dates, People, and Places (TEI Consortium, 2018). Končne datoteke TEI vsebujejo seznam oseb (<listPerson>), znotraj katere je vsaka žrtev iz slovenskih dežel Avstro-Ogrske označena znotraj elementa <person>. Na Sliki 3 je primer takšnega zapisa v skladu s smernicami TEI. Pri tem je čin označen z elementom <socecStatus> (socialno ekonomski status), pripadnost vojaškemu oddelku pa kot <affiliation>. Vojakovo usodo smo označili z elementom <event>, pri katerem vrsto usode določa atribut @type (death, pow, wounded, burial).

```
<person xml:id="vll.19181204_703.9082"
        facs="http://anno.onb.ac.at/cgi-content/anno?aid=vll&amp;datum=19181204&amp;seite=34&amp;zoom=33">
   <persName>
      <surname>Levstig</surname>
      <forename>Rudolf</forename>
   </persName>
   <socecStatus>Infst.</socecStatus>
   <affiliation>IR. Nr. 17, zugeteilt dem IR. Nr. 87, 1. Komp.</affiliation>
   <birth>
      <placeName key="SI" type="municipality" corresp="krn:c.040304" cert="high">
         <country>Krain</country>
         <region>Gurkfeld</region>
         <settlement>St. Ruprecht</settlement>
      </placeName>
      <date>1894</date>
   </birth>
   <event type="death">
      <label>gefallen</label>
      <desc>(<date when="1918-06-22">22./6. 1918</date>)</desc>
   </event>
</person>
```

Slika 3: Kodirane podatkov o žrtvi 1. svetovne vojne iz seznama izgub Avstro-Ogrske v skladu s Smernicami TEI.

Za uspešno izvedbo projekta so bili zelo pomembni podatki o kraju rojstva oziroma bivanja (<placeName>), saj smo v primeru nekdanjih avstrijskih dežel, ki so danes hkrati del Slovenije in sosednjih držav, potrebovali natančne geografske podatke o izvoru teh oseb. Pri tem smo se naslonili (atribut @corresp elementa <placeName>) na podatke iz avstrijskih krajevnih repertorijev (Pančur et al., 2017), ki smo jih prav tako kodirali v skladu s smernicami TEI. Od prvotnih 66.442 oseb v seznamu žrtev je na koncu ostalo 35.899 oseb iz slovenskih dežel, od tega 18.536 ranjenih, 9535 ujetih in 8790 mrtvih.

## 3.2 Dvojniki

Zaradi takšnega načina zbiranja manjkajočih podatkov, se je še dodatno povečalo število dvojnikov v bodoči skupni bazi. Ta problem smo rešili na dva načina:

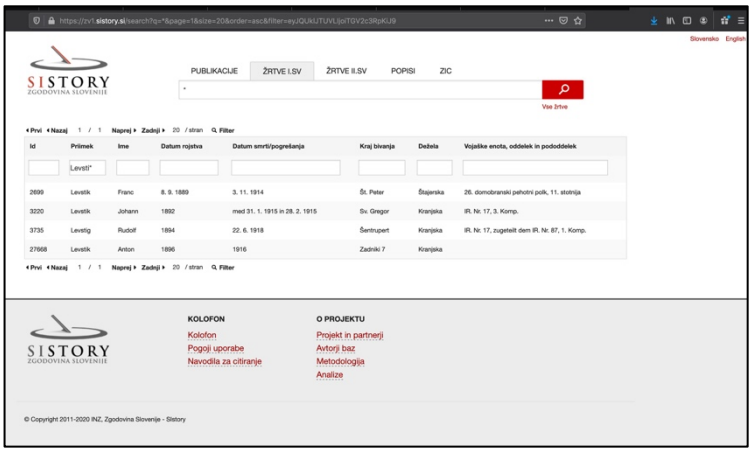
- V administracijo skupne baze smo vgradili osnovno orodje za povezovanje podatkov in deduplikacijo: ujemanje po imenu, priimku, rojstvu (datum in/ali kraj) in smrti (datum in/ali kraj).
- Za bolj napredno deduplikacijo zapisov o osebah uporabljamo:
    - R paket RecordLinkage
    - Fine-grained Record Integration and Linkage Tool (FRIL).

## 3.3 Neenotni podatki

Urejanje neenotnih podatkov preko administracije je ponavadi zelo zamudno in dolgotrajno, saj lahko na enkrat urejamo le podatke za nekaj oseb. Zato smo se odločili, da bomo podatke naknadno urejali in poenotili zunaj obstoječe baze. To urejanje tako poteka s pomočjo uvoza in izvoza vseh ali le dela podatkov iz baze v CSV formatu in naknadna obdelava teh podatkov v orodjih in okoljih (predvsem R), ki so specializirani za takšno urejanje podatkov.

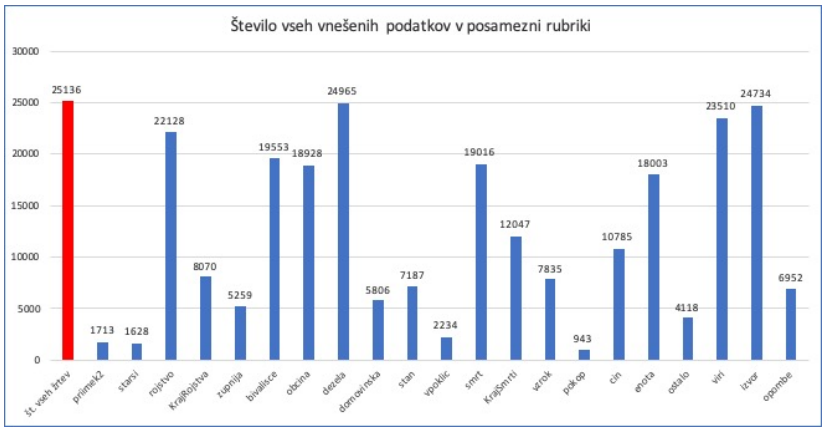
# 4 Baza žrtev

Skupna baza podatkov je dostopna na poddomeni portala Zgodovina Slovenije – SIstory kot samostojna spletna aplikacija.[13] Digitalno bazo žrtev prve svetovne vojne smo zasnovali s pomočjo klasičnih spletnih tehnologij: HTML5, CSS, MySQL, PHP7 (Laravel framework) in JavaScript. Za iskanje po bazi uporabljamo ElasticSearch.[14] Za tabelarni prikaz iskalnih rezultatov uporabljamo DataTables,[15] ki omogoča nadaljnjo filtriranje podatkov po izbranih stolpcih (glej Slika 4). Vsi podatki o žrtvi se izpišejo na ločeni strani, uporabnik si lahko te podatke naloži v formatu PDF.




Slika 4: Iskanje in filtriranje osnovnih podatkov spletne aplikacije.

Trenutno so v bazi podatki o 25.136 osebah. Redakcijska skupina projekta je določila enoten podatkovni model: priimek, druga oblika priimka, ime, očetovo/materino ime, datum rojstva, kraj rojstva, rojstna župnija, kraj bivanja, občina/okraj bivanja, dežela, domovinska pravica, stan (poklic, otroci), datum vpoklica, datum smrti/pogrešanja, kraj smrti, vzrok smrti, kraj pokopa, čin, vojaška enota, drugi podatki, viri, moje opomb, izvor. Zaradi raznovrstnosti izvornih podatkov projektnih partnerjev so podatki tega podatkovnega modela zelo različno zastopani (glej spodnji grafikon).




# 5 Načrti

[13] https://zv1.sistory.si/.
[14] https://www.elastic.co/elasticsearch/.
[15] DataTables: Table plug-in for jQuery, https://datatables.net/.

Podatki o žrtvah v zapisu TEI se trenutno prosto dostopni v GitHub repozitoriju.[16] Zato kratkoročno načrtujemo, da bomo to in nekatere ostale izvorne baze projektnih partnerjev objavili kot ločene digitalne izdaje. Prvi poskusni primer takšne ločene postavitve je že dostopen v GitHub repozitoriju[17] in spletni strani[18] ter poleg podatkov o smrtnih žrtvah vsebuje še podatke o vojnih ujetnikih in ranjencih.

V projektu smo za označevanje historičnih krajev uporabljali GeoNames. Hkrati načrtujemo še klasifikacijo dogodkov in vojaških enot. Po vzoru uspešnih Linked Open Data (LOD) projektov, ki so uporabili podatke o prvi svetovni vojni (Mäkelä et al., 2017), je naš dolgoročni cilj, da tudi podatke o žrtvah prve svetovne vojne iz Slovenije objavimo na semantičnem spletu.

## 6 Zahvala

Raziskavo je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije skozi program Raziskovalne infrastrukture slovenskega zgodovinopisja (I0-0013) in slovenske raziskovalne infrastrukture DARIAH-SI.

## Literatura

Frédéric Clavert. 2018a. Traces de mémoires collectives et bases de données de la Première Guerre mondiale. V: *L'histoire contemporaine à l'ère numérique (blog)*. https://histnum.hypotheses.org/2743.

Frédéric Clavert. 2018b. WW1 (fallen) soldiers on-line: databases as traces of collective memories. V: *Digital Humanities Benelux Conference.* International Institute for Social History, Amsterdam. http://2018.dhbenelux.org/wp-content/uploads/sites/8/2018/05/Clavert_FallenSoldiers_DHBenelux2018.pdf.

Eetu Mäkelä, Juha Törnroos, Thea Lindquist in Eero Hyvönen. 2017. WW1LOD – An application of CIDOC-CRM to World War 1 Linked Data. *International Journal on Digital Libraries* 18(4): 333-343.

Andrej Pančur, Neja Blaj Hribar in Barbara Györfi. 2017. *Krajevna imena na Slovenskem 1.0*. Zgodovina Slovenije – SIstory. http://hdl.handle.net/11686/36052.

Dorian Seillier. 2018. Assessing the survey on the Charter's principles. V: *Heritage Data Reuse Charter: Information and development around the Heritage Reuse Charter (blog)*. https://datacharter.hypotheses.org/179.

TEI Consortium (2018). *Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/P5/.

[16] https://github.com/SIstory/Verlustliste.
[17] https://github.com/sidih/verlustliste.
[18] https://sidih.github.io/verlustliste.

# Raziskovalna infrastruktura projekta RI-SI CLARIN

**Darinka Verdonik,* Matej Rojc,* Izidor Mlakar,* Danilo Zimšek,* Zdravko Kačič,* Milan Ojsteršek* in Tomaž Erjavec†**

* Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška 46, 2000 Maribor
darinka.verdonik@um.si; matej.rojc@um.si; izidor.mlakar@um.si; danilo.zimsek@um.si;
zdravko.kacic@um.si; milan.ojstersek@um.si
† Odsek za tehnologije znanja, Institut Jožef Stefan« Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

## 1 Uvod

Cilj evropske infrastrukture CLARIN (CLARIN ERIC, 2020) je spodbujanje raziskovalne dejavnosti na področju humanističnih in družbenih ved. Ta vizija se uresničuje z gradnjo in delovanjem raziskovalne infrastrukture v skupni uporabi, ki raziskovalnim skupnostim zagotavlja jezikovne vire, tehnologije in strokovno znanje. Slovenska raziskovalna infrastruktura CLARIN.SI (Erjavec et al. 2014) je vzpostavljena kot dolgoročni infrastrukturni projekt v okviru evropskega konzorcija CLARIN ERIC (European Research Infrastructures Consortium) ter naj bi olajšala in spodbujala v mednarodno okolje vpeto raziskovanje slovenščine, podporo razvoju digitalnih jezikovnih virov za slovenščino, digitalnih pripomočkov za komunikacijo med računalnikom in človekom ali namenjenih rabi v izobraževalnih okoljih ter lajšanju komunikacijskih zadreg oseb s posebnimi potrebami.

Konzorcij CLARIN.SI v mednarodnem projektu CLARIN zagotavlja prisotnost slovenskega jezika v virih in storitvah infrastrukture CLARIN. Institut »Jožef Stefan« (IJS) kot tehnični center konzorcija vzdržuje repozitorij jezikovnih virov, ki ustreza zahtevnim merilom, kakršna postavlja evropski CLARIN ERIC, in vsebuje že več kot 180 odprto dostopnih jezikovnih virov. CLARIN.SI ponuja tudi storitve za korpusno jezikoslovje in jezikovne tehnologije, na prvem mestu spletne konkordančnike z bazo več kot 50 jezikoslovno označenih korpusov. Center za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT UL) omogoča skozi portal viri.cjvt.si dostop do referenčnih korpusov slovenskega jezika (Gigafida 2.0 in Kres) tako za raziskave kot za poučevanje slovenščine, poleg tega pa njihovi strežniki ponujajo tudi dostop do slovarskih baz, npr. do Slovarja sopomenk sodobne slovenščine ter do Kolokacijskega slovarja sodobne slovenščine. S CLARIN povezano delo na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru (UM) se osredotoča na razvoj govornih tehnologij in virov.

Institut »Jožef Stefan« je skupaj s partnericama Univerzo v Ljubljani in Univerzo v Mariboru pridobil projekt »Razvoj raziskovalne infrastrukture za mednarodno konkurenčnost slovenskega RRI prostora – RI-SI – CLARIN« (s kratkim imenom »RI-SI CLARIN«) v skupni vrednosti 467.000 EUR za nakup nove raziskovalne opreme. Projekt traja od decembra 2018 do septembra 2021, financirata pa ga Ministrstvo za izobraževanje, znanost in šport in Evropski sklad za regionalni razvoj. Operacija je namenjena predvsem izboljšanju strojne opreme CLARIN.SI. Ta je pred začetkom projekta zajemala gručo treh računalnikov za namene portala www.slovenscina.eu in strežnik, ki je pokrival vse druge storitve CLARIN.SI (domača stran, repozitorij, konkordančniki, spletne storitve). Navedena strojna oprema ni omogočala posodabljanja infrastrukturnih storitev za znanstveno delo na področju računalniškega jezikoslovja in digitalne humanistike in je bila popolnoma neprimerna za eksperimentalno delo. Nova oprema, nabavljena v okviru RI-SI CLARIN, omogoča oz. bo omogočala bistveno hitrejše, intenzivnejše in visoko kakovostno vključevanje v velike mednarodne projekte. S tem prispeva h krepitvi mreže raziskovalne infrastrukture človeških virov v znanosti ter prostemu pretoku ljudi, zamisli in znanja v evropskem raziskovalnem prostoru.

Cilji že izvedenih in še načrtovanih nabav so med drugim:

- zagotoviti nadaljnje delovanje tehničnih storitev infrastrukture CLARIN.SI,
- omogočiti hranjenje velikih multimodalnih jezikovnih podatkov,
- omogočiti, da CLARIN.SI sledi paradigmi »velepodatkov« (angl. big data),
- omogočiti, da CLARIN.SI ponuja javno dostopne spletne storitve za obdelavo velikih količin slovenskih besedil,
- vzpostaviti namensko gručo računalnikov s pospeševalniki GPGPU za potrebe globokega učenja.

# 2 Nova oprema na Institutu »Jožef Stefan«

IJS v sklopu operacije nabavlja sledečo raziskovalno opremo:

1. Gruča za spletne storitve: dva strežnika z dvema 32-jedrnima procesorjema in dva strežnika z dvema 24-jedrnima procesorjema, vsi štirje z 2,3 GHz delovne frekvence, 1 TB pomnilnika, RAID krmilno kartico, 4 diski, diskovnim poljem s 24 TB in kartico za priklop na zunanja diskovna polja prek povezave FC z redundantnimi napajalniki.
2. Strežnik repozitorija: dva redundantna strežnika (eden kot produkcijski, drugi kot razvojni in pomožni) z dvema 16-jedrnima procesorjema, 2,3 GHz delovne frekvence in 768 GB pomnilnika, RAID krmilno kartico, diskovnim poljem s 24 TB in kartico za priklop na zunanja diskovna polja prek FC povezave ter redundantnimi napajalniki.
3. Diskovno polje za varnostne kopije: razširitveno diskovno polje z redundantnim priklopom in redundantnimi napajalniki s 16 diski po 10 TB kapacitete, certificirano za vgradnjo s strani proizvajalca diskovnega polja.
4. Stikalo za optični kanal: 2x nadgradnja stikala FC z 12 vrati HP SN 3000B na 24 vrat za povezavo obstoječih in novih diskovnih polj na nove strežnike.

# 3 Nova oprema na Centru za jezikovne vire in tehnologije UL

Center za jezikovne vire in tehnologije Univerze v Ljubljani je za potrebe izdelave in testiranja novih storitev nabavil strežniško rezino z dvema procesorjema po 12 jeder s 512 GB pomnilnika in predpomnilniški disk SSD velikosti vsaj 960 GB. Strežniška rezina je povezana z obstoječim diskovnim poljem s 10 TB efektivne kapacitete.

# 4 Nova oprema na Univerzi v Mariboru

Nova oprema na Univerzi v Mariboru, nabavljena do septembra 2020, vključuje gručo GPU-strežnikov, ki je optimirana za izvajanje aplikacij, temelječih na uporabi globokega učenja, strežnike za obdelavo velikih jezikovnih podatkov in diskovno polje za hranjenje velikih količin jezikovnih podatkov. V prihodnje je predvidena samo še ena nabava, tj. nadgradnja gruče GPU-strežnikov z novejšimi enotami GPU v letu 2021.

## 4.1 NVIDIA DGX-1

Sistem NVIDIA DGX-1-V100-EDU/32GB sestoji iz programske in strojne opreme, ki je optimirana za globoko učenje. Vključuje 8 grafičnih kartic Tesla V100 s tehnologijo NV-Link, vsaka grafična kartica ima 32 GB grafičnega pomnilnika, 2 centralni procesni enoti Intel Xeon E-2698 v4 20-core 2.20 GHz s skupno 40 procesorskimi jedri. Skupno zajema platforma 40.960 jeder NVIDIA CUDA® (FP32) in 20.480 jeder NVIDIA CUDA® (FP64). Poleg tega zagotovi 5120 Tensor jeder in 512 GB sistemskega pomnilnika DDR4 LRDIMM frekvence 2133 MHz. DGX-1 vključuje za shranjevanje podatkov še 1x 480 GB SSD (Intel S3610), na katerem je nameščen sistem, in 4x 1,92 TB SSD v polju RAID 0, ki zagotavlja hitro zapisovanje in branje uporabnikovih podatkov. Mrežna povezljivost je zagotovljena z dvema priključkoma 10 GbE. Infrastruktura teče na operacijskem sistemu Ubuntu Server Linux OS DGX-1 s priporočljivim grafičnim gonilnikom. Maksimalna poraba moči sistema pri polni obremenitvi je 3200 W.

## 4.2 Uporaba DGX-1

Dostop do DGX-1 imajo osebe, ki se uvrstijo na seznam uporabnikov. Na ta seznam se lahko uvrstijo zaposleni pri projektnih partnerjih (Univerza v Mariboru, Insitut !Jožef Stefan« in Univerza v Ljubljani), njihovi magistrski in doktorski študenti ter gostujoči raziskovalci, ki bodo opremo uporabljali v namene, skladne s projektom RI-SI CLARIN. Uporabniki so dolžni pri uporabi slediti navodilom uporabe in izvajati naloge skladno s tem, kar napovejo. Prav tako uporabniki na sistemu ne smejo hraniti podatkov, temveč jih morajo prenašati na lastne diskovne nosilce.

Dostop do sistema je omogočen preko upravljalca programskih bremen SLURM, ki omogoča upravljanje z viri in razvrščanje nalog v vrsto glede na parametre, ki jih določi uporabnik. Gre za sistem, pri katerem uporabnik kvantitativno določi vire, ki jih za izvajanje neke naloge potrebuje. Naloga se uvrsti v vrsto. Ko so na voljo specificirani viri, se naloga izvede.

Vsak uporabnik DGX-1 za vsako svojo nalogo napove potrebe po kapacitetah: število GPU-jev, število CPU-jeder, količino delovnega pomnilnika in čas trajanja naloge. Če viri niso na voljo, se uporabniki razvrščajo glede na to, koliko virov zahtevajo. Tisti, ki zahtevajo veliko virov, dobijo nižjo prioriteto.

### 4.3 Strežniška infrastruktura za obdelavo in hranjenje velepodatkovnih jezikovnih virov

Drugi sklop opreme na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru vključuje strežniško infrastrukturo za obdelavo velikih jezikovnih podatkov in diskovno polje za hranjenje velikih podatkov. Strežniška infrastruktura je sestavljena iz treh fizičnih strežnikov (vozlišč). Vsak izmed teh je opremljen z dvema procesorjema Intel Xeon Gold 5218 2.30 GHz (skupaj 32 jeder) in 512 GB (Dual Rank x4 DDR4-2933) sistemskega pomnilnika. Krmilnik diskov je Smart Array P408i-a/2 GB. Za sistemski disk sta uporabljena dva SSD-diska v polju RAID 1 in s skupno kapaciteto 800 GB. Za obdelavo podatkov je uporabljenih pet SSD-diskov v polju RAID 5 in skupno kapaciteto 7,6 TB ter en dodaten trdi disk kapacitete 2,4 TB. Povezljivost z omrežjem je zagotovljena s štirimi mrežnimi vmesniki. Na dveh strežnikih sta nameščena operacijska sistema Linux, medtem ko ima tretji strežnik nameščen operacijski sistem Windows Server. Infrastruktura vključuje še dve diskovni polji NAS Synology DS1819+ DiskStation za hrambo in obdelavo velikih jezikovnih podatkovnih zbirk. Skupna kapaciteta obeh diskovnih polj je 224 TB (16x NAS trdi disk 16 TB IronWolf v polju RAID 5).

S predlagano strežniško infrastrukturo se bodo izboljšale računske zmogljivosti, kar bo zelo povečalo število in obseg izvajanja raziskav na področju ugotavljanja pomena in ekstrakcije terminologije ter znanja iz besedil v slovenskem in drugih jezikih. Lotili pa se bomo lahko tudi najtršega oreha na področju detekcije plagiatov, ki je vezan na ugotavljanje direktnega prevajanja besedil iz drugih jezikov. Hranili bomo lahko precej več vhodnih jezikovnih virov, iz katerih bomo lahko pridobivali kvalitetne podatke in na njih izvajali analize ter iz njih ekstrahirali znanje.

## 5 Zaključek

Infrastrukturo CLARIN.SI smo načrtovali kot storitveno in razvojno infrastrukturo, ki je komplementarna SLING (SLING 2020) in HPC RIVR (Univerza v Mariboru 2019). CLARIN.SI bo razvil svojo obstoječo platformo stabilnih storitev in repozitorijev. Sistemi za razvoj in interaktivne obdelave pa so načrtovani kot komponente v omrežju SLING, ki bodo omogočile soopravilnost s sistemi v omrežju in bodo že v začetku lahko uporabljali kapacitete HPC RIVR za večje obdelave. Ker sta razvoj in vzdrževanje CLARIN.SI in HPC RIVR do določene mere sorodna, so potrebna sorodna znanja in veščine, zato lahko namestitev gruče GPU na Univerzi v Mariboru izkoristimo tudi za prenos znanja in tvorbo jedra znanja v okviru Univerze v Mariboru. Tako pričakujemo pozitivne sinergijske učinke sodelovanja, izmenjave znanja, skupne projekte ter visoko stopnjo soopravilnosti. CLARIN.SI bo prek HPC RIVR tudi uresničeval zahteve za dolgoročno varovanje svojih raziskovalnih podatkov na oddaljeni lokaciji, s člani SLING pa bo sodeloval prek skupnega vlaganja v infrastrukturni razvoj Slovenije prek izdelave tehničnih publikacij.

## Literatura

CLARIN ERIC (2020) CLARIN - European Research Infrastructure for Language Resources and Technology. Dosegljivo na https://www.clarin.eu/. Obiskano 1. 6. 2020.

Tomaž Erjavec, Jan Jona Javoršek, Simon Krek (2014) Raziskovalna infrastruktura CLARIN.SI. *Zbornik Devete konference Jezikovne tehnologije. Informacijska družba – IS 2014*, 9. – 10. 10. 2014, Institut "Jožef Stefan", Ljubljana.

SLING (2020) Slovensko nacionalno superračunalniško omrežje. Dosegljivo na http://www.sling.si/sling/. Obiskano 1. 6. 2020.

Univerza v Mariboru (2019) Projekt HPC RIVR. Dosegljivo na https://www.hpc-rivr.si/ Obiskano 1. 6. 2020.

# Ambivalence of LGBT Visibility in Croatian Video-Based Social Media Content

**Zoran Fijavž**

Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana

**Abstract**

We examine the influence of video content on the presence of socially inappropriate discourse in a dataset of annotated Facebook comments, representative of publicly accessible LGBT-related posts from major news sources in Croatian.
Even when controlling for post source and topicality we find posts with video content received significantly more negative LGBT-oriented than negative other-oriented comments. We interpret the results through previous discussions of LGBT visibility as both a necessity for effective advocacy via norm contestation and contact theory pathways as well as an opening for potential backlash in the form of abusive language.

## 1. Introduction

Discussions of visibility have been central to theorizing in LGBT and queer studies and created accounts of visibility that overreach the (un)desirability of visibility and foreground its ambivalence (Edenborg, 2019).[1] On one hand, visibility creates effects consistent with the contact hypothesis: intergroup interaction may, under optimal contact conditions, reduce intergroup bias (Allport, 1954). While the original formulations of the contact hypothesis focused on racial and ethnic groups, research within the paradigm has expanded to various minority groups and generally finds a inverse relalationship between prejudice and intergroup contact, even in non-optimal contact conditions (Pettigrew and Tropp, 2000). Furthermore, similar changes in prejudice have been demonstrated in imagined interactions (Crisp and Turner; 2009), parasocial (mediatized) video contact with non-fictional (Cooley and Burkholder, 2011) as well as fictional LGBT people (Schiappa, Gregg, and Hewes, 2007). The extended contact hypothesis posits and finds evidence for improved attitudes from merely knowing that in-group members have cross-group friendships (Zhou et al., 2018).

However, while public visibility of LGBT people may facilitate changes in attitudes via contact mechanisms, it may also cause various forms of backlash. In an exhaustive examination of LGBT norm diffusion in the European context, visibility was found to be central to the ability of activists to place demands on states and change attitudes, while simultaneously provoking contestation from various actors (Ayoub, 2016). In theorizing about backlash occurences, it is furthermore important to define in their particular form, as Bishin, et al. (2015) find no shifts in public opinion following the US supeme court ruling on gay marriage in 2015 as well as in experimental conditions: a "negative, large, and enduring reaction against a policy or group", as the authors define backlash, was thus not observed following legislative change. In addition, backlash in the form of banning pride parades may, provided there is external leverage, further increase the salience of the LGBT rights norm and mobilize new allies (O'Dwyer, 2012).

While research on LGBT visibility often focuses on the relationship between nation states and a transnational LGBT norm (Bosia, 2013; Lind, 2014; Edenborg, 2019), this article will adopt the lens of visibility and potential backlash to analyze engagement with LGBT content in online interactions. Drawing on discoursive theories, the discourse produced by nation states may be conceived as a single instance of speech with productive capacities among others, such as in cases of hate speech by individuals (Butler, 1996), which suggests that theories of visibility and backlash may also be applied to online user-generated content. Indeed, literature on computational propaganda demonstrates that, given the proliferation of possibilites to author public texts, discourse hyperproduction is a practicable mode of indirect state-orchestrated censorship (Wu, 2018): individual users can produce discourse that is ultimately pooled into a particular mode of state speech, ranging from innocuous-yet-strategic distraction (King et al, 2017), to targeted online harassment (Kargar and Rauchfleisch, 2019) and technological amplification (Abokhodair, Yoo, and McDonald, 2015). Furthermore, social media platforms serve a similar function as traditional media, with 55% of respondents in Croatia listing social media as a news source and 89% reporting receiving news online, which includes both social media and the websites of online news outlets (Newman et al., 2019). The functional similarity also warrants the expansion of (critical) discourse analysis approaches typically applied to traditional media (e.g. Gabrielatos and Baker, 2008) to user-generated content.

While there is an extensive body of research on various forms of abusive language, it is often motivated with the goal of (automatic) detection and or online coversation moderation (Pavlopoulos et al., 2017; Waseem and Hovy, 2017). However, the aforementioned line of research often aims at optimizing classifiers according to a particular typology or harmful speech, without investiganting the patterns in the sources used to train classifier in the first place. In particular, the notion of user-generated content may obscure the fact that there may not be a unitary, homogeneous user: according to Eurostat 2019 data, 56% of respondents aged 16 to 74 and 88% of those aged under 25 used social media EU-wide. In spite of wide-spread social media use, user-generated content (or behavioral data) can not be considered equivalent to population-level survey data. In Croatia, for example, 40% survey respondents report sharing news

[1] While some variant of LGBTI, LGBTIQ+, or LGBTIAQ+ or queer as an umbrella term could be used, we will use the acronym LGBT as used in Ljubešić et al. (2019) for the purpose of terminological consistency.

online and 25% comment on social media or websites (Newman et al., 2019).

## 2. Goal of the paper

This paper seeks to further analyze the effect of using video content as opposed to text in Facebook posts about LGBT people. We analyzed the Croatian LGBT section of the FRENK dataset (Ljubešić et al., 2019), which contains manually annotated Facebook comments. This subsection was particularly interesting because it contained both posts with text and links to news articles as well as embedded videos from three major Croatian news sites (Jutranji list, Index and 24sata).[2] The hypothesis was that posts containing video had a different distribution of comment tags with more negative LGBT-oriented posts compared to non-video posts.

## 3. Dataset decription and methodology

We analyzed the manually annotated LGBT-related Croatian FRENK dataset of socially unacceptable discourse. The following subsections explain the adjustments made to the dataset for analysis as well as post metadata we included in the analysis.

### 3.1. FRENK dataset and comment selection procedure

It is important to note that the FRENK dataset gathered Facebook posts of news sites in a relatively exhaustive manner. The Facebook posts of the most visited news sites (defined by web traffic to their main webpages) were retrieved using the public Graph API, and samples of 100 LGBT as well as non-LGBT-themed posts were used to train a SVM classifier. All retrieved posts were then classified. A sample of the classification results was checked manually, added to the posts from the first iteration and used to train the second version of the SVM classifier, giving the final selection of LGBT-related posts, which were then manually annotated for speech type and targets. While Ljubešić et al. (2019) do not analyze the content of the posts classified as LGBT-themed, we find in section 3.2. that most of the posts are related to pride parades in Croatia, which is consistent with previous studies that find these events draw large amounts for public LGBT visibility, particularly when organized for the first time (Fejes and Balogh, 2013).

In total, the dataset contained 5,787 comments from 22 Facebook posts. Out of which we removed 316 comments with multiple type-target annotations. Out of all possible type-target combinations, we only retained comments with the six most common type-target tags, comprising 97% of all comments in the dataset, which were:

1. Acceptable speech, No target (Ac);
2. Background offensive, LGBT (BO);
3. Background violence, LGBT (BV);
4. Inappropriate, No target (In);
5. Other offensive, Commenter (OC);
6. Other offensive, Other (OO).

For describing our results, we group these tags further into: neutral comments (*Acceptable speech, No target*), negative LGBT-oriented comments (*Background offensive, LGBT* or *Background violence, LGBT*) and negative other-oriented comments (*Other offensive, Commenter* or *Other offensive, Other*).

### 3.2. Metadata: video presence, post source, topicality

In addition to the comment tags, we examined the presence of embedded videos in the posts, the post source and post topicality. We marked whether posts included directly **embedded videos,** playable with Facebook's domain, as a binary category. The embedded videos varied from lengthy live stream recordings and short video interviews to a 360-degree video montage with snapshot moments. 10 out of 22 posts contained embedded videos. The posts originated from the Facebook profiles of three different **sources**: Jutranji list, Index.hr, and 24sata, all of which are major Croatian news sites. Comparing the source and video presence data reveals that in our dataset, 24sata exclusively posted video posts, Index.hr predominantly posted text posts (which nonetheless may link to video content outside of Facebook), and Jutranji lists posted the same number of both video and non-video posts. Notably, all except one were related to various pride parades in Croatia, which is why we further examined the **topicality** of the posts.

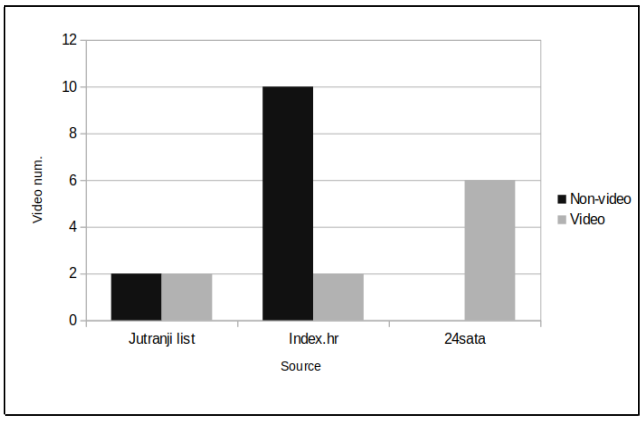



Figure 1: Video presence in posts by source

The topicality of posts was introduced to distinguish between posts focusing directly on pride parades and those not. As the video posts were nearly universally related to pride parades (with one exception), any differences identified between video and non-video posts may simply be due to the events they report and not video presence. Indeed, as seen in Figure 2, 18 of the 22 posts were published in June, when pride parades typically take

[2] Jutranji list: https://www.jutarnji.hr/
Index: https://www.index.hr/
24sata: https://www.24sata.hr/

place[3]. Manually examining the posts revealed that all posts in June were related to pride parades, while none in July or February were. While the number of video and non-video posts in June were the same, video posts in June received far more comments (3,261) compared to non-video comments (1,441), together comprising 81,3% of all comments in the dataset.

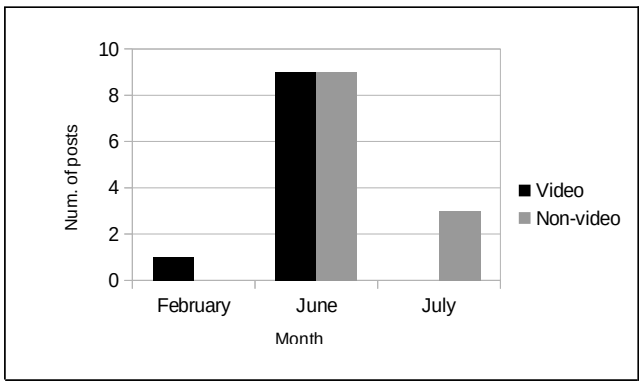



Figure 2: Video presence by month

## 4. Analysis

We tested whether the distribution of comment tags was statistically significantly different between posts with or without embedded videos, using a chi-square test of independence and calculated the standard residuals to determine which categories in particular differ between the two subsets and are over- or underrepresented. We used Cramér's V to measure the effect size of the results.[4]

The chi-square test of independence, Cramér's V and standardized residuals for each comment tag were also calculated on comment subsets to account for the effect of the difference between the sources and topicality of posts.[5]

### 4.1. General results

The results show a weak significant relationship between the video presence and comment tag distribution, $\chi^2(5, n = 5{,}357) = 175.133$, $p < 0.05$, $V = 0.180$. The distribution of all comment categories except *Inapproproate, no target* differed significantly. The differences were directional, with LGBT-oriented negative comment categories overrepresented in video posts, while neutral and other-oriented negative comments overrepresented in the non-video posts, as visualized in Figure 3 (i.e. their standardized residuals were positive or negative).

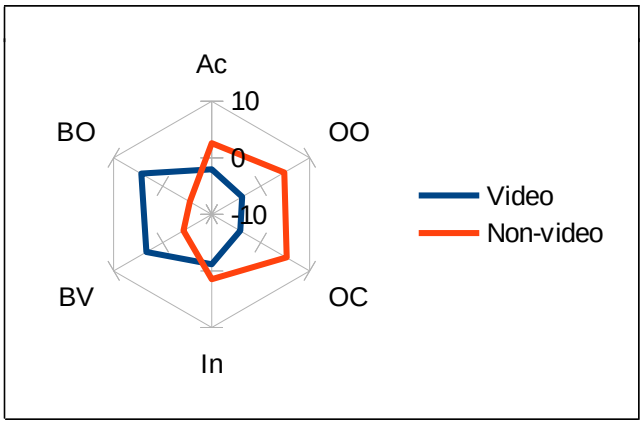



Figure 3: Standardized residuals per comment category

### 4.2. Controlling for source: Analysis of Jutranji list

As noted, not all sources used video posts to the same extent, so the results of 4.1. could influenced by the post source (e.g. more or less favorable reporting or social media audience). That is why we performed a separate analysis for Jutranji list, which had the same number of video and non-video posts. The difference in comment tag distributions were significant, $\chi^2(5, n = 1{,}705) = 58.304$, $p < 0.05$, $V = 0.184$. The directionality of the residuals was comparable to the results in 4.1. as seen in Figure 4: the video posts contained more negative LGBT-oriented comments and less other-oriented negative comments compared to non-video posts. In spite of the similar contours, the standardized residuals for neutral comments as well as comments expressing LGBT-related violence were close to zero and the differences overall were less pronounced (e.g. the overall residuals were smaller), with only 4 out of 12 residuals being over 1.645 (compared to 10/12 for the whole dataset).

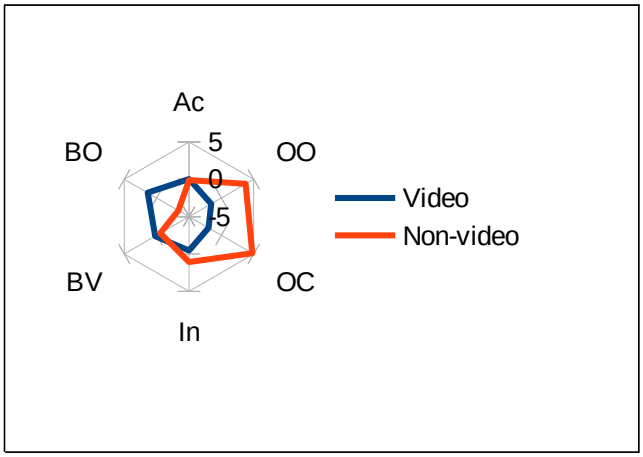



Figure 4: Standardized residuals per comment category (Jutranji list)

[3] For the purpose of our analysis, the particular year the article was published is not relevant, as the month of publishing corresponds to content related to pride parades. All posts were published in either 2011, 2016 or 2017.

[4] For df = 1 for all comparisons in this paper, the threshold value of V for small effects is 0.10, 0.3 for medium effects and 0.5 for large effects (Gravetter and Wallnau, 2016; 586).

[5] A standardized residual is a ratio: the difference between the observed count and the expected count divided by the standard deviation of the expected count in chi-square testing. Standardized residuals can be interpreted in the same manner as z-scores and were considered significant if greater than 1.645.

### 4.3. Controlling for topic orientation and moth of publishing

We also examined the influence of topicality on the distribution of comment tags. As discussed in section 3.2. nearly all video posts were related to pride parades and posted in June and none of the posts in July contained video. To disentangle the effect of video presence from topical orientation, two comparisons were made: firstly between the video and non-video posts published in June exclusively and secondly between non-video posts in July and June.

The results of separately analyzing posts in June mirror the main results. The difference between video and non-video posts was significant, $\chi^2(5, n = 4{,}379) = 66.198$, $p < 0.05$, $V = 0.122$. As seen in Figure 5, negative LGBT-oriented comments were above the expected frequencies in video posts and negative other-oriented comments were below it, while the inverse was true for non-video posts. One notable difference to the main results is that the absolute value of standardized residuals for neutral comments (*Acceptable speech, No target*), was less than 1, and the standardized residuals were smaller overall with 7/12 reaching a critical value over 1.645.

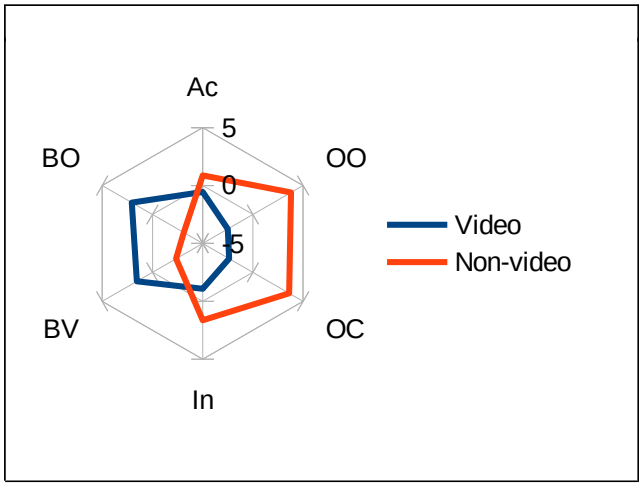



Figure 5: Standardized residuals (June only)

The comparison between non-video posts in June and July also had significant results, $\chi^2(5, n = 2{,}058) = 33.847$, $p < 0.05$, $V = 0.128$. As shown in Figure 6, there was a significant difference in offensive negative LGBT-oriented comments and the standardized residuals for LGBT-oriented comments including violence in June were slightly under the significance threshold at 1.545. Other-oriented negative comments, however, were more consistent across the months. In comparison the results for June alone, there was also a larger difference in the presence of neutral comments.

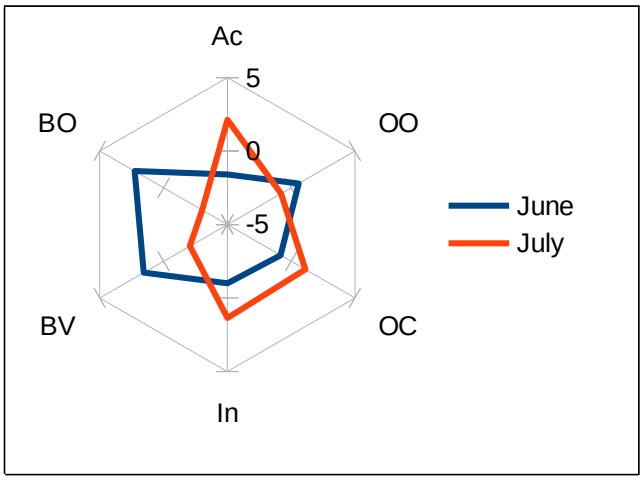



Figure 6: Standardized residuals (June and July, text)

The results from both comparisons in 4.3. provide further analysis of the main results. The results from June reflect a form of trade-off between negative other- and LGBT-related comments depending on the presence of video content, as observed in the main results. The comparison of text posts between June and July suggests that the main results can not be explained by topicality alone, as a different topicality (pride-related content and other content) lacks the difference in negative other-oriented comments seen in the main results. However, there is a significant overrepresentation of offensive negative LGBT-related comments and nearly significant overrepresentation of negative LGBT-related comments tagged as violent, which may be due to the overall increased visibility during a pride month.

## 5. Conclusion

This article hypothesized that LGBT-related posts in with video content received a different distribution of comment types compared to non-video ones in the FRENK dataset of socially unacceptable speech for Croatian. When comparing video and text posts, we found a weak positive and significant association between videos posts and negative LGBT-oriented comments and a weak and significant negative association between video posts and negative other-oriented comments. The directionality of these effects persisted when analyzing the source with the fewest posts and same number of video and non-video posts (Jutranji list), with the caveat that differences in standardized residuals did not reach critical values for all type-target tags. The main finding persisted when we only analyzed posts related to pride parades (all published in June). We also compared text posts published in June and in July (that is pride-related and non-pride-related text posts) and found increased presence of negative LGBT-related comments as well as a lower difference in other-oriented negative comments in comparison with the main results. Overall, the results show that what we considered a form of backlash in online LGBT visibility is more frequent in content related to pride parades and in particular to video content showing pride parades.

Such an interpretation woul be consistent with previous accounts of visibility (Ayoub, 2016) although on a micro-level in user-generated content: Visibility both allows for norm diffusion through multimedia content on social networks as well as potentially producing backlash in the form of socially unacceptable discourse including background-related threats. It is imporant to stress that these findings can not be read normatively as a reason to dissuade attemps at online LGBT visibility, as we merely analyzed the presence of negative LGBT-oriented comments and did not have external measures for their mid- or long-term effects: the results might well point towards a possibility to anticipate such comments and use them for countermobilization by LGBT and allied groups along the lines of O'Dwyer (2012).

Furthermore, the paper elaborates on the content of the posts gathered in the Croatian FRENK dataset: a large proportion of the posts was about pride parades in Croatia and thus published in June, consistent with qualitative studies on the role of pride parades in making LGBT communities visible in terms of media presence (Fejes and Balogh, 2013). This is an important fact for further use of the Croatian section of the FRENK dataset (and possibly others), as even the relatively exhaustive approach used for the FRENK dataset to gather LGBT-related posts mostly captured season-bound occurences (or alternatively, media articles on the topic are published nearly exclusively in relation to pride parades in June). 81,3% of all comments in the Croatian FRENK dataset originate from posts related to pride parades and the comments on pride-related videos alone comprise 56,4% of all comments in the dataset. This could be exploited in further research to compare other linguistic contexts, where the public online communication is less season-bound and more diffused and integrated into the reporting of news sites. This finding is also important for efforts regarding comment classification through machine learning approaches. Firstly, it is a cautionary tale against interpreting data gathered through automated means as a form of public opinion, as in our case the dataset clearly captures discourse of a non-random population around pride parades as a yearly peak of LGBT visibility in the context of Croatia. Secondly, the fact that the dataset is bound up with pride parades raises the question of generalizability of classifiers trained in this manner to other contexts: e.g. it is conceivable that headlines on a highly discussed legal case would produce comments that are not comparable to the ones in the FRENK dataset.

This paper may also provide some methodological insights. While the multi-variable analysis of categorical data is hardly novel, the development of the hypothesis on the effect of video content was formulated through the authors involvement in the comment annotation of the Croatian LGBT as well as other sections of the FRENK dataset, which required close reading of the comments and consideration of their pragmatic context, while theoretical frameworks of LGBT visibility, contact theory and backlash were incorporated after the statical analysis was performed. The author also had extensive background knowledge on LGBT-related topics, which eased the selection and interpretation of post metadata (i.e. introducing post topicality to further examine the main hypothesis). In this sense, the research process resembled grounded theory approaches, which favor hypothesis construction based on observations rather than fixed prior theoretical frameworks (Strauss and Corbin, 1994).

However, there are some limitations to this study. Firstly, there is a number of questions on the generalizability of the results. The results cannot be generalized outside of social media posts, as social media users differ significantly from the general population (Mellon and Prosser, 2017; Eurostat, 2019), a relatively low number of users actually write comments and there are demographic-related preferences for consuming content in video versus text form (Newman et al., 2019). Secondly, chi-square tests were used on data subsets and there was no assessment of the family-wise error rate (false discoveries while performing multiple hypothesis tests), which may be beyond the type-I error rate of each individual test at $p = 0.05$. That could be avoided by applying logit models (e.g. logistic regression), although the topicality and video presence would need to be added manually, as they are not part of the original FRENK dataset.

## 6. References

Amy Lind. 2014. 'Out' in International Relations: Why Queer Visibility Matters. *International Studies Review*, 16(4):601–604.

Anselm Strauss and Juliet Corbin. 1994. Grounded Theory Methodology: An Overview. In: Norman K. Denzin and Yvonna S. Lincoln, eds., *Handbook of Qualitative Research*, 1st edition, pages 273–285. Thousand Oaks, California, Sage Publications.

Benjamin Bishin, Thomas Hayes, Matthew Incantalupo, and Charles Anthony Smith. 2015. Opinion Backlash and Public Attitudes: Are Political Advances in Gay Rights Counterproductive? *American Journal Of Political Science*, 60(3):625–648.

Conor O'Dwyer. 2018. The Benefits of Backlash: EU Accession and the Organization of LGBT Activism in Postcommunist Poland and the Czech Republic. *East European Politics And Societies: And Cultures*, *32*(4):892–923.

Costas Gabrielatos and Paul Baker. 2008. Fleeing, Sneaking, Flooding. *Journal Of English Linguistics*, 36(1):5–38.

Edward Schiappa, Peter B. Gregg, and Dean E. Hewes. 2005. The Parasocial Contact Hypothesis. *Communication Monographs*, 72(1):92–115.

Emil Edenborg. 2019. Russia's spectacle of "traditional values": rethinking the politics of visibility. *International Feminist Journal Of Politics*, 22(1):106–126.

Eurostat, the statistical office of the European Union. 2020. ICT usage in households and by individuals [Data file]. Retrieved from: http://tiny.cc/c5z4oz (Accessed 15. 5. 2020)

Frederick Gravetter and Larry Wallnau. 2016. *Statistics for the Behavioral Sciences, Tenth Edition*. Boston: Cengage Learning.

Gary King, Jennifer Pan and Margaret Roberts. 2017. How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument. *American Political Science Review*, 111(3):484–501.

Gordon Willard Allport. 1954. *The nature of prejudice.* Boston: Addison-Wesley.

John Pavlopoulos, Prodromos Malakasiotis, Ion Androutsopoulos. 2017. Attention to Abusive User Content Moderation. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark, Association for Computational Linguistics.

Jonathan Mellon and Christopher Prosser. 2017. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3).

Jonny Cooley and Gary Burkholder. 2011. Using Video and Contact to Change Attitudes Toward Gay Men and Lesbians. *Journal of Social, Behavioral, and Health Sciences*, 5(1):83–90.

Judith Butler. 1996. *Excitable speech*. New York: Routledge.

Nárcisz Fejes and Andrea Balogh. 2013. *Queer visibility in post-socialist cultures*. Bristol: Intellect.

Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, and Rasmus Kleis Nielsen. 2019. *Reuters Institute Digital News Report 2019*. Oxford: Reuters Institute for the Study of Journalism.

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec. 2019. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. In: *Proceedings of International Conference on Text, Speech, and Dialogue, TSD 2019*, pages 103–114, Ljubljana, Slovenia.

Norah Abokhodair, Daisy Yoo, David Wright McDonald. 2015. Dissecting a Social Botnet. *Proceedings Of The 18th ACM Conference On Computer Supported Cooperative Work and Social Computing - CSCW '15*, pages 839–851, New York, United States, Association for Computing Machinery.

Phillip Ayoub. 2016. *When States Come Out: Europe's Sexual Minorities and the Politics of Visibility*. New York: Cambridge University Press.

Richard Crisp and Rhiannon Turner. 2009. Can imagined interactions produce positive perceptions? Reducing prejudice through simulated social contact. *American Psychologist*, 64(4), 231–240.

Shelly Zhou, Elizabeth Page-Gould, Arthur Aron, Anne Moyer, Miles Hewstone. 2018. The Extended Contact Hypothesis: A Meta-Analysis on 20 Years of Research. *Personality And Social Psychology Review*, 23(2):132–160.

Simin Kargar and Adrian Rauchfleisch. 2019. State-aligned trolling in Iran and the double-edged affordances of Instagram. *New Media and Society*, 21(7):1506–1527.

Thomas F. Pettigrew and Linda R. Tropp. 2000. Does intergroup contact reduce prejudice? Recent meta-analytic findings. In: Stuart Oskamp, ed., *Reducing prejudice and discrimination: Social psychological perspectives*, pages 93–114. Mahwah, New Yersey, Lawrence Erlbaum Associates.

Tim Wu. 2018. Is the First Amendment Obsolete? *Michigan Law Review*. 117(3).

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In: *Proceedings of the Student Research Workshop,* pages 88–93, San Diego, California, Association for Computational Linguistics.

# Zaznavanje sentimenta v novicah z globokimi nevronskimi mrežami

**Andraž Pelicon**

Odsek za tehnologije znanja
Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana
Andraz.Pelicon@ijs.si

**Povzetek**

Področje analize sentimenta v novicah se v zadnjem času vse pogosteje uporablja predvsem v okviru napovedovanja gibanja finančnih trgov, vendar je za slovenski jezik še dokaj slabo raziskano. V okviru te raziskave smo zasnovali arhitekturo na osnovi nevronskih mrež, ki za klasifikacijo uporablja kombinacijo samodejno generiranih značilk s pomočjo slojev s povratno zanko in TF-IDF obtežitev. Modeli, ki uporabljajo omenjeno arhitekturo, dosegajo primerljive rezultate z že obstoječimi modeli in so sposobni učinkovitega učenja na korpusih v velikosti okrog 10.000 dokumentov.

## 1. Uvod

Mnenja in subjektivni odnos ali sentiment do določenega subjekta so ljudem zelo pomembni pri vsakodnevnem odločanju in sodobne tehnologije omogočajo hitro in učinkovito izmenjavo mnenj kjerkoli po svetu. Na družbenih platformah kot sta Twitter in Facebook ljudje neprestano izražajo mnenja o različnih izdelkih ter svoj odnos do raznih tem in oseb, od politikov do podjetnikov. Spletne trgovine omogočajo, da kupci o določenem izdelku oddajo svoje mnenje in oceno, kar pomaga tako drugim potencialnim kupcem kot trgovini sami, da lahko lažje oceni priljubljenost prodajanih izdelkov. Novičarske platforme omogočajo objavo komentarjev, kjer bralci izražajo svoja mnenja glede prebranih novic in subjektivne poglede na obravnavane teme. Prek vseh teh kanalov se tako proizvedejo ogromne količine prosto dostopnih besedil, ki jih je praktično nemogoče ročno pregledati. Zato se je že precej zgodaj pojavilo zanimanje za izdelavo modelov in sistemov, ki bi s pomočjo tehnik strojnega učenja bili sposobni samodejno analizirati mnenja v krajših in daljših oblikah besedil.

Najbolj zgodnji modeli, ki so se ukvarjali z analizo mnenj, so se osredotočali predvsem na samodejno prepoznavanje mnenj o posameznih izdelkih v ocenah in komentarjih uporabnikov, kjer je mnenje uporabnika dokaj jasno izraženo v besedilu in kjer mnenje poleg besedila izraža tudi številčna ali znakovna ocena. Z eksplozijo družbenih omrežij je za posamezna podjetja postalo zanimivo sledenje priljubljenosti svoje znamke v javnih objavah uporabnikov na spletu. Kmalu pa se je zanimanje začelo usmerjati tudi na vrste besedil, ki naj bi načeloma nevtralno podajale informacije in ne bi izražale nobenega mnenja avtorja ali njegovega subjektivnega pogleda na obravnavano tematiko, kot so na primer novice. Analiza sentimenta v novicah se pogosto uporablja na novicah iz finančne domene, kjer številni raziskovalci poskušajo na podlagi analiziranega sentimenta napovedati gibanja na finančnih trgih. Van de Kauter et al. (2015) navajajo, da imajo novice neposreden vpliv na finančne trge, saj dobre novice načeloma spodbujajo, slabe novice pa omejujejo rast na trgu.

Večinoma se analiza mnenj in sentimenta zastavi kot klasifikacijski problem, kjer se posamezna besedila klasificira v eno od predhodno definiranih kategorij. Najbolj razširjene tehnike strojnega učenja na tem področju so metode podpornih vektorjev, vendar se v zadnjem času tudi na tem področju pojavlja vse več modelov globokih nevronskih mrež, ki dosegajo primerljive rezultate.

V zvezi s količino podatkov, ki je potrebna za učenje sposobnega modela, ni enotnega mnenja, saj je slednje precej odvisno od samega problema in velikosti modela. Kljub temu empirični rezultati iz posameznih študij kažejo, da se lahko na podlagi večje količine podatkov modeli strojnega učenja, predvsem modeli globokega učenja, naučijo bolj informativnih značilk, ki pomagajo pri boljši klasifikaciji. Yang et al. (2016) so razvili arhitekturo globokih nevronskih mrež za klasifikacijo daljših dokumentov. Poročali so o izboljšanju rezultatov na vseh šestih preskusnih zbirkah podatkov, pri čemer sta dve manjši zbirki podatkov vsebovali več kot 300.000 primerov, tri večje podatkovne zbirke so vsebovale več kot milijon primerov, največja pa več kot tri milijone primerov. Devlin et al. (2018) so na podlagi variante nevronskih mrež, tako imenovane pozornosti (ang. attention), razvili velik jezikovni model, imenovan BERT. V članku poročajo, da je bil model naučen na dveh sicer neoznačenih korpusih, BookCorpus, ki vsebuje 11.038 knjig v celoti in 800 milijonov besed, in korpusu člankov angleške Wikipedije, ki vsebuje 2.500 milijonov besed. Pridobivanje tako velikih količin podatkov, predvsem označenih, je pogosto zelo drag in zamuden postopek, zato je učenje uspešnih klasifikacijskih modelov na omejenih virih aktualen problem.

Raziskovalni prispevki našega dela so naslednji. Razvili smo nov sistem za zaznavanje senitmenta v novicah, ki temelji na arhitekturi globokih nevronskih mrež in pokazali, da dosega primerljive rezultate z dosedanjimi modeli, ki temeljijo na tradicionalnih metodah strojenga učenja. Nadalje smo pokazali, da lahko naš nevronski model učinkovito učimo tudi na manjših podatkovnih množicah reda do 10.000 primerov.

## 2. Namen članka

Čeprav se analiza sentimenta v novicah vse pogosteje izvaja za večje jezike, je to področje za slovenščino zaenkrat še precej neraziskano. Edini modeli, ki po naših informacijah obstajajo, so modeli, naučeni na podlagi metode podpornih vektorjev. Cilj te raziskave je, da bi razvili model globokih nevronskih mrež, ki bi uspešno napovedoval sentiment v novicah, kot ga dojema povprečen slovenski bralec. Pri tem si kot prvo hipotezo postavljamo trditev, da je mogoče z modeli nevronskih mrež doseči primerljivo klasifikacijsko točnost kot z drugimi tehnikami strojnega učenja.

Za naš problem trenutno še ni na voljo velike količine kakovostno označenih podatkov. Tako bomo poskušali naše modele naučiti na korpusu, ki vsebuje okrog 10.000 primerov novic. Naša druga hipoteza je, da je tudi s tako omejeno količino podatkov mogoče izdelati model globokih nevronskih mrež, ki uspešno modelira dani problem.

## 3. Pregled raziskav

Analiza sentimenta, v literaturi imenovana tudi analiza mnenj, je večdisciplinarno področje na preseku računalništva in jezikoslovja, ki spada v širše področje obdelave naravnega jezika (NLP). Ukvarja se z analizo in prepoznavanjem odnosa, čustev in mnenj ljudi do določenega objekta. Zajema skupek tehnik iz računalniške obdelave naravnega jezika, s katerimi lahko iz besedila izvlečemo subjektivne informacije (Beigi et al., 2016).

Tradicionalno se sentiment v besedilih razpoznava v odnosu do objekta sentimenta, tj. predmeta, pojma ali osebe, na katero se izraženi sentiment nanaša (Mejova, 2009). Primeri objektov sentimenta v tvitih so izdelki določenega podjetja, v ocenah filmov pa filmi.

Čeprav je tako razmerje med sentimentom v jezikovni strukturi in objektom sentimenta precej razširjeno, pa se nekateri avtorji problema lotijo z druge perspektive. Namesto vprašanja, kakšno mnenje ima avtor do nekega objekta v besedilu, se sprašujejo, kakšen sentiment zaznavajo bralci v določenem besedilu. Primer takega razumevanja vloge sentimenta najdemo v Lin et al. (2008). Naša raziskava se osredotoča na slednjo perspektivo sentimenta, in sicer na nivoju dokumenta.

Na področju analize sentimenta v novicah avtorji pogosto kombinirajo pristope z leksikoni sentimenta in algoritmi strojnega učenja. Taj et al. (2019) so sestavili svoj korpus iz novic, objavljenih na portalu BBC, ter s pomočjo TF-IDF obtežitev v vsaki novici poiskali ključne besede. Nato so izključno na podlagi podatkov o sentimentu ključnih besed iz leksikona sentimenta poskušali analizirati sentiment v posameznih novicah. V okviru tehnik strojnega učenja se za učenje modelov najpogosteje uporablja metoda podpornih vektorjev. Kaur in Kaur (2015) sta s pomočjo metode podpornih vektorjev analizirala sentiment v novicah, napisanih v jeziku pandžabi, pri čemer sta besedila obdelala s standardnimi metodami predobdelave, in sicer tokenizacijo, odstranjevanjem neinformativnih besed in krnjenjem. Pogosto se v kombinaciji s strojnim učenjem uporabljajo tudi leksikoni sentimenta za podajanje dodatnih informacij o sentimentu v model. Lin et al. (2008) so analizirali sentiment v kitajskih novicah, kjer so uporabili metodo podpornih vektorjev, pri čemer so kot vhod uporabili več vrst značilk, ki so jih pridobili s kombinacijo predobdelave besedil v korpusu in informacij iz leksikona sentimenta. Kot značilke so uporabili bigrame pismenk, posamezne besede, metapodatke o posamezni novici, pripone besed in sentiment posameznih besed, ki so ga pridobili v leksikonu sentimenta. Li et al. (2014) so podobno metodo uporabili za preverjanje vpliva novic na donosnost delnic. Naloga modelov je bila predvideti gibanje cen delnic, pri čemer je bila učna množica sestavljena iz novic, kot napovedni cilj pa so uporabili historične cene delnic iz istega petletnega obdobja, v katerem so izšle novice iz učnega korpusa. Da bi v model vnesli podatke o sentimentu, so s pomočjo leksikona sentimenta sestavili matriko sentimenta za celotno besedišče. Nato so matriko sentimenta uporabili za preslikavo TF-IDF matrike, ki so jo sestavili na podlagi pogostosti izrazov v korpusu novic, v nov vektorski prostor. Dobljeno matriko so nato uporabili za učenje modela s pomočjo metode podpornih vektorjev. Kumar et al. (2017) so novicam poskušali določiti sentiment kot zvezno vrednost v intervalu med -1 in 1, pri čemer je vrednost -1 predstavljala najbolj negativen sentiment, vrednost 1 pa najbolj pozitiven sentiment. Tudi oni so za učenje uporabili metodo podpornih vektorjev s to razliko, da so problem opredelili kot regresijski problem. Za napoved sentimenta so kot značilke uporabili uni- in bigrame besed, obtežene s TF-IDF utežmi, značilke, izdelane na podlagi informacij iz leksikona sentimenta, in 300-dimenzijske vektorske vložitve besed word2vec.

V zadnjem času se tudi na področju analize sentimenta vedno bolj pojavljajo modeli globokega učenja, ki temeljijo na nevronskih mrežah. Mansar et al. (2017) so za analizo sentimenta uporabili konvolucijske nevronske mreže, ki se sicer pogosteje uporabljajo v robotskem vidu. S pomočjo konvolucijskih plasti so pridobili notranje predstavitve posamezne novice iz učnega korpusa in jih nato združili z oceno sentimenta posamezne novice, ki so jo pridobili s preprostim modelom VADER, ki temelji na pravilih. Tako pridobljene značilke so nato uporabili kot vhod v končni klasifikator, ki je temeljil na polnopovezani nevronski mreži. Razen tokenizacije se metod predobdelave besedil niso posluževali. Njihov model je bil leta 2017 najboljši v eni od nalog na letnem mednarodnem tekmovanju o računalniški semantični analizi SemEval. Moore in Rayson (2017) sta za namene analize sentimenta iz naslovov novic izdelala dva modela, enega na podlagi metode podpornih vektorjev z ročno generiranimi značilkami, drugega pa na podlagi dvosmernih nevronskih mrež s povratno zanko z dolgo-kratkoročno pomnilno celico (ang. bidirectional recurrent neural networks with long-short term memory ali BILSTM). V svojem eksperimentu poročata, da je njihov nevronski model dosegel za 4–6 odstotkov višjo točnost v primerjavi z metodo podpornih vektorjev.

Kljub temu, da številne od navedenih raziskav uporabljajo predhodno naučene besedne vektorske vložitve za predstavitev vhodnih besedil, Yu et al. (2017) navajajo, da slednje načeloma ne zajamejo dovolj podatkov o sentimentu, saj nosijo samo informacijo o kontekstu posamezne besede. Tako imamo lahko besede, ki se uporabljajo v zelo podobnih kontekstih, izražajo pa povsem nasproten

sentiment. Vzemimo na primer protipomenki *dober* in *slab*. Besedi izražata povsem nasproten sentiment, vendar ker sta obe pridevnika in se pojavljata v zelo podobnih kontekstih, imata zelo podobno vektorsko vložitev. Vnašanja informacije o sentimentu v vektorske vložitve so se lotili tako, da so vzeli javno dostopne predhodno naučene vektorske vložitve in jih na podlagi informacij iz leksikona sentimenta prilagodili. Pri tem so uporabili leksikon E-ANEW, v katerem je sentiment kodiran z realnim številom na intervalu od 1 do 9, kjer 1 predstavlja najbolj negativen, 5 najbolj nevtralen in 9 najbolj pozitiven sentiment. Za vsako besedo v korpusu so poiskali njenih deset najbližjih sosedov glede na kosinusno razdaljo njihovih vektorskih vložitev. Nato so besede v tej množici razvrstili po sentimentu, in sicer od bolj do manj podobne glede na absolutno razliko v sentimentu iz leksikona sentimenta med besedo iz množice in ciljno besedo. Ko je bil seznam besed sestavljen, so vektorsko vložitev ciljne besede posodobili tako, da je bila bliže vektorskim vložitvam na vrhu seznama, ki so ji bile po sentimentu podobne, in dlje od vektorskih vložitev besedam na dnu seznama, ki so se po sentimentu od ciljne besede razlikovale. Da bi preprečili preveliko spremembo izvirnega vektorskega prostora in posledično izgubo kontekstualnih informacij, so dodali omejitev, ki je določala, koliko se lahko posodobljeni vektor ciljne besede največ razlikuje od izvirnega vektorja. Metodo posodabljanja vektorjev so preizkusili na dveh predhodno naučenih javno dostopnih vektorskih vložitvah Word2vec in GloVe. Posodobljene in izvirne vektorske vložitve so nato uporabili na problemih binarne in petrazredne klasifikacije sentimenta, pri čemer so modele učili na podatkih iz korpusa SST (Socher et al., 2013). V primerjavi rezultatov avtorji poročajo o 1,5-odstotni izboljšavi za vektorje GloVe in 1,7-odstotni izboljšavi za vektorje Word2vec na problemu binarne klasifikacije ter o 1,5-odstotni izboljšavi za oboje vektorje na problemu petrazredne klasifikacije.

Področje analize sentimenta v novicah v slovenskem jeziku je zaenkrat še precej neraziskano. Bučar et al. (2018) so na korpusu novic, ki so ga sami sestavili in tudi javno objavili, naučili več modelov na podlagi metod naivnega Bayesa, podpornih vektorjev in k najbližjih sosedov, pri čemer so kot značilke uporabljali n-grame besed s TF-IDF obtežitvami. Najbolj učinkovite modele, ki so bili naučeni za razlikovanje sentimenta glede na tri kategorije, negativno, nevtralno in pozitivno, smo v tej raziskavi uporabili kot osnovo za primerjavo naših naučenih modelov.

## 4. Metodologija dela

V tem poglavju predstavimo metodološki pristop, ki smo ga uporabili za izvedbo eksperimenta. Najprej opišemo podatkovno množico, ki smo jo uporabili za učenje modelov in njihovo evalvacijo. Nato opišemo arhitekturo nevronskih mrež, ki smo jo uporabili za naše modele, ter postopek učenja vseh petih modelov analize sentimenta v novicah. Poglavje zaključimo z opisom postopka validacije naučenih modelov.

### 4.1. Opis podatkov

Za namene razvoja modelov za zaznavanje sentimenta v slovenskih novicah smo uporabili korpus slovenskih novic SentiNews (Bučar et al., 2018), ki je sestavljen iz 10.427 novic z ekonomsko, finančno in politično vsebino. Novice so bile zbrane na petih najbolj branih slovenskih spletnih novinarskih portalih, in sicer 24ur, Dnevnik, Finance, Rtvslo in Žurnal24, v obdobju med 1. septembrom 2007 in 31. decembrom 2013. Novice je šest označevalcev ročno označilo na treh stopnjah granularnosti, in sicer na nivoju besed, povedi in celotnega dokumenta. Med označevanjem so morali označevalci sentiment v novicah označiti s stališča povprečnega slovenskega bralca, pri čemer so odgovorili na vprašanje, kakšne občutke so doživljali po branju novice. Za učenje modelov smo uporabili končne oznake, ki posamezno novico v korpusu označijo kot negativno, nevtralno ali pozitivno. Označeni del korpusa tako vsebuje 3337 negativnih, 5425 nevtralnih in 1665 novic, kar kaže na dokaj drastično neuravnoteženost kategorij. Ker so mere strinjanja med označevalci najvišje pri oznakah na ravni dokumenta, smo se odločili, da bomo za namene nadzorovanega učenja modela uporabili oznake na tej ravni.

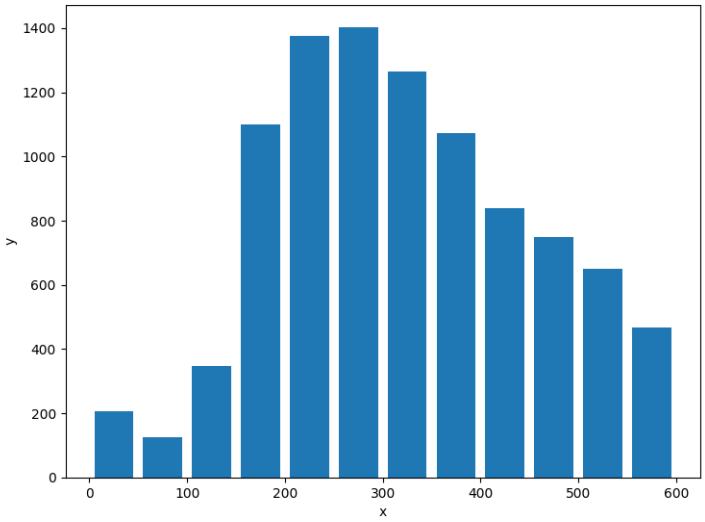
Po podrobnejšem pregledu podatkov smo ugotovili, da so določeni primeri v korpusu izjemno kratki. Načeloma pri teh primerih večinoma ne gre za celotno novico, temveč samo za podpise pod slikami, ki so sestavljeni iz ene do treh povedi. Takih priemrov je sicer malo, vendar se po sami zgradbi in informativnosti močno razlikujejo od ostalih primerov. Da v modele ne bi vnašali nepotrebnega šuma, smo se odločili, da bomo zato iz korpusa odstranili vse primere, ki so sestavljeni iz manj kot petih povedi. Takih primerov je bilo v korpusu 307, tako da je naš končni korpus vseboval 10.120 označenih primerov.

### 4.2. Zasnova eksperimenta

Za modeliranje sentimenta novic smo naučili klasifikacijski model na osnovi nevronskih mrež. Model kot vhod tako prejme besedilo kot celoto in ga v nadaljevanju obdela v dveh obdelovalnih cevovodih. V prvem obdelovalnem cevovodu smo vsako novico najprej tokenizirali. Nato smo vsako besedo, ki se pojavi v korpusu, preslikali v ustrezno vektorsko vložitev. V našem primeru smo uporabili predhodno naučene 300-dimenzionalne vektorske vložitve Fasttext (Grave et al., 2018). Tako smo dobili matriko velikosti $R^{dxm}$, kjer je $d$ število besed v novici, $m$ pa dimenzija vektorske vložitve.

Novice v korpusu so različne dolžine, za zagotavljanje računsko učinkovitih matričnih operacij med učenjem modela pa moramo nevronski mreži s povratno zanko posredovati vhodne podatke enake dolžine. Zato smo posamezne novice po potrebi skrajšali ali podaljšali. Posamezno novico smo podaljšali tako, da smo konec matrike zapolnili z ničelnimi vektorji, ki ne predstavljajo besed. Dolžino novic smo glede na porazdelitev dolžin novic v korpusu, predstavljeni na Sliki 1, določili na 500 besed.

Predobdelane novice smo tako poslali v sloj s povratno zanko z LSTM celicami s 120-dimenzionalnimi skritimi plastmi. V vsakem koraku obdelave sloj s povratno zanko prejme kot vhod vektorsko vložitev posameznega elementa vhodne sekvence in jo preslika v 120-dimenzionalno notranjo vektorsko predstavitev. Za namene regularizacije smo na sloju LSTM uporabili izpust neposredno na nevronih kot



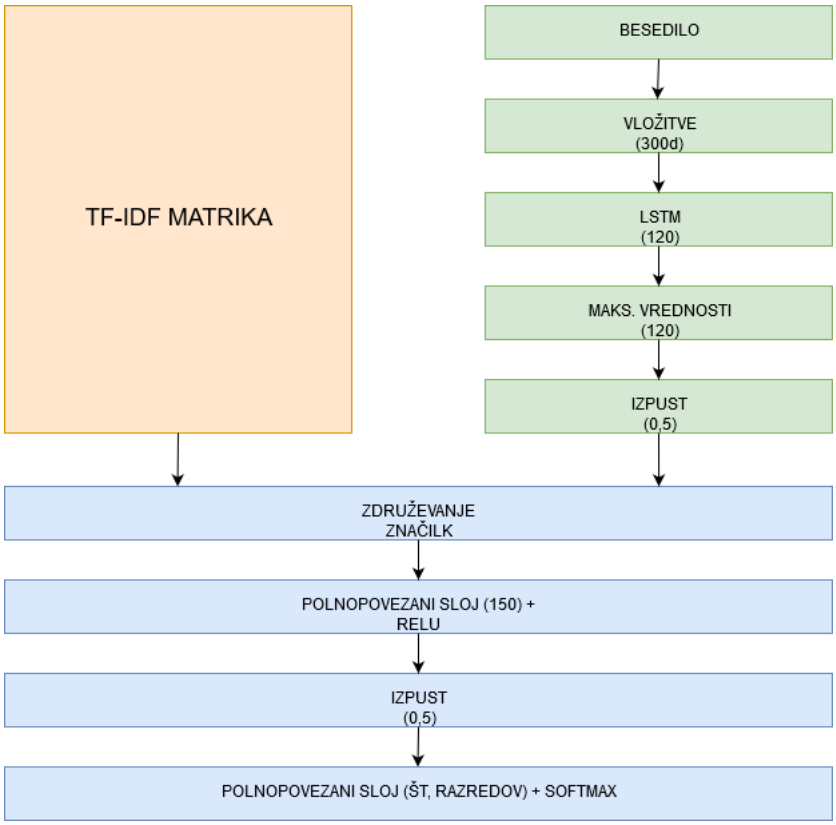

Slika 1: Porazdelitev dolžin besedil v korpusu glede na število besed

tudi na povratnih zankah. Parameter $p$, ki določa verjetnost izpusta posameznega nevrona ali zanke, smo za obe vrsti izpusta nastavili na 0,4. Na izhodu iz sloja s povratno zanko smo nato uporabili operacijo globalne izbire maksimalnih vrednosti (ang. global max pooling) prek vseh notranjih vektorskih predstavitev, ki jih mreža generira pri obdelavi sekvence. Na ta način smo želeli, da model med samodejno generiranimi notranjimi predstavitvami besedila uporabi samo tiste, ki najbolje predstavljajo vhodno besedilo. Na ta način smo dodatno zmanjšali dimenzionalnost vhoda, tako da smo izbrali zgolj najbolj diskriminativne značilke.

Tudi v drugem obdelovalnem cevovodu smo vsako vhodno besedilo iz učne množice najprej tokenizirali. Nato smo izdelali dve matriki s TF-IDF obtežitvami. V prvi matriki smo obtežili vse besedne N-grame dolžine od 1 do 5, pri čemer smo dodali omejitev, da v končno matriko vključimo zgolj tiste N-grame, ki se v učni množici pojavijo vsaj petkrat. Na tak način smo nekoliko zmanjšali dimenzionalnost matrike, pri čemer smo predpostavljali, da tak poseg odstrani vse izjemno redke besede in besedne zveze, ki ne dodajo veliko teže pri klasifikaciji, ter besede in besedne zveze, ki so morda zatipkane. Vse tako pridobljene N-grame smo obtežili s sublinearno obtežitvijo. V drugi matriki smo s TF-IDF utežmi obtežili vse N-grame, sestavljene iz znakov dolžine od 1 do 7. Tudi v tem primeru smo iz končne matrike odstranili vse N-grame, ki se pojavljajo manj kot 5-krat, preostale N-grame pa obtežili s sublinearno obtežitvijo. Nato smo obe matriki združili v eno TF-IDF matriko dimenzije $R^{nxm}$, kjer $n$ predstavlja število besedil v učni množici, $m$ pa število značilk. Matriki, ki smo jih pridobili iz obeh obdelovalnih cevovodov, smo nato združili in posredovali polnopovezani nevronski mreži, ki skrbi za končno klasifikacijo. Mreža je sestavljena iz ene skrite plasti in ene izhodne plasti. Skrita plast vsebuje 150 nevronov, kot aktivacijska funkcija pa je uporabljena funkcija Relu.

Kot regularizacijsko metodo smo tudi na tem sloju uporabili izpust z vrednostjo parametra $p$ 0,5. Izhodno plast sestavljajo trije nevroni, torej po en nevron za vsako ciljno oznako. Na izhodnem sloju uporabimo aktivacijsko funkcijo softmax, ki vrne verjetnostno porazdelitev vhodnega parametra glede na ciljne oznake. Kot napovedani razred nato vzamemo tisto ciljno oznako z največjo verjetnostjo. Za bolj celovit pregled je opisana arhitektura grafično predstavljena na Sliki 2.

Opisano arhitekturo si deli vseh pet modelov, ki smo jih naučili v sklopu te raziskave. Modeli se med seboj razlikujejo v predobdelavi vhodnih podatkov in uporabljenih vektorskih vložitvah besed. Pri modelu LSTM_BOW smo vhodna besedila minimalno obdelali, in sicer smo zgolj vse velike črke v celotni novici spremenili v male. Pri modelu LSTM_BOW+TPROC so bila vhodna besedila nekoliko temeljiteje predobdelana, in sicer smo iz besedil odstranili vse številke, neinformativne besede in ločila. Pri modelih LSTM_BOW+REF in LSTM_BOW+TPROC+REF smo vektorske vložitve na podlagi metode, opisane v poglavju 3., obogatili z informacijami o sentimentu besed. Sentimente slovenskih besed smo pridobili v leksikonu sentimenta JOB (Bučar et al., 2018). Leksikon JOB je bil sestavljen na istem korpusu kot smo ga uporabili za učenje našega modela. Vendar so informacije iz leksikona bile pridobljene s pomočjo oznak na nivoju posameznih povedi besedil v korpusu, medtem ko smo za učenje naših modelov uporabili oznake na nivoju besedil. Modela se razlikujeta po predobdelavi besedil, in sicer je bila za model LSTM_BOW+TPROC+REF uporabljena temeljitejša predobdelava, ki je bila uporabljena tudi za model LSTM_BOW+TPROC. Pri zadnjem modelu LSTM_BOW+REF2 so bila besedila enako minimalno obdelana kot pri modelu LSTM_BOW, uporabljene pa so bile s sentimentom obogatene vektorske vložitve. Za razliko od ostalih modelov, kjer se vektorske vložitve med učenjem niso posodabljale, smo jih v tem primeru med sa-




Slika 2: Arhitektura modelov nevronskih mrež.

mim učenjem modela dodatno prilagajali. S tem smo upali, da bodo na podlagi podatkov iz učnega korpusa zajele dodatne informacije, ki so specifične za podani problem, kar bi pozitivno vplivalo na zmogljivost modela.

Izvirno podatkovno zbirko smo razdelili na učno in testno množico v razmerju 80:20. Vse modele smo naučili s pomočjo stohastičnega gradientnega spusta z velikostjo paketov 32. Uteži modelov smo med učenjem prilagajali z algoritmom ADAM (Kingma in Ba, 2014), pri čemer je bila začetna učna stopnja nastavljena na 0,001. Mdele smo učili 10 epoh, pri čemer smo v vsaki epohi model učili na celotni učni množici. Uspešnost naučenih modelov smo na testni množici ocenili s pomočjo dveh standardnih mer, in sicer klasifikacijsko točnostjo in povprečno oceno F1.

Tabela 1: Primerjava rezultatov modelov, naučenih v sklopu te raziskave, in osnovnih modelov.

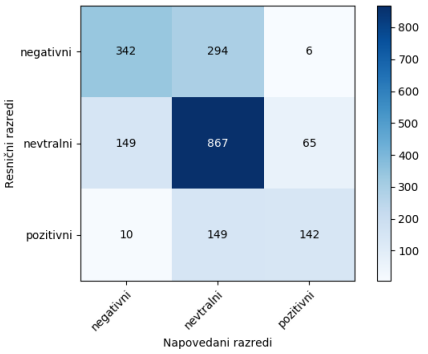
| Model | Točnost | F1 |
|---|---|---|
| SVM (poročano v (Bučar et al., 2018)) | 66.5 | 63.4 |
| NBM (poročano v (Bučar et al., 2018)) | 64.3 | 65.9 |
| SVM (ponovitev poizkusa) | 62.7 | 59.0 |
| NBM (ponovitev poizkusa) | 53.8 | 24.57 |
| LSTM_BOW | 66.5 | 61.6 |
| LSTM_BOW+TPROC | 66 | 61.1 |
| LSTM_BOW+TPROC+REF | 65.9 | 61 |
| LSTM_BOW+REF | 66.4 | 62.1 |
| LSTM_BOW+REF2 | 66.7 | 62.5 |

## 5. Rezultati

V tem poglavju predstavljamo rezultate vseh petih naučenih modelov. Kot osnovo za primerjanje zmogljivosti modelov smo uporabili modela iz Bučar et al. (2018). Tudi ta modela sta naučena za prepoznavanje sentimenta v novicah na isti podatkovni zbirki, pri čemer model na podlagi metode podpornih vektorjev (SVM) dosega višjo klasifikacijsko točnost, model, naučen na podlagi metode naivnega Bayesa (NBM), pa dosega višjo oceno F1.

Za testiranje modelov so avtorji v izvirnem članku izvedli 5 ponovitev 10-kratnega križnega preverjanja. Zaradi znatno višje računske zahtevnosti naših modelov, enakega načina testiranja nismo mogli ponoviti, zato predpostavljamo, da neposredna primerjava med dvema študijama ni povsem primerljiva. Zato smo modela na podlagi metod podpornih vektorjev in naivnega Bayesa, kot sta opisana v izvirnem članku, znova naučili na naši razdelitvi podatkovne zbirke. Na naši razdelitvi sta oba modela dosegla nižje rezultate. Rezultate modelov nevronskih mrež in modelov, ki smo jih vzeli za osnovo, predstavljamo v Tabeli 1.

V primerjavi z osnovnimi modeli, ki so bili naučeni na naši razdelitvi podatkovne zbirke, dosegajo vsi modeli nevronskih mrež opazno boljše rezultate tako po klasifikacijski točnosti kot po oceni F1. V primerjavi z modeli iz izvirnega članka je zmogljivost naučenih modelov nevronskih mrež primerljiva s poročanimi rezultati. Modela, ki najslabše klasificirata, LSTM_BOW+TPROC in LSTM_BOW+TPROC+REF, se glede na točnost povsem približata SVM-ju, medtem ko po oceni F1 zaostajata za približno 2 %. Dober rezultat doseže najenostavnejši model




Slika 3: Matrika zamenjav najuspešnejšega nevronskega modela LSTM_BOW+REF2.

LSTM_BOW, ki je po natančnosti izenačen s SVM-jem, izboljša pa tudi oceno F1 v primerjavi z našima najšibkejšima modeloma. Najboljše rezultate dosegata nevronska modela s popravljenimi vektorskimi vložitvami LSTM_BOW+REF in LSTM_BOW+REF2. Glede na klasifikacijsko točnost sta oba modela povsem izenačena s SVM-jem, pri čemer ga drugi model celo preseže za 0,2%.

Zanimiva je tudi primerjava modelov nevronskih mrež med sabo. Kljub temu, da se metode predobdelave besedila smatrajo kot zelo učinkovite pri učenju klasifikacijskih modelov, se je v naši raziskavi izkazalo ravno nasprotno. Ne samo, da pri modelih LSTM_BOW+TPROC in LSTM_BOW+TPROC+REF, ki uporabljata bolj izrazito predobdelavo vhodnih besedil, ni opaziti izboljšav, zdi se tudi, da sama predobdelava besedila rezultate celo nekoliko poslabša. Vnašanje informacij o sentimentu v vektorske vložitve je imelo nekoliko boljši vpliv. Oba modela, ki sta uporabljala popravljene vektorske vložitve, LSTM_BOW+REF in LSTM_BOW+REF2, sta bila po klasifikacijski točnosti primerljiva z osnovnim SVM-jem, v primerjavi z ostalimi modeli nevronskih mrež pa sta izboljšala oceno F1 za 0,5-1 %.

S pregledom matrik zamenjav posameznih naučenih modelov (glej Sliko 3) pridobimo boljši vpogled v napake, ki jih delajo modeli pri napovedovanju sentimenta. V članku smo vključili zgolj matriko zamenjav najuspešnejšega nevronskega modela LSTM_BOW+REF2, vendar matrike zamenjav vseh nevronskih modelov, ne glede na razlike v predobdelavi podatkov, kažejo dokaj konsistentno sliko. Modeli najmanj napak delajo pri klasifikaciji nevtralnih primerov, nekoliko slabše se odrežejo pri klasifikaciji negativnih primerov, najslabše pa klasificirajo pozitivne primere. Ta razporeditev je konsistentna z neuravnoteženostjo primerov v uporabljenem korpusu podatkov, ki v primerjavi z nevtralnimi primeri vsebuje skoraj dvakrat manj negativnih primerov in kar trikrat manj pozitivnih primerov. Nadalje modeli ne delajo veliko napak med dvema skrajnima sentimentoma, negativnim in pozitivnim. Pri vseh modelih je skupno število takih napak na testni množici namreč manjše od 20. Večino napak tako modeli naredijo pri klasifikaciji negativnih in pozitivnih primerov v nevtralne. To je še posebej očitno pri pozitivnem razredu, ki je tudi edini razred, pri katerem vsi modeli več primerov klasificirajo napačno kot pravilno. Iz teh rezultatov lahko torej sklepamo, da je razpoznavanje skrajnih sentimentov (pozitivnega in negativnega) v novicah dokaj preprosta naloga za nevronske modele, precej težje pa se odločajo pri vmesnem nevtralnem razredu. Glede na to, da so informacije v novicah podane veliko objektivneje kot v nekaterih drugih vrstah besedil, na primer ocenah izdelkov, lahko sklepamo, da vsebujejo tudi manj besed, ki izražajo močan negativen ali pozitiven sentiment. Vsled pomanjkanja takih očitnih nosilcev sentimenta, modeli tako pogosto klasificirajo novice v vmesni nevtralni razred, čeprav so morda te novice pri bralcih izzvale čustveno reakcijo.

Upoštevati moramo tudi, da je vsak učni (in testni) primer označen s konsenzom šestih označevalcev. Poleg strinjanja označevalcev na celotni podatkovni zbirki, bi bilo zato vredno raziskati nestrinjanje tudi na nivoju posameznih razredov. Če bi se izkazalo, da je nestrinjanje med označevalci precej veliko za novice, ki spadajo v nevtralni razred, bi lahko slabšo sposobnost modela na takih primerih pripisali tudi šumu v podatkih. Na podlagi tega bi lahko sklepali, da je konsistentno označevanje nevtralnih novic trd oreh tudi za ljudi.

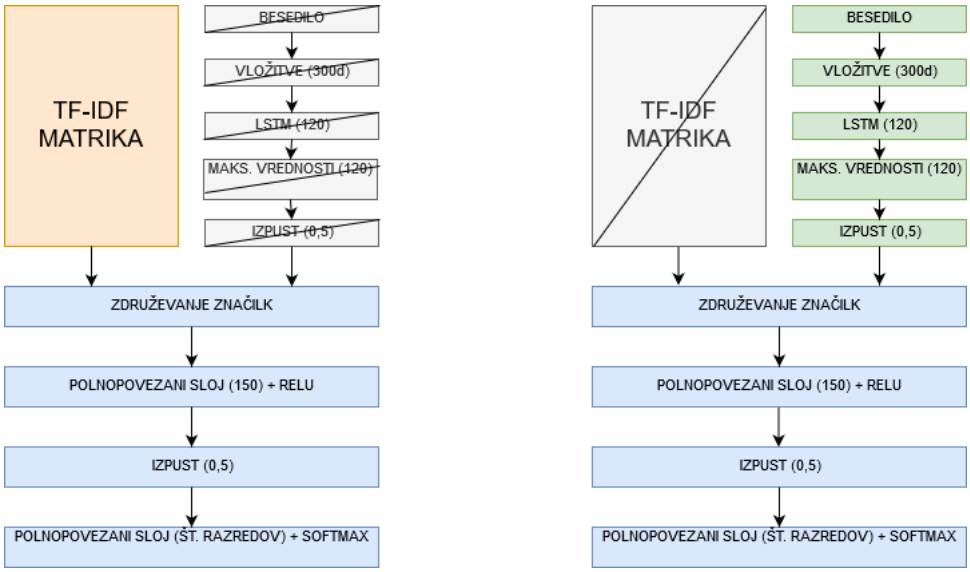
## 6. Primerjava napovedne učinkovitosti posamezne skupine značilk

V zadnjem delu tega članka bomo predstavili še rezultate primerjave napovedne učinkovitosti posamezne skupine značilk nevronskih modelov. Vsi naučeni modeli si delijo isto arhitekturo z dvema obdelovalnima cevovodoma. En cevovod je sestavljen iz nevronskih mrež s povratno zanko z dolgoročnimi in kratkoročnimi pomnilnimi celicami, ki iz vhodnih besedil samodejno izluščijo značilke, drugi cevovod pa vhodna besedila predstavi v obliki TF-IDF matrike. Napovedno učinkovitost posamezne vrste značilk smo preverjali tako, da smo iz arhitekture odstranili en obdelovalni cevovod in model na učni množici naučili samo s pomočjo preostalih značilk. Za vsak model smo nato izračunali klasifikacijsko točnost in oceno F1 ter rezultate primerjali. Shemo eksperimenta predstavlja Slika 4.

Rezultati kažejo, da so TF-IDF uteži za podani problem nekoliko bolj učinkovite, saj model s TF-IDF utežmi doseže za 3,7 % višjo klasifikacijsko točnost in za 5,8 % višjo oceno F1, kot če značilke generira nevronska mreža s povratno zanko. Opazimo lahko tudi, da sta oba modela zelo primerljiva z modeli iz glavne študije, pri čemer model, ki uporablja zgolj TF-IDF uteži kot značilke celo preseže modele iz glavne študije za 0,5 % po oceni F1.

## 7. Zaključek

V raziskavi smo se osredotočili na problem samodejnega zaznavanja sentimenta v slovenskih novicah, problema, ki zaenkrat še ni bil podrobno raziskan. Cilj naloge je bil razvoj modela, ki bi uspešno napovedal sentiment v novicah, kot ga dojema povprečen slovenski bralec. Problem smo zastavili kot klasifikacijski problem, kjer smo poskušali novice kategorizirati v eno od treh kategorij: negativno, pozitivno in nevtralno. V ta namen smo razvili arhitekturo nevronskih mrež, ki posamezno novico predstavi na dva načina, in sicer s pomočjo vektorskih vložitev in kot




Slika 4: Priprava arhitekture modelov za primerjavo vpliva značilk na klasifikacijo.

Tabela 2: Rezultati primerjave napovedne učinkovitosti posamezne vrste značilk.

| Značilke | Predobdelava besedila | Točnost | F1 |
|---|---|---|---|
| Samodejno generirane z LSTM plastmi | Sprememba velikih črk v male | 61,4 % | 57,1 % |
| TF-IDF matrika | Sprememba velikih črk v male | 65,1 % | 62,9 % |

matriko n-gramov s TF-IDF utežmi. Za učenje modelov smo uporabili korpus novic SentiNews, ki vsebuje 10.427 novic, ročno označenih s sentimentom. Skupaj smo naučili pet modelov z enako arhitekturo, ki so se rahlo razlikovali v predstavitvi besedil. Kot osnovo za primerjavo rezultatov smo vzeli rezultate modelov iz Bučar et al. (2018), ki so po naših informacijah tudi edini modeli, ki modelirajo ta fenomen na slovenskih besedilih. Modeli v izvirnem članku so bili evalvirani s petkratno ponovitivjo 10-kratne križne validacije, česar zaradi računske zahtevnosti nevronskih modelov nismo mogli ponoviti. Da bi zagotovili primerljivost rezultatov, smo modele, kot so opisani v izvirnem članku, tudi sami naučili na naši razdelitvi učne in testne množice. Modele nevronskih mrež smo nato primerjali tako z rezultati iz ponovljenega poskusa kot z rezultati modelov iz izvirnega članka.

Rezultati raziskave so vzpodbudni, saj so naši modeli rezultate iz ponovljene študije presegli. Poleg tega so dosegli povsem primerljivo klasifikacijsko točnost z najboljšimi osnovnimi modeli iz izvirnega članka ter se jim po oceni F1 povsem približali. Tako smo potrdili prvo hipotezo, da so lahko nevronske mreže konkurenčne tudi na področju analize sentimenta, kjer navadno prevladujejo bolj tradicionalne metode strojnega učenja, predvsem metoda podpornih vektorjev. Hkrati smo potrdili tudi drugo hipotezo, in sicer, da se lahko modeli na osnovi nevronskih mrež, ki navadno zahtevajo večje količine podatkov, uspešno učijo tudi na manjših korpusih. Vse naše modele smo naučili na korpusu, ki vsebuje zgolj okrog 10.000 novic.

Za konec smo opravili še pregled vpliva posamezne vrste značilk na rezultate klasifikacije. Študijo smo opravili tako, da smo naučili dva modela, pri vsakem pa izločili eno od vrst značilk. Rezultati so pokazali, da so n-grami s TF-IDF utežmi nekoliko bolj diskriminativni za klasifikacijo kot samodejno generirane značilke plasti s povratno zanko, pri čemer model s tako vrsto predstavitve besedil celo preseže modele iz glavne študije glede na oceno F1.

Verjamemo, da arhitektura, uporabljena v tej raziskavi, še ni dosegla mej zmogljivosti. Za začetek bi bilo treba hiperparametre modela dodatno optimizirati. Treba je namreč pripomniti, da smo vrednosti hiperparametrov izbrali glede na vrednosti, ki v praksi dosegajo ugodne rezultate, niso pa nujno optimalne za dani problem. S pomočjo tehnike mrežnega ali naključnega iskanja bi lahko poskusili te vrednosti dodatno optimizirati.

Trenutni nevronski modeli so naučeni in preizkušeni na novicah z ekonomsko, finančno ali politično vsebino, ni pa iz trenutnih rezultatov razvidno, kako bi se ti modeli obnašali pri klasifikaciji novic z drugačno vsebino. Preizkus modelov na novicah iz drugih domen bi bolje ocenil možnost prenosa in uporabe naših modelov na različnih vrstah novic, vendar ker taka označena podatkovna zbirka po naših informacijah ne obstaja, prepuščamo tak preizkus modelov za nadaljnje raziskave.

Najnovejši nevronski klasifikacijski modeli so osnovani na predhodno naučenih jezikovnih modelih, med katerimi sta najbolj znana modela BERT in XLM. Ti modeli so pomembno izboljšali rezultate na številnih nalogah na področju računalniške obdelave naravnega jezika, zato bi bila njihova uporaba za modeliranje sentimenta v slovenskih novicah zanimivo izhodišče za nadaljnje raziskave.

## 8. Zahvala

Avtor članka se zahvaljuje doc. dr. Petri Kralj Novak in izr. prof. dr. Biljani Milevi Boshkoski za vse nasvete, vodenje in mentorstvo, ki sta mu jih nudila med potekom te raziskave.

## 9. Literatura

Ghazaleh Beigi, Xia Hu, Ross Maciejewski in Huan Liu. 2016. An overview of sentiment analysis in social media and its applications in disaster relief. V: *Sentiment analysis and ontology engineering*, str. 313–340. Springer.

Jože Bučar, Martin Žnidaršič in Janez Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52(3):895–919.

Jacob Devlin, Ming-Wei Chang, Kenton Lee in Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin in Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Gagandeep Kaur in Kamaldeep Kaur. 2015. Sentiment analysis on punjabi news articles using svm. *Int. J. Sci. Res., 6 (8)*, str. 414–421.

Diederik P Kingma in Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Abhishek Kumar, Abhishek Sethi, Md Shad Akhtar, Asif Ekbal, Chris Biemann in Pushpak Bhattacharyya. 2017. Iitpb at semeval-2017 task 5: Sentiment prediction in financial text. V: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, str. 894–898.

Xiaodong Li, Haoran Xie, Li Chen, Jianping Wang in Xiaotie Deng. 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23.

Kevin Hsin-Yih Lin, Changhua Yang in Hsin-Hsi Chen. 2008. Emotion classification of online news articles from the reader's perspective. V: *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, zvezek 1, str. 220–226. IEEE.

Youness Mansar, Lorenzo Gatti, Sira Ferradans, Marco Guerini in Jacopo Staiano. 2017. Fortia-fbk at semeval-2017 task 5: Bullish or bearish? inferring sentiment towards brands from financial news headlines. *arXiv preprint arXiv:1704.00939*.

Yelena Mejova. 2009. Sentiment analysis: An overview. *University of Iowa, Computer Science Department*.

Andrew Moore in Paul Rayson. 2017. Lancaster a at semeval-2017 task 5: Evaluation metrics matter: predicting sentiment from financial news headlines. *arXiv preprint arXiv:1705.00571*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng in Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. V: *Proceedings of the 2013 conference on empirical methods in natural language processing*, str. 1631–1642.

Soonh Taj, Baby Bakhtawer Shaikh in Areej Fatemah Meghji. 2019. Sentiment analysis of news articles: A lexicon based approach. V: *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, str. 1–5. IEEE.

Marjan Van de Kauter, Diane Breesch in Véronique Hoste. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with applications*, 42(11):4999–5010.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola in Eduard Hovy. 2016. Hierarchical attention networks for document classification. V: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, str. 1480–1489.

Liang-Chih Yu, Jin Wang, K Robert Lai in Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. V: *Proceedings of the 2017 conference on empirical methods in natural language processing*, str. 534–539.

# Variante in modifikacije (iz)biblijskih frazemov

**Eva Trivunović**

ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša
Novi trg 2, 1000 Ljubljana
eva.trivunovic@zrc-sazu.si

**Povzetek**

V prispevku so analizirani (iz)biblijski frazemi, njihove variante, prenovitvene in neprenovitvene modifikacije v korpusih *Gigafida 2.0*, *Janes* in *slWaC*. Vsi izbrani frazemi in njihove prenovitve so najpogostejši v korpusu *Janes*, kar dokazuje nujnost vključevanja večjega števila raznovrstnih korpusov v jezikoslovne raziskave. Večje število (iz)biblijskih frazemov se vse bolj prilagaja sodobnemu jeziku in stvarnosti, s čimer se odmikajo od izvornega besedila, v nekaterih primerih sta zaradi sestavinskih variant prikrita biblijski izvor in verska konotacija.

**Variants and Modifications of Biblical Phraseological Units**

The paper analyses biblical phraseological units, their variants and modifications, especially phraseological innovations in the Gigafida 2.0, Janes and slWaC corpora. All of the examined phraseological units and innovations are most common in the Janes corpus, which proves the need for including diverse corpora in linguistic studies. In many of the analysed biblical phraseological units a significant change from the biblical text toward modern language and reality can be observed, and in some cases lexical variants result in concealment of their biblical origin and religious connotation.

## 1. Uvod

Ena od definicijskih lastnosti frazemov je stalnost oblike in pomena, vendar je ta pogosto razrahljana in ohlapnejša, kot se zdi na prvi pogled, saj se frazemi v besedilih uresničujejo na zelo različne načine. Mnogi frazemi obstajajo v različnih variantah in oblikah (prim. Meterc, 2019), pogosto je tudi »razhajanje med uslovarjenimi variantami in stanjem v korpusih« (Meterc, 2017: 89). Bistveno je ločevati med izhodiščno obliko, oblikami rabe, normiranimi variantami ter modificiranimi besedilnimi rabami. Izhodiščno obliko tvorijo vrsta in zaporedje sestavin, njihovo minimalno število in razmerje med njimi (Kržišnik, 1996: 133, po Čermák, 1985: 184), oblike rabe vključujejo oblikoslovne prilagoditve in pretvorbe. Od izhodiščne oblike je treba ločiti normirane variante, ki so v jeziku ustaljene in so lahko med seboj enakovredne ali stilno zaznamovane. Od variant ločen pojav je modificirana raba, ki je za razliko od variant neustaljena. Modifikacije so lahko prenovitvene ali neprenovitvene. Neprenovitvene modifikacije so nefunkcionalne spremembe in so z vidika normirane stalnosti neke vrste napake (Kržišnik, 1996: 133–134).

Prenovitvene modifikacije se od neprenovitvenih ločijo po besedilni vlogi. »Prenovitve so namerne in sobesedilno pogojene spremembe pomena in/ali oblike frazeologemov« (Kržišnik, 1990: 399, prim. tudi Kržišnik, 1987; Kržišnik, 1996; Kržišnik, 2006). Vezane so na konkreten kontekst oziroma situacijo, v kateri jih avtor tvori, zato za razliko od izhodiščnega frazema obstajajo le na besedilni ravni, ne pa tudi na jezikovni (Kržišnik, 1990: 401). Pogoj za učinkovitost prenovitve je vzpostavljanje razvidnega razmerja s pomenom izhodiščnega frazema, hkrati pa prenovitev vsebuje nove pomenske razsežnosti, zato je za prenovitve bistveno ustvarjanje napetosti med izhodiščnim pomenom in pomenom prenovitve (Kržišnik, 1990: 401). Prenovitve so v besedilu stilno opazne, pogosto postanejo aktualni pomenski prenosi, ki so v frazemu postali že bolj ali manj prikriti (Kržišnik, 1996: 140–141, po Pogorelec, 1976), velikokrat se nanašajo na aktualne družbene, politične ali osebne dogodke (Justin et. al., 2015: 37).

Prenovitve so zelo pogoste prav pri frazemih zaradi večsestavinskosti, ki v primerjavi z besedo omogoča več možnosti za izrabo semantične potence, in zaradi visoke stopnje prepoznavnosti frazemov, ki izvira iz njihove dolgotrajne prisotnosti v jeziku in jezikovni skupnosti. Predpogoj za prenovitve je tako ustaljenost jezikovnih enot (Kržišnik, 2006: 266), izbira frazemov za prenovitev in možnosti za prenovitve pa so neomejene (Kržišnik, 2006: 271).

Kljub jasno zastavljeni tipologiji in terminologiji je pomembno zavedanje, da je vsako postavljanje mej umetno, saj je jezik kontinuum z mnogimi mejnimi primeri (prim. Meterc, 2019: 37), ločevanje med različnimi tipi pojavitev in razvrščanje zgledov iz avtentičnih besedil tudi ob jasnih definicijah jezikovnih pojmov ni vedno enostavno, predvsem ločevanje med variantami in neprenovitvenimi modifikacijami je pogosto zelo težavno (Kržišnik, 1996: 134). Čeprav je neprenovitvena modifikacija neke vrste napaka, to ne privede vedno do razpada frazeološkega pomena. V teh primerih se uveljavlja t. i. potencialna norma, znotraj katere se gibljeta tako sprejemnik kot tudi tvorec besedila (Kržišnik, 1996: 149). Neprenovitvene modifikacije lahko napovedujejo nove variante in spremembo norme (Kržišnik, 2006: 164).

Pri preučevanju frazeologije so nepogrešljivi korpusi, saj velika količina avtentičnih besedil povečuje verodostojnost raziskav (Gantar, 2007: 22–23), zlasti so pomembni prav pri »ugotavljanju regularnosti pretvorbenih procesov in variantnosti« (Gantar, 2006: 154), ki so skupaj s frazeološkimi pretvorbami osrednja tema prispevka.

## 2. Raziskovalna vprašanja

V raziskavi so predstavljene variante ter prenovitvene in neprenovitvene modifikacije sedmih frazemov v

sodobni slovenščini.[1] Ker prenovitve vzpostavljajo napetost med pomenom izhodiščnega frazema in pomenom prenovitve, je za njihovo razumevanje bistveno, da naslovnik pozna tudi izhodiščni frazem. Ob tem dejstvu se postavlja vprašanje, ali so prenovitvam bolj podvrženi frazemi, ki so tudi sicer (v svoji izhodiščni obliki) pogostejši. Pregledane so bile pojavitve izhodiščnih frazemov in njihove prenovitve, ob tem pa je bilo treba določiti tudi, kaj so ustaljene variante, ki v nekaterih primerih še niso predstavljene v slovarjih. Ustaljenost variante dokazujejo predvsem pogosta raba, uvajalna sredstva in neprenovitveni učinek.

Za ugotavljanje, v katerih primerih gre za ustaljeno variantnost in v katerih za neprenovitvene modifikacije, kjer frazeološki pomen ne razpade zaradi uveljavljanja potencialne norme, je ključnega pomena količina gradiva, ki jo imamo na voljo. Da bi bilo umeščanje pojavitev v ti dve pojavnosti zanesljivejše, so bili v raziskavo vključeni trije korpusi, ki prinašajo različne tipe besedil. Hkrati je to omogočilo tudi raziskovanje vprašanj, ki so vezana na določen tip besedil, in sicer ali so frazeološke prenovitve (ali pa določeni tipi prenovitev) pogostejše v določenih tipih besedil ter ali so med korpusi razlike v pogostosti rabe izhodiščnih frazemov in prenovitev.

Poleg splošnih vprašanj o variantah in modifikacijah so nas zanimale specifike (iz)biblijskih frazemov[2] in njihov razvoj v sodobnem jeziku. Že pri raziskovanju izbiblijskega frazema *videti iver/pezdir v očesu koga/česa, bruna v svojem/lastnem (očesu) pa ne*, ki je bil predmet proučevanja prejšnjih raziskav (Trivunović, 2019), so se pokazale nekatere značilnosti, ki bi lahko bile skupne večjemu številu (iz)biblijskih frazemov. To so predvsem prilagajanje sodobnemu jeziku in stvarnosti ter s tem odmikanje od izhodiščnega besedila (*Biblije*). Nadaljnje vprašanje je, ali ti frazemi izgubljajo očitno versko konotacijo, predvsem z ustaljenimi sestavinskimi variantami, kjer so besede iz verskega diskurza nadomeščene z versko nezaznamovanimi besedami.

## 3. Izbor frazemov in metodologija

V raziskavo je bilo vključenih sedem frazemov (šest pregovorov, ki so spodaj zapisani z veliko začetnico, in en glagolski frazem, ki je zapisan z malo začetnico). Ker je za uspešno prenovitev bistveno poznavanje izhodiščnega frazema, so bili štirje pregovori (*Oko za oko, zob za zob*; *Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega*; *Kdor je brez greha, naj prvi vrže kamen vanjo*; *Beseda je meso postala*) izbrani iz seznama najbolj poznanih in pogostih slovenskih pregovorov in sorodnih paremij (Meterc, 2017: 237–59).

Izhodiščne oblike frazemov so izpisane iz *Slovarja slovenskega knjižnega jezika, druga izdaja* (v nadaljevanju: *SSKJ2*),[3] pri analizi vsakega frazema je opozorjeno na morebitno razhajanje v slovarski obliki s *Slovarjem slovenskih frazemov* (v nadaljevanju: *SSF*) in dejanskim stanjem v korpusih.

Ker je prenavljanje frazemov kreativen proces z neomejenimi možnostmi, se ne da predvideti vseh možnih prenovitev. V raziskavi so obravnavani le naslednji tipi prenovitev.[4]

A. Vzorec, ki bi se lahko nadaljeval ali zamenjal (2):
   1. Oko za oko, zob za zob
      - oko za oko, x za x
      - zob za zob, x za x
   2. Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega
      - dati cesarju, kar je cesarjevega, in (dati) x, kar je x-ovega
      - dati Bogu, kar je božjega, in (dati) x, kar je x-ovega
      - dati x, kar je x-ovega
B. Frazemi z dvodelno strukturo, kjer je lahko en del ohranjen in drugi zamenjan, pri enem frazemu so zajete tudi sestavinske prenovitve (3):
   1. Človek obrača, Bog obrne
      - človek obrača ...
      - ... Bog obrne
      - x obrača, y obrne
   2. Kdor je brez greha, naj prvi vrže kamen vanjo
      - ... naj prvi vrže kamen / naj prvi vrže kamen ...
      - kdor je brez greha ... / ... kdor je brez greha
   3. Duh je sicer voljan, ali meso je slabo
      - duh je voljan ... / … duh je voljan
      - ... meso je slabo / meso je slabo …
C. Zamenjava samostalniške sestavine (2):
   1. Beseda je meso postala
      - beseda je x postala
      - x je meso postal
   2. metati bisere svinjam
      - metati x svinjam
      - metati bisere x

### 3.1. Metodologija

Ker je raziskovalni poudarek na razširjenosti in rabi frazemov in prenovitev v različnih diskurzih, so bili uporabljeni trije korpusi: *Gigafida 2.0*, *Janes* in *slWaC*, za iskanje je bilo uporabljeno orodje noSketch Engine infrastrukture CLARIN.SI.[5] Vsa iskanja so bila izvedena v vseh treh korpusih.

[1] Redki zadetki v korpusih iz starejših besedil zato v raziskavo niso bili vključeni (npr. v korpusu *slWaC* najdemo tudi primer: *Bogu, kar je božjega, cesarju, kar je cesarjevega in narodu, kar je narodovega*, ki je iz časopisa *Novice gospodarskih, obrtniških in narodnih stvari* iz leta 1882).

[2] V diahronem raziskovanju se je izkazalo nujno ločevanje biblijskih (frazemi, ki so bili prevzeti v slovenski knjižni jezik iz jezikov biblijskih predlog kot že izoblikovane frazeološke enote) in izbiblijskih frazemov (frazemi, ki so svojo stalnost in/ali preneseni pomen dobili šele s procesom frazeologizacije v slovenskem jeziku) (Trivunović, 2019: 50–51), vendar je za to raziskavo to zanemarljivo. Zapis »(iz)biblijski« zajema oba tipa frazemov: tako biblijske (*Oko za oko, zob za zob*) kot izbiblijske frazeme (vsi ostali), ki so obravnavani v prispevku.

[3] Razlika v zapisu izhodiščne oblike v primerjavi s *SSKJ2* je le v sestavini *bog/Bog*, ki je v slovarju v vseh primerih zapisana z malo, tukaj pa se uporablja zapis z veliko začetnico. V uporabljenih korpusih se pojavljata oba zapisa, a prevladuje zapis z veliko. Tudi drugi raziskovalci opozarjajo na pravopisno variantnost (prim. Meterc 2017). Glede na izvor frazemov se zdi utemeljen zapis z veliko začetnico.

[4] Drugi tipi prenovitev, ki tem vzorcem ne sledijo (npr. *Zob za zob, oko za oko ... in rit za povišico*; *Kdor še nikoli ni nič privatnega stiskal v službi, naj prvi vrže printer vame*) niso obravnavani.

[5] https://www.clarin.si/noske/index.html

*Gigafida 2.0* je referenčni korpus standardne slovenščine, saj so bila iz prejšnje različice korpusa odstranjena besedila z nestandardnimi jezikovnimi značilnostmi. V primerjavi s prejšnjo različico vsebuje večji delež leposlovnih in šolskih besedil, dodanih je tudi več besedil spletnih besedilodajalcev z večjo produkcijo (npr. novičarski portali, dnevni časopisi). Največji delež besedil v korpusu predstavljajo časopisi (47,8 %), sledijo spletna besedila (28,0 %), revije (16,5 %), stvarna besedila (3,8 %) in leposlovje (3,5 %). Besedila drugih tipov predstavljajo le 0,3 % (Krek et. al., 2019: 1–7; prim. tudi Krek et. al., 2016). Korpus vsebuje 1.333.360.653 pojavnic in je največji od uporabljenih korpusov.

*slWaC* je korpus slovenskega spleta. V prvo različico korpusa so bila vključena besedila samo z domene .si, v novejšo različico pa so vključene tudi spletne strani drugih domen (Erjavec in Ljubešić, 2014: 50). Korpus vsebuje 895.903.321 pojavnic.

Korpus *Janes* sestavljajo besedila slovenskih družbenih omrežij. Najnovejša različica korpusa vsebuje pet zvrsti javno objavljenih uporabniških spletnih vsebin: tvite, forume, komentarje na spletne novice, bloge in komentarje nanje ter uporabniške in pogovorne strani na *Wikipediji* (Fišer et. al., 2020: 225; prim. tudi Fišer et. al., 2016: 68–69). Največji je delež tvitov, ki predstavljajo več kot polovico celotnega korpusa, najmanj pa je komentarjev z *Wikipedije* (Fišer et. al., 2020: 232). Korpus vsebuje 252.904.238 pojavnic.

Najprej so bila izvedena ločena iskanja za pojavitve izhodiščnega frazema in prenovitve, vendar se je kasneje pri večjem številu frazemov izkazalo, da je bolje iskati samo po dveh ali treh najmanj variantnih sestavinah ter nato pregledati rezultate in s pomočjo različnih opcij filtriranja rezultatov ločiti pojavitve izhodiščnega frazema od prenovitev. Za zelo učinkovito se izkazalo tudi iskanje po novo odkritih delih (prim. Meterc, 2019).

Za ponazoritev vzemimo frazem *Oko za oko, zob za zob*. Za iskanje prenovitev sta bili izvedeni dve iskanji, ki sta zajeli nadaljevanje dela *oko za oko* (prvo iskanje) oziroma *zob za zob* (drugo iskanje). Iskanje ni zajelo izhodiščnega frazema, temveč samo prenovitve, saj so bile z iskalnimi pogoji izločene že uslovarjene variante s samostalniki *oko, zob* in *glava* v drugem delu (gl. 4.1).

1. [lemma_lc="oko"] [word="za"] [lemma_lc="oko"] []{0,2} 1:[tag="S.*" & lemma_lc!="oko" & lemma_lc!="zob" & lemma_lc!="glava"] [word="za"] 2:[tag="S.*" & lemma_lc!="oko" & lemma_lc!="zob" & lemma_lc!="glava"] & 1.lemma_lc=2.lemma_lc
2. [lemma_lc ="zob"] [word="za"] [lemma_lc ="zob"] []{0,2} 1:[tag="S.*" & lemma_lc!="oko" & lemma_lc!="zob" & lemma_lc!="glava"] [word="za"] 2:[tag="S.*" & lemma_lc!="oko" & lemma_lc!="zob" & lemma_lc!="glava"] & 1. lemma_lc =2. lemma_lc

Za iskanje pojavitev izhodiščnega frazema sta bili prav tako izvedeni dve iskanji: [word="oko"] [word="za"] [word="oko"] in [word="zob"] [word="za"] [word="zob"]. Pri drugem iskanju so bili iz okolice izločeni zadetki z besedo *oko*, saj je te primere zajelo že prvo iskanje. Iz zadetkov so bile izločene že prej zabeležene prenovitve in nefrazeološke pojavitve zveze, hkrati pa so bile izpisane prenovitve, ki jih iskanje prenovitev ni zajelo, a ustrezajo tipu, ki so predmet raziskave. To so primeri, kjer se v prenovljenem delu ne pojavi samostalnik, temveč beseda pripada drugi besedni vrsti (npr. *prav za prav*), in kjer se pojavi več kot samo *za* med besedama (npr. *šnops za še en šnops*).

Proces iskanja je bil podoben za vse frazeme, pogosto je zahteval veliko iznajdljivosti, nekaterim naključno najdenim primerom je sledilo popravljanje iskalnih pogojev, pogosto v smer večje odprtosti, dopuščanja več prostih mest ter odstopov od pravopisne in frazeološke norme. Predvsem je bilo težavno predvidevati, koliko prostih mest naj se pusti v zapisu in med katerimi sestavinami, zato so se pogosto za učinkovitejša izkazala bolj odprta iskanja (npr. iskanje [lemma="biser"] []{0,2} [lemma="svinja"] je prineslo več relevantnih zadetkov kot [lemma="metati"] [lemma="biser"] [lemma="svinja"], saj je poleg predvidevanja potencialnih mest med sestavinami frazema zajelo tudi glagolske variante).

V nekaterih primerih je bilo treba določen tip iskanja opustiti, ker se iskalih pogojev ni dalo oblikovati tako, da bi našli obvladljivo število zadetkov. Tako v raziskavo niso bili vključene prenovitve *x za x*, ki se pojavljajo brez vsaj enega dela izhodiščnega frazema, saj je pojavitev, ki ustrezajo tem pogojem, preveč. Podobno tudi v primeru (*dati) x, kar je x-ovega*. Izhodiščni frazem se pojavlja tudi brez glagola *dati*, vendar se je iskanje [lemma="dati"]? [tag="S.*" & lemma_lc!="bog" & lemma_lc!="cesar"] []{0,1} [word="kar"] [word="je"] [tag="P.*"] within <s/> izkazalo za preveč odprto, saj se ne da zapisati ukaza, ki bi iskal samo pojavitve, kjer pridevnik izhaja iz samostalnika v prvem delu. Zato je bilo iskanje prenovitev omejeno na pojavitve z glagolom *dati* in prenovitve, ki nadaljujejo enega od delov izhodiščnega frazema.

# 4. Analiza

## 4.1. *Oko za oko, zob za zob*

*SSKJ2* prinaša le eno varianto frazema, v *SSF* pa so zabeležene naslednje variante frazema: *oko za oko, zob za zob*; *oko za oko*; *zob za zob* in *zob za zob, glava za glavo*. Poleg tega je v vseh treh korpusih pogosto tudi zaporedje *Zob za zob, oko za oko*,[6] ki ga *SSF* ne navaja. Meterc (2017: 90) temu dodaja še variante s *kri za kri* ter opozarja, da je za ta frazem značilno podaljševanje z ustaljenimi in neustaljenimi sestavinami, kar je povezano z dejstvom, »da je paremija že v izvornem biblijskem odlomku del daljše verige« (Meterc, 2017: 90). Po pregledu primerov v vseh treh korpusih bi med ustaljene variante lahko šteli še variante z *življenje za življenje*. Med preučevanimi frazemi je bil ta najpogostejši, saj se v korpusih pojavi 780-krat (*Gigafida 2.0*: 330, *Janes*: 200, *slWaC*: 250; relativne frekvence:[7] *Gigafida 2.0*: 0,2476, *Janes*: 0,7905, *slWac*: 0,279), poleg tega se pogosto pojavlja tudi kot citat.

Iskanje prenovitev je zajelo 35 primerov (*Gigafida 2.0*: 17, *Janes*: 10, *slWaC*: 8; relativne frekvence: *Gigafida 2.0*: 0,0128, *Janes*: 0,0395, *slWaC*: 0,0089). V prvi skupini prenovitev (5; *Gigafida 2.0*: 2, *Janes*: 1, *slWaC*: 2) je izhodiščni frazem nadaljevan z različnimi deli telesa ali s samostalnikom *telo*, npr. *oko za oko, koleno za koleno*;

[6] V *Gigafida 2.0* se to zaporedje pojavi približno 50-krat, v *slWaC* 30-krat in v *Janes* 15-krat.

[7] Relativna frekvenca pomeni število pojavitev na milijon pojavnic. Ta podatek omogoča primerjavo med korpusi različnih velikosti.

*Oko za oko, zob za zob, noht za noht, slepič za slepič*.[8] Te prenovitve se pojavijo v literarnih besedilih (2), v publicistiki (2) in v spletnih komentarjih (1).

Dve prenovitvi (*slWaC*) sta zaznamovani v smeri slabšalnosti oziroma vulgarnosti, kar je za (iz)biblijske frazeme nenavadno, saj zanje na splošno velja, da niso zaznamovani v smer neknjižnosti, v slovarjih tudi nimajo oznak, kot so pogovorno, vulgarno ali neknjižno (Kržišnik, 2000: 68). Prvi primer *Oko za oko, češpljo za češpljo* se pojavi kot naslov članka o svingerstvu v reviji Playboy, drugi *zob za zob al pa jajca za jajca* pa se pojavi v komentarju na spletu.

Ostali primeri prenovitev niso vezani na dele telesa, samostalnik izhaja iz sobesedila. Pojavijo se **a)** prenovitve z dvema deloma (14; *Gigafida 2.0:* 8, *Janes*: 5, *slWaC*: 1): *oko za oko, bomba za bombo*; *Zob za zob ali biser za biser*; **b1**) s tremi deli, kjer sta dva dela ustaljena in en prenovitven (7; *Gigafida 2.0*: 3, *Janes*: 3, *slWaC*: 1): *oko za oko, zob za zob, ignoranca za ignoranco*; **b2)** s tremi deli, kjer je en ustaljen del in dva prenovitvena (3; *Gigafida 2.0*: 2, *slWaC*: 1): *zob za zob, udarec za udarec, sramotenje za sramotenje*; ter **c)** s štirimi deli (1; *slWaC*), ki je zanimiv tudi zaradi preobrata od negativnih dejanj k pozitivnim: *zob za zob, udarec za udarec, ljubezen za ljubezen, pozdrav za pozdrav*. V korpusu *Gigafida 2.0* se dvakrat pojavi prenovitev z delom *usluga za uslugo* (enkrat v strokovnem besedilu in enkrat v publicističnem), podobni sta si tudi prenovitvi s *tvit za tvit* in *retvit za retvit* iz korpusa *Janes*. Glede na tip besedila izstopa primer *Kaj imaš raje, zob za zob ali uslugo za uslugo*, ki se pojavi v strokovnem besedilu. V literarnih besedilih se pojavita dve prenovitvi, v publicističnih in spletnih besedilih pa po enajst.

Poleg tega se v dveh primerih prenovitev frazema pojavi v verigi frazemov: *kako bi jima vrnil milo za drago, zob za zob, balon za balon*; *klin se s klinom zbija, oko za oko, zob za zob, smrt za smrt*. Oba primera sta iz korpusa *Gigafida 2.0*, prvi se pojavi v literarnem besedilu, drugi pa v publicističnem besedilu. Zanimiv je tudi tvit iz korpusa Janes, ki najprej nadaljuje vzorec, nato pa z njim prekine: *Zob za zob, ljubezen za ljubezen ... seks pa samo še za denar*.

## 4.2. *Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega*

Izhodiščni frazem se v korpusih pojavi približno 180-krat (*Gigafida 2.0*: 55, *Janes*: 50, *slwaC*: 75; relativne frekvence: *Gigafida 2.0*: 0,0413, *Janes*: 0,1976, *slWaC*: 0,0837) in izkazuje variantnost v zaporedju delov, poleg tega se oba dela osamosvajata in rabita samostojno.[9] Opazno je tudi, da se frazem ne rabi samo v velelniku 2. osebe množine, temveč v različnih oblikah npr. *Naša oblast nas je obvestila, da je čas, da damo cesarju, kar je cesarjevega*; *Daš Bogu kar je božjega, Cesarju, kar je cesarjevega*; *Naj da Bogu, kar je božjega, in cesarju, kar je cesarjevega*.

Le v korpusu *Gigafida 2.0* se pojavita dva primera podaljšave izhodiščnega frazema (obe v publicističnih besedilih), tako da ima prenovitev tri dele. Pri prvem ostaja prenovitev znotraj pomenskega polja verske oblasti, drugi pa je pomensko dlje od izhodiščnega frazema: *Bogu kar je božjega, cesarju kar je cesarjevega, papežu kar je papeževega*; *Vendar, bogu kar je božjega, cesarju, kar je cesarjevega, in kulturi, kar je kulturnega*.

Pojavi se več tipov dvodelnih prenovitev. V treh primerih je enemu od izhodiščnih delov frazema dodan en prenovljen del, kjer je pridevnik tvorjen iz samostalnika, npr. *Dal je cesarju kar je cesarjevega in cestarju kar je cestarjevega*. Dva primera sta iz korpusa *Janes* in se pojavita v tvitih, tretji pa je iz korpusa *slWaC* iz literarnega besedila. V dveh primerih je enemu od izhodiščnih delov frazema dodan prenovljen del, kjer se samostalnik in pridevnik ne ujemata: *Daj bogu, kar je božjega, in Cerkvi, kar je slovenskega*; *Daj cesarju kar je cesarjevega in državi kar je sosedovega*. Prvi primer je iz publicističnega besedila (*Gigafida 2.0*), drugi pa iz spletnega foruma (*slWaC*). Drugemu primeru je podoben še tvit iz korpusa *Janes*, kjer se prav tako v obeh delih pojavi samostalnik povezan s posvetno oblastjo: *Cesarju kar je cesarjevega, državi, kar je državnega*.

Le ena prenovitev je enodelna, torej brez navedbe katerega od delov izhodiščnega frazema: *sva pila kavo na balkonu in dala jutru, kar je jutranjega*, pojavi se v publicističnem besedilu v korpusu *Gigafida 2.0*. Poleg tega se pojavijo še štirje primeri enodelnih zvez (*Gigafida 2.0*: 2, *Janes*: 1, *slWac*: 1), vendar zaradi premalo konteksta ni mogoče z gotovostjo trditi, ali so to prenovitvene ali neprenovitvene modifikacije (napake ali slabo poznavanje frazema) ali pa morda že (skoraj) ustaljene variante. Z največjo gotovostjo lahko o neprenovitveni modifikaciji govorimo v primeru tvita, kjer je samostalnik *cesar* nadomeščen s samostalnikom *car*, saj gre za sopomenki s podobnim slušnim vtisom, tudi pojavitev je samo ena. Drugače je v primerih, kjer je *cesar* nadomeščen z *državo* (enkrat v publicističnem besedilu in enkrat na forumu), *Bog* pa s *papežem* (enkrat v publicističnem besedilu), npr. *Dajo državi kar je državnega*; *Dajmo papežu, kar je papeževega*.

Pri naslednjih štirih pojavitvah (*Gigafida 2.0*: 2, *slWaC*: 2) iz publicističnih (3) in spletnih besedil (1) ostaja konotativni pomen skoraj nespremenjen, saj je kljub zamenjavi sestavin ohranjeno razmerje med posvetno oblastjo (*cesar*, *gospod*, *gospodar*) in Bogom oziroma cerkveno oblastjo (*Bog*, *papež*): *daj cesarju, kar je cesarjevega, in papežu, kar je papeževega*; *srednjeveške mantre o Bogu, kar je božjega, gospodu, kar je gospodovega*; *dati Bogu, kar je božjega, in gospodarju, kar je gospodarjevega* (prenovitev se pojavi v dveh korpusih: *Gigafida 2.0* in *slWaC*). Če vse zgornje primere obravnavamo kot prenovitve, se skupno pojavi 17 prenovitev (*Gigafida 2.0*: 8, *Janes*: 4, *slWaC*: 5; relativne frekvence: *Gigafida 2*.0: 0,006, *Janes*: 0,0158, *slWaC*: 0,0056).

Po definiciji so prenovitve enkratne in namerne tvorbe, vendar se nekatere zveze pojavijo večkrat (predvsem z *država* in *papež*) tako v samostojnih enodelnih zvezah kot v dvodelnih in trodelnih zvezah. V nekaterih kontekstih delujejo prenovitveno (npr. kot tretji del), v nekaterih ne ali pa je konteksta premalo, da bi lahko ugotovili, ali gre za nameren odstop od izhodiščnega frazema ali napako ali morda že za (skoraj) ustaljeno variantnost. V teh primerih gre za odmik v smeri sodobne stvarnosti, predvsem pri pogostem nadomeščanju samostalnika *cesar* s

[8] Primeri so predstavljeni tako, kot so zapisani v korpusu, popravljeni so le presledki ob ločilih.

[9] Krajšanje frazemov in osamosvajanje posameznih delov je v frazeologiji dokaj pogost pojav. V *SSF* je pri (iz)biblijskih frazemih npr. zabeleženo krajšanje frazema: *umiti si roke* (*kot Pilat*).

samostalnikom *država*. Tem prenovitvam sta podobni dve pojavitvi, kjer je skupaj navedenih več izrazov, kar kaže na njihovo izenačitev: *daj cesarju kar je cesarjevega in bogu, papežu, kleru kar je božjega*; *daj cesarju (državi), kar je cesarjevega in bogu, (cerkvi) kar je božjega*.

### 4.3. *Človek obrača, Bog obrne*

Frazem je obravnavan v *SSKJ2* samo v obliki *človek obrača, bog obrne*, vendar je pregledano korpusno gradivo pokazalo, da se frazem odmika od verske konotacije, saj je samostalnik *Bog* pogosto nadomeščen s samostalniki *usoda*, *življenje*, redkeje tudi s *čas* v taki meri, da lahko govorimo o ustaljeni variantnosti in ne o modificirani rabi. Da so to že ustaljene variante izhodiščnega frazema, dokazujejo tudi uvajalna sredstva: *kot pravi pregovor »človek obrača, usoda obrne«*; *Kako gre že tisti rek ... "Človek obrača, življenje obrne"* (podčrtala E. T.). Varianta s samostalnikom *usoda* se pojavi šestkrat v publicističnih besedilih in enkrat na spletu, varianta s *čas* se pojavi le dvakrat, in sicer v publicističnih besedilih, najpogostejša je varianta z *življenje*, ki se pojavi petnajstkrat (šestkrat v publicistiki in devetkrat na spletu). V nekaterih primerih se skupaj pojavi več variant, npr. *Pravijo, da človek obrača, življenje ali bog pa obrne*. Pogosto je frazem rabljen tudi v množini *ljudje obračajo* ali *ljudje obračamo*. V več primerih je samostalnik *človek* nadomeščen z osebnim zaimkom *ti*, *mi* ali *vi*, vendar brez prenovitvenega učinka. Frazem se v korpusih pojavi približno 250-krat (*Gigafida 2.0*: 110, *Janes*: 65, *slWaC*: 75; relativne frekvence: *Gigafida 2.0*: 0,0825, *Janes*: 0,2569, *slWaC*: 0,0837).

V korpusih se pojavi 39 prenovitev (*Gigafida 2.0*: 18, *Janes*: 9, *slWaC*: 12; relativne frekvence: *Gigafida 2.0*: 0,0135, *Janes*: 0,0356, *slWaC*: 0,0134), ki ustrezajo iskanim tipom. V dveh primerih sta skupaj navedena izhodiščni frazem in prenovitev, kjer je opozorjeno, da gre za prenovitev: *Pozabljajo pa na star pregovor, da človek obrača, bog pa obrne, ki se v tem primeru glasi: človek obrača, narava pa obrne*; *Metafizično prepričanje Človek obrača, bog obrne, je modificirano v Človek obrača, zgodovina obrne*. Prvi primer je iz publicističnega besedila (*Gigafida 2.0*), drugi pa iz strokovnega (*slWaC*).

Zabeležili smo pet prenovitev celih delov, v vseh ostalih primerih so prenovitve tvorjenje z zamenjavo samostalnika *Bog* ali obeh samostalniških sestavin. Ena prenovitev je tipa *človek obrača ...*: *človek obrača, župnik spregleda*, ki se pojavi v publicističnem besedilu v korpusu *Gigafida 2.0*. Ostale štiri prenovitve so tipa ... *Bog obrne*. Tri od teh pomensko ostajajo zelo blizu izhodiščnemu frazemu: *ti načrtuješ, bog obrne* (pojavi se dvakrat); *kakorkoli obrneš, bog obrne*. Četrti pa se od izhodiščnega frazema bistveno oddaljuje: *Bog obrne, Joža z štilerco pa požaga*. Zadnje dva primera sta iz korpusa *Janes* (tvita), ostala dva pa iz korpusa *Gigafida 2.0* (publicistični besedili).

Prenovitev tipa *človek obrača, x obrne* je 16 (*Gigafida 2.0*: 7, *Janes*: 3, *slWaC*: 6), vsaka se pojavi samo enkrat in se nanaša na konkretno situacijo (kdo/kaj je nekomu prekrižalo načrte). Samostalnik *Bog* zamenjujejo različni samostalniki iz sobesedila, in sicer osebe (predvsem, a ne samo poklici, npr: *sprevodnik*, *trgovci*, *Tajka*), živali (npr. *opica*), predmeti (npr. *motor*), imena podjetij (npr. *Microsoft*) in pojmi (npr. *trg*, *smrt*); zanimiva je pojavitev, kjer *Boga* nadomešča *Alah* (*Človek obrača, Alah obrne*). Sedem prenovitev je iz publicističnih besedil, po dve iz tvitov in blogov, ena se pojavi v komentarju bloga ter štirje drugje na spletu.

Prenovitev tipa *x obrača, y obrne* je prav tako 16 (*Gigafida 2.0*: 7, *Janes*: 4, *slWaC*: 5). Velika večina primerov se nanaša na nemoralna dejanja ali družbene probleme (npr. *Pohvale obračajo, denar obrne?!*), veliko jih je s področja politike (npr. *politiki obračate, volivci obrnejo*; *Stranke obračajo, mediji pa obrnejo*). V tem pogledu izstopa pojavitev: *Alpinist obrača, gora obrne*. Le ena prenovitev se pojavi dvakrat v istem korpusu (*slWaC*: *Država obrača, državljan obrne*). Večina prenovitev je iz publicističnih besedil, le štiri se pojavijo v tvitih.

### 4.4. *Kdor je brez greha, naj prvi vrže kamen vanjo*

Izhodiščni frazem se je izkazal za zelo variantnega. Že Meterc (2017: 247) opozarja, da zaimek *vanjo* ni obvezna sestavina frazema, temveč fakultativna, kar potrjuje tudi pregled primerov v korpusu. Pojavlja se tudi z obratnim vrstnim redom delov. Po pregledu primerov v vseh treh korpusih so bile kot ustaljene variante upoštevane različice prvega dela s sestavinami iz besednih družin *greh*, *kriv* in *nedolžen*, npr. *kdor (nič) ne greši*, *kdor ni kriv*, *kdor je brez krivde*, *kdor je nedolžen*. Drugi del je manj varianten, najbolj trdno so povezane sestavine *prvi*, *vreči* in *kamen*, zato tudi ni nenavadno, da se ta del pojavlja tudi osamosvojeno kot glagolski frazem *prvi vreči kamen* 'obsojati napake drugih'. Če seštejemo vse pojavitve pregovora (v vseh variantah) in glagolskega frazema, je pojavitev približno 335 (*Gigafida* 2.0: 100, *Janes*: 100, *slWaC* 135; relativne frekvence: *Gigafida 2.0*: 0,075, *Janes*: 0,3953, *slWaC*: 0,1507).

Od vseh preučevanih frazemov se pri tem pojavi največ prenovitev, in sicer 70 (*Gigafida 2.0*: 12, *Janes*: 36, *slWaC*: 22; relativne frekvence: *Gigafida 2.0*: 0,009, *Janes*: 0,1423, *slWaC*: 0,0246). Največ je prenovitev, kjer ostaja nespremenjen del *naj prvi vrže kamen* (60), ki je tudi pri izhodiščnem frazemu manj varianten in se osamosvaja. Večina teh prenovitev ustreza enemu od spodnjih tipov:

A. prenovitve, ki ustrezajo ustaljeni varianti *kdor je brez greha*, torej primeri, ki vsebujejo zanikano negativno dejanje (41; *Gigafida 2.0*: 6, *Janes*: 24, *slWaC*: 11): *Kdor izmed vas je brez slabega singla, naj prvi vrže kamen vanj*; *Tisti, ki je odrastel, ne da bi zajokal, naj prvi vrže kamen v jokajočega otroka*; *Kdor ni prepisal diplomske, naj prvi vrže kamen*. Največ prenovitev je iz tvitov (21), sledijo publicistična besedila (10). Po pet ali manj prenovitev se pojavi v komentarjih na spletu (5), na forumih (3) in v blogih (2).

B. prenovitve, ki ustrezajo ustaljeni varianti *kdor je nedolžen*, torej primeri, ki vsebujejo pozitivno dejanje (19; *Gigafida 2.0*: 4, *Janes*: 6, *slWaC*: 9): *Kdor trdi, da se je odločil za nakup svojega vozila na osnovi takšnega ali drugačnega podatka o obremenitvi okolja, naj prvi vrže kamen*; *tisti, ki je nezmotljiv, naj prvi vrže kamen*; *naj prvi vrže kamen tisti, ki ob VSAKEM znaku 60 VEDNo zavre na 60*. Največ prenovitev je v publicističnih besedilih (7) in nato v tvivih (6), manj jih je v komentarjih na spletu (3), blogih (2) in forumih (1).

Pomensko izstopa prenovitev, ki izrablja dobesedni pomen besede *kamen*: *Naj prvi vrže kamen tisti, ki ga ne*

*misli metat v begunce*, ki se pojavi v tvitu (*Janes*). Prav tako od tega modela odstopa naslednja prenovitev: *V tem trenutku je malo verjetno, da bi prvi vrgel kamen tisti, katerega roka se od dalec vidi, cigava je*, ki se pojavi v komentarju na spletu v korpusu *slWaC*.

Kljub upoštevanju različnih variant smo zabeležili samo šest prenovitev (*Gigafida 2.0*: 1, *Janes*: 4, *slWaC*: 1) z ohranjenim prvim delom frazema. Pri štirih je zveza *vreči kamen* zamenjana z drugim nasilnim ali negativnim dejanjem (*streljati*, *vreči petardo*, *vreči bombo*, *vreči očitek*). Dve prenovitvi sta iz tvitov, ena iz bloga in ena iz publicističnega besedila. Od pomena izhodiščnega frazema se najbolj oddaljujeta dve prenovitvi iz korpusa *Janes*, obe se pojavita na Twitterju: *kdor je brez greha, naj prvi .. dobi provizijo*; *kdor je brez greha je ateist*.

## 4.5. *Duh je sicer voljan, ali meso je slabo*

Že v obstoječih slovarskih virih se frazem pojavi v nekoliko različnih variantah: *duh je sicer voljan, ali meso je slabo* (*SSKJ2*) in *duh je voljan, a meso je slabo* (*SSF*), vendar korpusno gradivo kaže na še večjo variantnost. Samostalnik *duh* je pogosto nadomeščata samostalnika *volja* in *um*. Veznik *sicer* se pojavlja pogosto, a ne vedno. Pridevnik *voljan* je pogosto nadomeščen s pridevniki *močan*, *dober* in *čvrst*. Od veznikov se pojavijo: *a*, *ampak*, *toda*, *le*, *samo*, *ali*, *pa*. V drugem delu se namesto pridevnika *slab* pojavita tudi pridevnika *slaboten* in *šibek*. Pojavi se tudi obraten vrstni red delov. Težko se določi mejo, kaj je ustaljena variantnost in kaj je neprenovitvena modifikacija, saj se variante prepletajo med sabo, zagotovo pa niso prenovitve, saj spremembe niso hotene in tudi pomen ostaja nespremenjen. Če k pojavitvam izhodiščnega frazema štejemo vse zgoraj naštete variante, se frazem pojavi približno 50-krat (*Gigafida 2.0*: 15, *Janes*: 15, *slWaC*: 20; relativne frekvence: *Gigafida 2.0*: 0,0113, *Janes*: 0,0593, *slWaC*: 0,0223). Če pa bi k ustaljenim variantam šteli samo že uslovarjene, se frazem pojavi približno 20-krat.

Zanimale so nas prenovitve z ohranjenim enim delom izhodiščnega frazema. Čeprav je iskanje zajelo tudi pogostejše sestavinske variante, so bile zabeležene le tri prenovitve (v vsakem korpusu ena; relativne frekvence: *Gigafida 2.0*: 0,0008, *Janes*: 0,004, *slWaC*: 0,0011). Prva prenovitev se pojavi kot tvit (*Janes*), vendar ni dovolj konteksta, da bi lahko razbrali natančen pomen prenovitve: *želja je vroča, meso je šibko*. Drugi primer se pojavi v publicističnem besedilu (*Gigafida 2.0*), kjer je samostalnik *duh* zamenjan s samostalnikom *vlada*, ponovno je aktualiziran dobesedni pomen samostalnika *meso*: *VLADA JE VOLJNA, (RDEČE) MESO JE SLABO*. V tretjem primeru je v publicističnem besedilu (*slWaC*) poleg prenovitve enega od delov izhodiščnega frazema prisotno še križanje s frazemom *jekleni živci*: *Meso je šibko, duh še šibkejši, jekleni živci pa zmagujejo*.

## 4.6. *Beseda je meso postala*

Izhodiščni frazem se v korpusih pojavlja v zelo raznolikih oblikah rabe: v različnih časih in številih (*besede so meso postale*, *beseda bo meso postala*, *beseda meso postane*), zanikano (*besede niso meso postale*), pooziraljeno (*besede, ki so meso postale*) ipd. Pojavi se približno 405-krat (*Gigafida 2.0*: 160, *Janes*: 100, *slWaC*: 145; relativne frekvence: *Gigafida 2.0*: 0,12, *Janes*: 0,3953, *slWac*: 0,1618).

Obravnavane prenovitve so razdeljene v dva tipa: *x je meso postala* in *beseda je x postala*, upoštevani so različni spoli in števila. Skupno je iskanje zajelo 50 prenovitev (*Gigafida 2.0*: 25, *Janes* 14, *slWaC*: 11; relativne frekvence: *Gigafida 2.0*: 0,0188, *Janes*: 0,0553, *slWaC*: 0,0123). Bistveno bolj produktiven je prvi tip prenovitev (37; *Gigafida 2.0*: 18, *Janes*: 9, *slWaC*: 10). Nekatere sestavinske menjave se pojavijo večkrat: *govorica*/*govorice* (*Gigafida 2.0*: 5, *Janes*: 1), *ideja* (*Gigafida 2.0*: 2, *slWaC*: 2), *želja* (*Gigafida 2.0*: 1, *slWaC*: 2), *inflacija* (*Gigafida 2.0*: 1, *slWaC*: 1), *ženska*/*ženske* (*Janes*: 1, *slWaC*: 1), vendar kljub več pojavitvam delujejo prenovitveno. Samostalnik *beseda* je najpogosteje zamenjan z eno besedo, a pojavijo se tudi daljše besedne zveze različnih tipov: *robotska rakovica*, *ljubezen do soseda*, *zgodovinski posel v Ligi MLB*, *povest davnih dedov stricev in tet*. Največ prenovitev se pojavi v publicističnih besedilih (27), veliko manj jih je v tvitih (8), po ena prenovitev se pojavi na forumih in v spletnih komentarjih.

Prenovitev tipa *beseda je x postala* je bistveno manj (13; *Gigafida 2.0*: 7, *Janes*: 5, *slWaC*: 1), najpogosteje se pojavi zamenjava s samostalnikom *konj* (trikrat; *Janes*: 2, *slWaC*: 1), morda zaradi vpliva frazema *beseda ni konj*. Dvakrat se pojavi prenovitev *beseda je nota postala*, obakrat v *Gigafida 2.0*. Pogosta je izraba prenovitve za naslove ali vabila na umetniške dogodke, npr. *beseda je nota postala*, *beseda je risba postala*, vendar niso vse prenovitve vezane na umetniško področje, saj se pojavi tudi zamenjava s *tofu* in *motor*. Šest prenovitev se pojavi v publicističnih besedilih, pet v tvitih ter po ena v leposlovju in drugje na spletu.

## 4.7. *Metati bisere svinjam*

V *SSKJ2* je frazem obravnavan samo v obliki *metati bisere svinjam*, *SSF* pa navaja še pretvorbo z glagolnikom *metanje* in opozarja, da se v tem primeru frazem rabi tudi v krajši obliki *metanje biserov*, torej brez samostalnika *svinja*. Izhodiščni frazem se v korpusih pojavi približno 145-krat (*Gigafida 2.0*: 35, *Janes*: 60, *slWaC*: 50; relativne frekvence: *Gigafida 2.0*: 0,0263, *Janes*: 0,2372, *slWaC*: 0,0558) v zelo različnih oblikah rabe, iskanje je zajelo tudi posamostaljene primere z glagolnikom *metanje*. Kot varianti bi lahko opredelili pojavitve s predlogoma *pred* in *med*: *metati bisere pred svinje* in *metati bisere med svinje*, pogost je tudi povratno svojilni zaimek *metati svoje bisere svinjam*. Namesto glagola *metati* se po večkrat pojavljajo: *vreči*, *dati*, *dajati*, *trositi*, po enkrat tudi *deliti* in *sipati*, ki nista ustaljeni varianti, temveč neprenovitveni modifikaciji.

Iskanje, ki je upoštevalo tudi zgoraj navedene variante, je zajelo tri prenovitve (*Gigafida 2.0*: 2, *slWaC*: 1; relativne frekvence: *Gigafida 2.0*: 0,0015, *slWaC*: 0,0011). Pri dveh je samostalnik *biser* zamenjan z nečim dragocenim iz konteksta, obe prenovitvi se pojavita v korpusu *Gigafida 2.0* v leposlovju: *KAJ NAJ IMAM S SLOVENCEM, KI MEČE AVTORJA SONETNEGA VENCA SVINJAM?*; *Ni vedel, da je iz rodu Enakovih, in ni ga zanimalo, da meče obresti porekla svinjam*. V enem primeru iz bloga (*slWaC*) je samostalnik *svinja* zamenjan z nekom, ki ne ceni ponujenega: *Vse bolj besno mečejo bisere občinstvu v glave! – Ne ceni jih (občinstvo) – kako bi tudi jih, če se sami ne?*

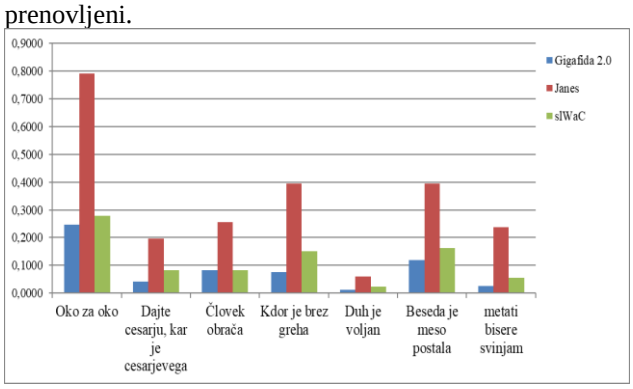
## 5. Diskusija

V tabeli 1 je v prvi vrstici pri vsakem frazemu prikazano približno število pojavitev izhodiščnega frazema in natančno število prenovitev, ki so bile vključene v raziskavo. V drugi vrstici so prikazane relativne frekvence izhodiščnih frazemov in prenovitev v korpusih.

| | Izhodiščni frazem | | | | Prenovitve | | | |
|---|---|---|---|---|---|---|---|---|
| | *GF 2.0* | *Janes* | *slWaC* | **Skupaj** | *GF 2.0* | *Janes* | *slWaC* | **Skupaj** |
| *Oko za oko, zob za zob* | 330 | 200 | 250 | **780** | 17 | 10 | 8 | **35** |
| | 0,2476 | 0,7905 | 0,279 | **0,3143** | 0,0128 | 0,0395 | 0,0089 | **0,0141** |
| *Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega* | 55 | 50 | 75 | **180** | 8 | 4 | 5 | **17** |
| | 0,0413 | 0,1976 | 0,0837 | **0,0725** | 0,006 | 0,0158 | 0,0056 | **0,0068** |
| *Človek obrača, Bog obrne* | 110 | 65 | 75 | **250** | 18 | 9 | 12 | **39** |
| | 0,0825 | 0,2569 | 0,0837 | **0,1007** | 0,0135 | 0,0356 | 0,0134 | **0,0157** |
| *Kdor je brez greha, naj prvi vrže kamen vanjo* | 100 | 100 | 135 | **335** | 12 | 36 | 22 | **70** |
| | 0,075 | 0,3953 | 0,1507 | **0,135** | 0,009 | 0,1423 | 0,0246 | **0,0282** |
| *Duh je sicer voljan, ali meso je slabo* | 15 | 15 | 20 | **50** | 1 | 1 | 1 | **3** |
| | 0,0113 | 0,0593 | 0,0223 | **0,0201** | 0,0008 | 0,004 | 0,0011 | **0,0012** |
| *Beseda je meso postala* | 160 | 100 | 145 | **405** | 25 | 14 | 11 | **50** |
| | 0,12 | 0,3953 | 0,1618 | **0,1632** | 0,0188 | 0,0553 | 0,0123 | **0,0201** |
| *metati bisere svinjam* | 35 | 60 | 50 | **145** | 2 | 0 | 1 | **3** |
| | 0,0263 | 0,2372 | 0,0558 | **0,0584** | 0,0015 | 0 | 0,0011 | **0,0012** |
| **Skupaj po korpusih** | **805** | **590** | **750** | **2145** | **83** | **74** | **60** | **217** |

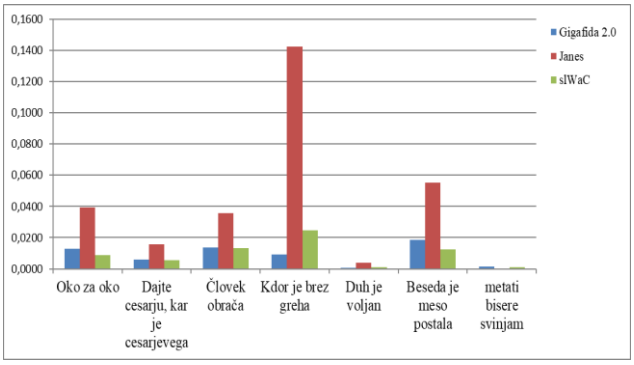
Tabela 1: Zastopanost izhodiščnih frazemov in prenovitev v korpusih

Zanimalo nas je, ali obstaja povezava med pogostostjo izhodiščnega frazema in številom prenovitev. Štirje frazemi z največ pojavitvami izhodiščnega frazema (*Oko za oko, zob za zob*; *Beseda je meso postala*; *Kdor je brez greha, naj prvi vrže kamen vanjo*; *Človek obrača, Bog obrne*) imajo največ prenovitev in trije frazemi z najmanj pojavitvami izhodiščnega frazema (*Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega*; *metati bisere svinjam*; *Duh je sicer voljan, ali meso je slabo*) imajo tudi prenovitev najmanj, vendar so med njimi pomembne razlike. Za frazem *metati bisere svinjam* je tako približno 100 pojavitev več kot za frazem *Duh je sicer voljan, ali meso je slabo*, prenovitev pa imata oba enako (le 3), čeprav bi jih pri prvem pričakovali več. Izstopata tudi frazema *Oko za oko, zob za zob* in *Kdor je brez greha, naj prvi vrže kamen vanjo*, kjer bi več prenovitev pričakovali pri prvem, vendar je prenovitev bistveno več pri drugem. Torej le delno drži, da so pogostejši frazemi pogosteje prenovljeni.




Slika 1: Relativne frekvence izhodiščnih frazemov v korpusih

Če upoštevamo relativne frekvence, so vsi izhodiščni frazemi najpogostejši v korpusu *Janes* in najredkejši v korpusu *Gigafida 2.0*. Razlika v pogostosti je največja pri frazemu *Oko za oko, zob za zob*, najmanjša pa pri frazemu *Duh je sicer voljan, ali meso je slabo*, ki je tudi po številu pojavitev najredkejši.




Slika 2: Relativne frekvence prenovitev v korpusih

Pri frazemih *Duh je sicer voljan, ali meso je slabo* in *metati bisere svinjam* je prenovitev tako malo, da težko govorimo o razlikah med korpusi. Tako kot pri izhodiščnih frazemih je tudi pri prenovitvah opazno, da so najpogostejše v korpusu *Janes*. Največja razlika z ostalimi korpusi je pri prenovitvah frazema *Kdor je brez greha, naj prvi vrže kamen vanjo*, najmanjša pa pri prenovitvah frazema *Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega*. Prenovitve frazema *Človek obrača, Bog obrne* so v korpusih *Gigafida 2.0* in *slWaC* približno enako pogoste. Prenovitve frazemov *Oko za oko, zob za zob*; *Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega* in *Beseda je meso postala* so najredkejše v korpusu *slWaC*, prenovitve frazema *Kdor je brez greha, naj prvi vrže kamen vanjo* pa so najredkejše v korpusu *Gigafida 2.0*. Pri prenovitvah tega frazema je presenetljivo, da je tudi številsko prenovitev največ v korpusu *Janes* (36), sledi *slWaC* (22), najmanj prenovitev je v korpusu *Gigafida 2.0* (12).[10]

[10] O specifičnosti frazemov in njihovih prenovitev za spletna besedila prim. Justin el. al., 2015.

Eno od raziskovalnih vprašanj je bilo vezano tudi na tipe besedil, v katerih so pojavljajo prenovitve, in sicer ali so frazeološke prenovitve in določeni tipi prenovitev pogostejši v določenih tipih besedil. V spodnji tabeli je prikazano, koliko primerov prenovitev je bilo zabeleženih v posameznem tipu besedil; združeni so podatki vseh treh uporabljenih korpusov.

| | Strok. | Lit. | Publ. | Splet. |
|---|---|---|---|---|
| *Oko za oko, zob za zob* | 1 | 5 | 15 | 14 |
| *Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega* | 0 | 1 | 9 | 7 |
| *Človek obrača, Bog obrne* | 1 | 0 | 22 | 16 |
| *Kdor je brez greha, naj prvi vrže kamen vanjo* | 0 | 0 | 20 | 50 |
| *Duh je sicer voljan, ali meso je slabo* | 0 | 0 | 2 | 1 |
| *Beseda je meso postala* | 0 | 1 | 33 | 16 |
| *metati bisere svinjam* | 0 | 2 | 0 | 1 |
| **Skupaj** | **2** | **9** | **101** | **105** |

Tabela 2: Zastopanost prenovitev v tipih besedil

Pričakovano je, da je v strokovnih besedilih prenovitev malo, saj je frazeologija vezana predvsem na neformalne govorne položaje in nestrokovna besedila (Gantar, 2001: 212). Največ prenovitev se pojavi v publicističnih in spletnih besedilih, vendar je to delno posledica sestave izbranih korpusov, saj v njih prevladujeta ta dva tipa besedil. Prav sestava korpusov bi lahko pojasnila tudi majhno število prenovitev v literarnih besedilih, saj ta besedila predstavljajo le majhen delež besedil v izbranih korpusih. Skoraj dve tretjini prenovitev iz spletnih besedil se pojavi v tvitih (67), bistveno manj jih je v komentarjih (13), blogih (11), forumih (7) in drugje na spletu (7).

Razlike se pojavijo tudi pri določenih tipih prenovitev. Pri frazemu *Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega* se prenovitve z dodajanjem tretjega dela pojavijo le v korpusu *Gigafida 2.0* v publicističnih besedilih. Prav tako močno prevladuje raba v publicističnih besedilih pri prenovitvah tipa *x obrača, y obrne* frazema *Človek obrača, Bog obrne*, saj se kar 12 od skupno 16 prenovitev te vrste pojavi v tem tipu besedil. Ne samo, da se prenovitve frazema *Kdor je brez greha, naj prvi vrže kamen vanjo* pojavljajo predvsem na spletu, kar 33 prenovitev (skoraj polovica vseh prenovitev) se pojavi v tvitih. Še posebej je razlika med prenovitvami v tvitih in vseh ostalih virih izrazita pri tipu *kdor* + zanikano negativno dejanje, *naj prvi vrže kamen*, saj je kar 21 primerov od 41 s Tweeterja.

Predvsem dva frazema sta se izkazala za bistveno bolj variantna kot prikazujejo slovarji: *Duh je sicer voljan, ali meso je slabo* in *Kdor je brez greha, naj prvi vrže kamen vanjo*. Pri prvem je variantnih več sestavin v obeh delih, v drugem pa je varianten predvsem prvi del, zaimek *vanjo* v drugem delu pa je neobvezen; oba se pojavljata tudi z obratnim vrstnim redom delov. Poleg frazema *Kdor je brez greha, naj prvi vrže kamen vanjo* je težaven za ločevanje med ustaljenimi variantami in modifikacijami tudi *Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega*, saj večina sestavinskih menjav ostaja znotraj pomenskega polja posvetne in cerkvene oblasti, zato tudi konotativni pomen ostaja blizu izhodiščnemu frazemu, poleg tega pa pogosto zaradi pomanjkanja konteksta ni razvidno, ali gre za hoteno spremembo (prenovitveno modifikacijo), napako (neprenovitveno modifikacijo) ali pa celo za že ustaljeno variantnost. K ustaljenim variantam frazema *Človek obrača, Bog obrne* bi lahko šteli zamenjave samostalnika Bog s samostalniki *življenje*, *čas* in *usoda* ter rabo v množini s samostalnikom *ljudje*. K ustaljenim variantam frazema *Oko za oko, zob za zob* bi poleg variant z *glava* in *kri* lahko prišteli še variante z *življenje*. Pri frazemu *metati bisere svinjam* bi kot varianto gotovo lahko obravnavali pojavitve z glagolom *vreči*, pogojno tudi z *dati*, *dajati*, *trositi* ter predlogoma *med* in *pred*. Bistveno je, da novonastajajoči slovarji obravnavajo te in podobne spremembe, variante in premike v rabi, ki se kažejo v sodobnem gradivu, ter da prinašajo podatek tudi o pravopisni variantnosti (v našem primeru predvsem pri samostalniku *bog*/*Bog*), kjer to podpira korpusno gradivo.

Pri več frazemih je opazno krajšanje in s tem nastajanje novih npr. glagolskih frazemov ali osamosvajanje posameznih delov frazema. To je zlasti opazno pri frazemih *Oko za oko, zob za zob*; *Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega* ter *Kdor je brez greha, naj prvi vrže kamen vanjo*. Pri teh frazemih bi bilo smiselno razmisliti o uslovarjenju novonastalih krajših frazemov kot samostojnih izhodiščnih oblik ali vsaj opozoriti na možnost samostojne rabe posameznega dela.

Pri nekaterih frazemih je opazno oddaljevanje od izvornega biblijskega odlomka oziroma frazema in izgubljanje verske konotacije. Najbolj očitno je to pri frazemu Č*lovek obrača, Bog obrne* pri variantah s samostalniki *življenje*, *usoda* in *čas*. Za frazem *Duh je sicer voljan, ali meso je slabo* so tipične zamenjave sestavin, ki frazem bližajo sodobnemu jeziku in stvarnosti ter s tem oddaljujejo od prvotnega besedila, saj pridevnik *voljan* pogosto nadomeščajo bolj vsakdanji pridevniki *čvrst*, *močan* in *dober*. Odmik k sodobni stvarnosti je opazen tudi pri frazemu *Dajte cesarju, kar je cesarjevega, in Bogu, kar je božjega*, ko pogosto *cesarja* nadomešča *država*, *Boga* pa *papež*, vendar pri teh še ni mogoče z gotovostjo trditi, da so to ustaljene variante.

## 6. Sklep

Glavna prednost uporabe treh korpusov je, da je bilo zajetih veliko raznovrstnih primerov iz različnih tipov besedil in diskurzov, saj je za ugotavljanje mej med ustaljenostjo in neustaljenostjo frazeoloških variant poleg velike količine podatkov ključnega pomena tudi heterogenost virov. Če bi za raziskavo uporabili le korpus *Gigafida 2.0* kot največji in referenčni korpus, bi spregledali mnogo relevantnih pojavitev. Kot je pokazala ta raziskava, so spletna besedila (predvsem družbena omrežja) zelo bogata s frazemi in frazeološkimi prenovitvami, zato je bila njihova vključitev v analizo ključnega pomena. Samo zelo velik referenčni korpus splošne slovenščine za res kvalitetno jezikoslovno raziskovanje ni dovolj, temveč potrebujemo tudi več specializiranih korpusov.

Za celovitejši odgovor glede razmerja pogostosti izhodiščnega frazema in števila prenovitev bi bilo treba pregledati veliko večje število frazemov, na podlagi

analiziranih frazemov lahko predpostavko, da se pogosteje prenavljajo frazemi, ki so tudi v svoji izhodiščni obliki pogostejši, potrdimo le delno. Verjetno imajo večji vpliv na prenovitvene možnosti frazema njegove pomenske lastnosti in dejstvo, da so frazemi kolektivne metafore, skupne celotni jezikovni skupnosti in zato predstavljajo dobro izhodišče za tvorbo novih metafor (prim. Justin et al., 2015: 37). Predvsem pri pojavitvah tvitov v korpusu *Janes* je pogosto premalo konteksta, da bi primere z gotovostjo umestiti v enega od tipov pojavitev, še posebej je to problem pri zamenjavi sestavin znotraj istega pomenskega polja, kjer je meja med ustaljeno variantnostjo in neprenovitveno modifikacijo zabrisana, kar otežuje ustrezen slovarski prikaz frazema in njegovih variant.

Potrdila se je predpostavka, da je za večje število (iz)biblijskih frazemov značilno prilagajanje sodobnemu jeziku in stvarnosti, s čimer se odmikajo od izvornega besedila, a hkrati ne pride do razpada frazeološkega pomena. Nekateri frazemi izgubljajo versko konotacijo z zamenjavo sestavin, ki so v nekaterih primerih že ustaljene variante, pri drugih pa se morda bodo v nadaljnjem jezikovnem razvoju nekatere sedaj neprenovitvene modifikacije tako ustalile in razširile v rabi, da bodo postale ustaljene in normirane variante.

## Zahvala

Prispevek je nastal v okviru usposabljanja mladih raziskovalcev, ki ga financira ARRS pri programu P6-0038.

## 7. Literatura

Tomaž Erjavec in Nikola Ljubešić. 2014. The slWaC 2.0 Corpus of the Slovene Web. V: T. Erjavec, J. Žganec Gros, ur., *9. Konferenca jezikovne tehnologije*, str. 50–55.

Darja Fišer, Tomaž Erjavec in Nikola Ljubešić. 2016. JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. V: *Slovenščina 2.0*, 4(2):67–99.

Darja Fišer, Nikola Ljubešić in Tomaž Erjavec. 2020. The Janes project: language resources and tools for Slovene user generated content. V: *Language resources and evaluation*, 54(1):223–246.

Polona Gantar. 2001. Slovenska frazeologija v dosedanjih slovarjih glede na aktualna slovaropisna načela. *Jezikoslovni zapiski*, 7(1–2): 207–223.

Polona Gantar. 2006. Korpusni pristop v frazeologiji in slovarske aplikacije. V: A. Vidovič Muha, ur., *Slovensko jezikoslovje danes*, str. 151–163. Slavistično društvo Slovenije.

Polona Gantar. 2007. *Stalne besedne zveze v slovenščini: korpusni pristop*. Založba ZRC, ZRC SAZU.

Martin Justin, Nejc Hirci in Polona Gantar. 2015. Rana ura, slovenskih fantov grob: analiza frazeoloških prenovitev v spletni slovenščini. V: D. Fišer, ur., *Slovenščina na spletu in v novih medijih*, str. 33–37. Znanstvena založba Filozofske fakultete.

Janez Keber. *Slovar slovenskih frazemov*, www.fran.si, dostop 20. 4. 2020.

Simon Krek, Polona Gantar, Špela Arhar Holdt in Vojko Gorjanc. 2016. Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres. V: T. Erjavec, D Fišer, ur., *Jezikovne tehnologije in digitalna humanistika 2016*. 200–202.

Simon Krek, Špela Arhar Holdt, Jaka Čibej, Andraž Repar in Nikola Ljubešić. 2019. *Specifikacije izdelave korpusa Gigafida 2.0*.

Erika Kržišnik. 1987. Prenovitev kot inovacijski postopek. *Slava I*(1):49–56.

Erika Kržišnik. 1990. Tipologija frazeoloških prenovitev v Cankarjevih proznih delih. *Slavistična revija*, 38(4):399–420.

Erika Kržišnik. 1996. Norma v frazeologiji in odstopi od nje v besedilih. *Slavistična revija*, 44(2):133–154.

Erika Kržišnik. 2000. Biblične stalne zveze v Slovarju slovenskega knjižnega jezika in v slovenskem jeziku. *Razprave razreda za filološke in literarne vede*(XVII):67–80.

Erika Kržišnik. 2006. Izraba semantične potence frazemov. V: A. Vidovič Muha, ur., *Slovensko jezikoslovje danes*, str. 259–279. Slavistično društvo Slovenije.

Matej Meterc. 2017. *Paremiološki optimum: najbolj poznani in pogosti pregovori ter sorodne paremije v slovenščini*. Založba ZRC, ZRC SAZU.

Matej Meterc. 2019. Analiza frazeološke variantnosti za slovarski prikaz v eSSKJ-ju in SPP-ju. *Jezikoslovni zapiski*, 25(2):33–45.

*Slovar slovenskega knjižnega jezika, druga, dopolnjena in deloma prenovljena izdaja*, www.fran.si, dostop 20. 4. 2020.

Eva Trivunović. 2019. Diahrono raziskovanje biblijskih in izbiblijskih frazemov. *Jezikoslovni zapiski*, 25(2):47–61.

# Korpus metafor KOMET 1.0

**Špela Antloga**

Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška cesta 46, 2000 Maribor
s.antloga@um.si

## 1 Uvod

Prepoznavanje vrednosti in razširjenosti metaforičnih izrazov v jeziku je v zadnjih dvajsetih letih vodilo k povečanemu zanimanju za sistematično identifikacijo metaforičnih izrazov in metafor v posameznih jezikih, kljub temu pa je identifikacija metaforično rabljenega jezika še vedno eden od izzivov, s katerim se soočata sodobno korpusno in računalniško jezikoslovje.

Po sodobni definiciji metafore niso samo jezikovni izraz, ampak v metaforah tudi razmišljamo. Med različnimi teoretičnimi pristopi, ki so se konec 70. let prejšnjega stoletja v času t. i. kognitivnega preobrata posvečali preučevanju metafore, je še danes ena najvidnejših zagotovo 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 (Bratož 2010: 19). 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. Jezikovne, vizualne in multimodalne uresničitve konceptualnih metafor pa nam omogočajo vpogled v te medpodročne preslikave.

Na prelomu tisočletja so kritiki kognitivistom očitali subjektivnost, neempiričnost in nekonsistentnost pri prepoznavanju (iskanju) in razlagi metafor. Kot odziv na tovrstne očitke je skupina raziskovalcev[1] za angleščino[2] razvila postopek za ugotavljanje metaforičnih besed v besedilu, poimenovan MIPVU (Steen in drugi, 2010),[3] 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. 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.[4]

## 2 Korpus metafor KOMET 1.0

Korpus KOMET (KOrpus METafor) 1.0 je korpus slovenskih publicističnih, leposlovnih in spletnih besedil v obsegu 200.622 besed, v katerem so ročno označene metaforične besede, torej besede, katerih pomen v danem kontekstu odstopa od osnovnega pomena te besede. Označevanje metaforičnih besed je temeljilo na omenjenem postopku za identifikacijo metafor MIPVU, kar je pomenilo označevanje vseh jezikovnih elementov, povezanih s pomenskimi prenosi na osnovi medpodročnih preslikav. Označenim metaforičnim besedam je bilo nato pripisanih 69 pomenskih shem, ki funkcionirajo kot sistem kategorij, ki so strukturirane glede na določen kontekst, ki jih motivira. To nam omogoča, da znotraj določene pomenske kategorije (npr. naravni pojavi, čas, prostorska orientacija, družina, premikanje itd.) prepoznamo metaforične besede, ki so lahko potencialno uresničitev neke konceptualne strukture.

[1] Skupina se je poimenovala Pragglejaz po prvih črkah imen raziskovalcev, ki so sodelovali projektu: Peter Crisp, Ray Gibbs, Alan Cienki, Graham Low, Gerard Steen, Lynne Cameron, Elena Semino, Joe Grady, Alice Deignan in Zoltan Kövecses.

[2] Postopek je s prilagoditvami značilnostim posameznih jezikov sprožil zanimanje za identifikacijo metaforičnih izrazov in metafor v češčini, litovščini, madžarščini, poljščini, srbščini, španščini ter za izdelavo korpusov metafor v angleščini (Steen G. J. idr., 2010), ruščini (Yulia Badryzlova in Olga Lyashevskaya, 2017) in hrvaščini (Kristina Despot idr., 2019).

[3] Metaphor Identification Procedure Vrije Universität.

[4] Projekt izdelave korpusa je bil financiran v okviru projekta CLARIN.si 2019. Korpus je dostopen na naslovu http://hdl.handle.net/11356/1293.

## 2.1 Metodološka izhodišča za izdelavo korpusa

Osnovna baza je bila oblikovana z uravnoteženim naborom zvrstno različnih besedil iz korpusa MAKS.[5] Za označevanje metaforičnih besed in pomenskih shem je bilo uporabljeno orodje WebAnno, kjer je postopek identifikacije metaforičnih besed potekal po naslednjih korakih:

1. Branje celotnega besedila za splošno razumevanje pomena.
2. Za vsako leksikalno enoto v besedilu je bil določen njen pomen v kontekstu. Upoštevano je bilo okolje, v katerem se leksikalna enota nahaja (kaj je pred in za njo).
3. Za vsako leksikalno enoto je bilo ugotovljeno, ali ima osnovnejši pomen v drugih kontekstih. Osnovni pomeni so običajno:
   • konkretnejši (to pomeni, da si je tisto, kar evocirajo, lažje predstavljati, videti, slišati občutiti, vohati ali okusiti),
   • povezani s telesnim delovanjem,
   • bolj natančni in jasni.
4. Če se je kontekstualni pomen razlikoval od osnovnega, je bila beseda označena kot metaforična beseda (MRW).

V naslednjem koraku je bila označenim metaforičnim besedam dodana tudi oznaka 'frame' za določanje pomenske sheme. S tem je bila označena beseda uvrščena v določeno pomensko področje, ki je bilo opredeljeno glede na osnovni pomen označene besede in funkcionira kot »izhodiščna domena« v pojmovanju konceptualne metafore.

### 2.1.1 Postopek določanja metaforičnih besed (oznaka 'MRW')

Pri postopku identifikacije so bile označene leksikalne enote (glagoli, samostalniki, pridevniki, prislovi in predlogi oz. stalne besedne zveze), katerih kontekstualni pomen se je razlikoval od njihovega osnovnega pomena. Osnovni pomen je bil določen na podlagi hierarhije slovarskih pomenov danega leksema.[6] Označenim metaforičnim besedam je bila nato pripisana informacija o tem, ali gre za (1) indirektno metaforo, (2) direktno metaforo ali (3) mejni primer. Označeni so bili tudi (4) metaforični signalizatorji.

(1) Z oznako 'indirektna metafora' (MRWi) so bili označeni kontekstualni pomeni besed, ki niso osnovni pomeni teh besed in jih lahko pojasnimo kot preslikavo z ene domene na drugo. Npr. v povedi

*Lepi, slavni in bogati naravnost obožujejo knežjo družino, ki vsakič poskrbi za pikantne govorice*

sta bili z oznako MRWi označeni besedi *naravnost* in *pikantne*. Osnovni pomen besede *naravnost* namreč izraža gibanje ali položaj, osnovni pomen besede *pikanten* pa prekomerno začinjenost. Kontekstualni pomen je sicer že lahko konvencionaliziran in zato opredeljen v slovarju, npr. pomen intenzivnosti (tretji pomen iztočnice *naravnost*) ali privlačnosti (drugi pomen iztočnice *pikanten*). Vendar to ne pomeni, da je že vsaka pomenska enota z več pomeni označena kot metaforična, če v konkretnem primeru ni rabljena v slovarskem prvem pomenu. Kontekstualni pomen mora biti dovolj oddaljen od osnovnega pomena, da je označen kot metaforičen.

(2) Z oznako 'direktna metafora' (MRWd) so bile označene besede, ki so sicer rabljene v svojem osnovnem pomenu, a je kontekstualni pomen lahko pojasnjen kot preslikava z ene domene na drugo. Npr. v povedi

*Obnašate se kot sestradani golob, ki kljuva kot zmešan, prepričan, da je samo za las oddaljen od cilja*

sta besedi *sestradan* in *golob* sicer rabljeni v svojem osnovnem pomenu 'ki je zaradi pomanjkanja hrane zelo izčrpan, oslabel' in 'domača ali divja ptica z majhno glavo, kratkim vratom in krepkim telesom', vendar sta

[5] Korpus MAKS v obsegu približno 10 milijonov besed z besedili iz publicistike, leposlovja in spletnih portalov je bil izdelan v okviru projekta Slovenščina na dlani. Več na http://projekt.slo-na-dlani.si/sl/.

[6] V uvodu v SSKJ (§ 75) je navedeno, da je na prvem mestu večpomenskih iztočnic osnovni pomen, to je tisti, ki je v sodobnem knjižnem jeziku najbolj nevtralen oziroma prevladujoč. Nadaljnji pomeni so razporejeni po stopnji odvisnosti od osnovnega pomena. Včasih na prvem mestu ni najmočnejši, ampak pomensko izhodiščni pomen iztočnice, zlasti če je to potrebno zaradi razvrstitve nadaljnjih pomenov (§ 76). Kadar ima beseda več pomenov, ki so enako močni, stoji na prvem mestu pomen, ki ima največ ali najmanj pomenskih prvin. Nadaljnji pomeni si sledijo, kakor število teh elementov ali upada ali narašča (§ 77). Manjši pomenski premiki so obdelani kot podpomeni ali pomenski odtenki v okviru nadrejenega pomena (§ 78). Pri slovničnih besedah odloča o razvrščanju razlag pogostnost, ob enaki pogostnosti pa ustaljeno slovnično zaporedje (§ 79).

skupaj s primerjalnim veznikom *kot* (metaforičnim signalizatorjem) zaradi preslikave pomena z izhodiščne živalske domene na človeka označeni kot direktno rabljeni metaforični besedi. Enotam, ki so bile označene kot MRWd, ni bila dodana oznaka 'frame' za pomensko shemo, saj je ta že neposredno izražena.

(3) Kot mejni primeri (WIDLI) so bile besede označene takrat, ko iz konteksta ni bilo mogoče določiti, ali gre za metaforični ali osnovni pomen besede. Npr. v izseku intervjuja, kjer nogometni igralec na vprašanje odgovori: *»Do derbija se je že nakopičila utrujenost ob nizu zaporednih tekem,«* novinar doda, da se je *mladi hrvaški reprezentant tako kot večina najprej dotaknil bolečih mišic*. V tem primeru ni mogoče določiti, ali gre za osnovni pomen glagola *dotakniti se* ('približati se tako, da pride do dotika, stika') ali metaforični ('na kratko, nekoliko spregovoriti o čem'), zato je enota označena kot mejni primer.

(4) Z oznako metaforični signalizatorji (MFlag) so bile označene besede, ki signalizirajo posamezne besede in slovnične strukture, npr. *kot, kakor, enačiti* (kaj) z, *metaforično* (*rečeno*) ipd.

Posebej so v korpusu označene še frazeološke enote (ki so razumljene kot ena leksikalna enota z enim pomenom, ki pa je metaforičen), metonimije (pri čemer ne gre za preslikavo pomena z ene domeno na drugo, ampak se preslikava dogaja znotraj ene (iste) domene) in poosebitve.

### 2.1.2 Postopek določanja pomenskih shem (oznaka 'frame')

Z oznako 'frame' je bila predhodno označena metaforična beseda uvrščena v določeno pomensko shemo, ki je bila opredeljena glede na osnovni pomen označene besede (pomen, iz katerega izhaja preslikava). Določanje pomenske sheme označenim metaforičnim besedam omogoča, da znotraj določene pomenske kategorije (npr. čas, prostorska orientacija, premikanje, čustvovanje itd.) prepoznamo metaforične izraze, ki so lahko potencialno uresničitev določene konceptualne strukture. Trenutno je označenim metaforičnim besedam dodanih 69 pomenskih shem. Seznam pomenskih shem ni končen in se lahko ob dodajanju novih besedil še dopolnjuje. Npr. v povedi

*Zato je telekomunikacijska družba pripravila napad na oba monopolista na evropskem trgu*

je bila beseda *napad* zaradi kontekstualnega pomena označena kot indirektna metaforična beseda. Pripisana ji je bila pomenska shema 'boj', kar izhaja iz osnovnega pomena označene besede 'oborožena vojaška akcija z namenom uničiti, onesposobiti nasprotnikove sile'.

### 2.1.3 Splošna vodila pri označevanju metaforičnih besed glede na besedno vrsto

Za označevanje metaforičnih besed so bila določena splošna vodila za posamezne besedne vrste. Osnovna enota za označevanje je leksikalna enota, ki je opredeljena kot beseda, stalna besedna zveza, zveza glagola in členka/predloga ali frazem z enim samostojnim pomenom.

Kontekstualni pomen **glagolov** je bil določen glede na hierarhijo pomena posameznega glagola v SSKJ, pri čemer kontekstualni pomen, ki sicer v slovarju ni označen kot osnovni pomen tega glagola, a vsebuje kvalifikator »s širokim pomenskim obsegom«, ni označen kot metaforična beseda.

Kontekstualni pomen **samostalnikov** in **pridevnikov** je bil določen glede na hierarhijo pomena posamezne besede v SSKJ.

Večina označenih **prislovov** je zaenkrat dobila oznako 'WIDLI' (mejni primeri), saj slovar načeloma eksplicitno ne hierarhizira pomenov prislova, ampak skupaj s primeri navaja zgolj glagol ali pridevnik, iz katerega je tvorjen. Zgolj na podlagi tega osnovnega pomena posameznega prislova ni mogoče določiti, saj ne predpostavljam, da se hierarhija pomenov ohranja pri prehodu v drugo besedno vrsto (enako velja za glagolnike, če nimajo določenih pomenov).

Določanje kontekstualnega pomena **predlogov** izhaja iz tega, da primarno poimenujejo prostorsko razmerje, čeprav je pri hierarhiji pomenov v slovarju upoštevana pogostnost (gl. opombo 6). Za razliko od označevalne sheme za angleščino je za vsak predlog osnovni pomen določen še po sklonu, s katerim se veže, npr. *pod* + tožilnik (za izražanje premikanja, usmerjenosti k spodnji strani česa) ali npr. v metaforičnem pomenu 'biti *pod* vplivom koga' oziroma *pod* + orodnik (za izražanje položaja na spodnji strani česa) ali npr. v metaforičnem pomenu 'zlomiti se *pod* pritiskom'.

## 2.2 KOMET 1.0 v številkah

V korpus metafor KOMET 1.0 v obsegu 200.622 besed je vključenih 42 publicističnih besedil v obsegu 95.531 besed (48 % celotnega korpusa), 8 strokovnih besedil v obsegu 23.732 besed (12 % celotnega korpusa), 5 leposlovnih besedil v obsegu 38.030 besed (19 % celotnega korpusa) in spletna besedila v obsegu 43.329

besed (21 % celotnega korpusa). Od 10.632 označenih metaforičnih besed (5,3 % vseh besed) je bilo največ implicitnih metafor (MRWi), in sicer 86,3 %, kot mejnih (WIDLI) je bilo označenih 11,6 % vseh označenih metaforičnih besed, najmanj je označenih direktnih metafor (MRWd), in sicer 2,1 %. Besednovrstno je 39,2 % vseh označenih metaforičnih besed glagolov, 27,6 % samostalnikov, 13,1 % pridevnikov, 12,7 % predlogov in 7,4 % prislovov. Med pripisanimi 69 pomenskimi shemami prevladujejo (po vrsti): premikanje, vidno zaznavanje, telo, razsežnost, življenjski procesi, čas, boj, smer, oblika, tekočina, gib, potovanje, ogenj, objekti, lov itd.

# 3 Zaključek

Za vpogled v metaforično konceptualizacijo stvarnosti je zaradi kulturne razsežnosti metafore smiselno združevanje različnih teoretičnih pristopov. Tradicionalni kognitivni vidik preučevanja konceptualnih struktur smiselno nadgrajuje korpusni pristop, ki omogoča natančnejšo in bolj sistematično analizo metaforičnih izrazov v različnih besedilih. Namen izdelave korpusa metafor ni bilo iskanje niti interpretacija metafor. Postopek identifikacije metafor je namreč temeljil na predpostavki, da ima vsak pomen besede, ki v določenem kontekstu odstopa od osnovnega pomena te besede, potencial, da predstavlja jezikovno uresničitev neke (konceptualne) strukture. Za nadaljnje delo s korpusom bodo še potrebne izboljšave, predvsem pri prepoznavanju metaforičnosti v povezavi s posameznimi skloni v smislu »osnovnega pomena«, ki ga izraža posamezen sklon, pri opazovanju metaforičnosti zvez glagola s členki in predlogi ter pri natančnejši opredelitvi pomenskih shem. Že sedaj pa korpus omogoča kvantitativno analizo, primerjavo metaforično rabljenih besed in pomenskih shem glede na besedilni tip/žanr, analizo izrazov znotraj določenega pomenske sheme, analizo posameznih izrazov, ki so potencialno uresničitev več konceptualnih struktur ipd.

# 4 Literatura

Špela Antloga. 2010. Metaphor corpus KOMET 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1293.

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.

Silva Bratož. 2010. *Metafore našega časa*. Fakulteta za management, Koper.

Lyanne Cameron. 1999. Operationalising 'metaphor' for applied linguistic research. V: *Researching and applying metaphor*, str. 3–28. Cambridge University Press.

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 LT4DH workshop at COLING 2016*. https://webanno.github.io/webanno/.

Alan Cienki. 2008. Why study metaphor and gesture. V: *Gesture Studies*, str. 5–25. John Benjamins, Amsterdam/Philadelphia.

Kristina Despot, Mirjana Tonković, Mario Brdar, Benedikt Perak, Ana Ostroški Anić, Bruno Nahod, 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. John Benjamins Publishing Company.

Raymond W. Gibbs. 1999. Researching Metaphor. V: *Researching and applying metaphor*, str. 29–47. 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.

Graham Low. 1999. Validating metaphor research projects. V: *Researching and applying metaphor*, str. 48–68. Cambridge University Press.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol* 22 (1): 1–39.

Elena Semino. 2017. Corpus linguistics and metaphor. V: *The Cambridge Handbook of Cognitive Linguistics*, str. 463–476. Cambridge University Press.

*Slovar slovenskega knjižnega jezika, druga, dopolnjena in deloma prenovljena izdaja*. www.fran.si.

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*. John Benjamins, Amsterdam/Philadelphia.

# Working with Digital Devices in Teacher Training Using the Example of the Teaching-Learning-Lab

**Magdalena Schlintl †, Kerstin Pawluch‡, Mara Rader^, Verena Novak-Geiger ***

† School of Education, University of Klagenfurt
Universitätsstraße 65-67, 9020 Klagenfurt
Magdalena.Schlintl@aau.at
‡ School of Education, University of Klagenfurt
Universitätsstraße 65-67, 9020 Klagenfurt
Kerstin.Pawluch@aau.at
^Department of German Studies, Faculty of Humanities, University of Klagenfurt
Universitätsstraße 65-67, 9020 Klagenfurt
Mara.Rader@aau.at
* School of Education, University of Klagenfurt
Universitätsstraße 65-67, 9020 Klagenfurt
Verena.Novak@aau.at

## 1. Introduction

In recent decades, the widespread use of digital technologies has changed the world dramatically. The impact on the economy, social life, and the learner is enormous. Schools, too, are now increasingly trying to use digital technologies and digital tools in the classroom and for learning at home. Among other subjects, language teaching in particular has become a field where digital devices and online tools, as well as apps, can be used in a variety of ways to intensify and monitor the learning progress.

The use of digital technologies has already found its way into continuing professional education and adult education. Even before the COVID-19 crises, the use of digital technology in the educational field increased. Curran et al. (2019) examined the possibility of self-directed learning in adulthood by the support of technology increased. The latter was described as an important resource in many professional fields. Especially, the own laptop and mobile phone are used for self-directed learning by workers in the health sector for professional and adult education and there is still a need to identify creditable sources in the internet. However, the use of digital technologies is not limited to lifelong education. Among others, Altuna and Lareki (2015) dealt with the implementation of digital technologies in schools. The study results show that teachers training, comprising theoretical aspects of education and training as well as focusing on the use of digital technology, is needed. Where implemented, positive effects of the use of digital devices were found. For example, Sung et al.'s (2015) meta-analysis showed that language instruction supported by mobile devices has resulted in a meaningful improvement with mean effect size (d= .55). Mobile devices, such as mobile phones and tablets had a larger effect on the improvement for language learning than laptops or computers. Furthermore, better achievement was seen in second language learning than in first language learning and the integration of mobiles together with multiple learning and teaching strategies produced higher effects.

Another description of the innovations and changes in teaching and learning of pupils was conducted by Brandhofer et al. (2018) in Austria's National Education Report. The authors examine mobile learning with notebooks and tablets, the digital textbook, augmented reality, e-portfolio, mind and concept maps. In addition, they describe gamification, game-based learning, and educational robotics, Maker Education and the use of the cloud in schools. Brandhofer et al. emphasise that a digitisation strategy can be beneficial for all parties involved if teachers are convinced of the importance of digital media, want to use them in class and are motivated to promote the digital skills of their students (2018, 342). In addition, Sung et al. (2015) point to the importance of enabling language teachers to adapt to mobile assisted language learning (MALL) activities and to reduce the workload of innovative teaching practices. Consequently, it is important that all teachers acquire a certain level of digital literacy and are able to teach with and about digital media. This field of expertise must therefore be increased in teacher education and professional teacher training (Brandhofer et al. 2018, 344). The School of Education at the University of Klagenfurt has regarded these recommendations seriously and provided infrastructure that supports students and teachers at the university in meeting these requirements.

The Teaching-Learning-Lab (TLL) was founded about two years ago aiming to support teachers and students in planning and developing lessons. The teaching-learning lab can be divided into four segments: laboratory, teaching, digitization and infrastructure, and interface and public relation.

Firstly, it can be described as a laboratory. It offers students and teachers infrastructure for conducting empirical surveys and validating empirical methods. The lab also serves as a development environment, where workshops for students and other participants are held. Secondly, it is integrated into regular teaching courses at the university and provides the academics with both analogue and digital teaching media. It offers them the possibility of carrying out the lessons on the premises of TLL, where in cooperation with students and teachers a pool of ideas for innovative academic research papers (Bachelor, Master, PhD) is built. Thirdly, the TLL contributes to digitalisation with its services and provides the necessary technology. It offers equipment with interactive devices for groups in combination with wireless image and sound transmission. Fourthly, it functions as an interface to practice, stimulates the exchange of experience with teachers and paves the way for cooperative projects.

The target group of the Teaching-Learning-Lab service is very broad: students, especially those in teacher training, teacher educators as well as academics from different fields of expertise. Visitors receive counselling on issues of education, the use of digital technology in their classrooms and are given the opportunity to get to know some of the digital devices in a closed space. In this way, many participants overcome their fears and lose their prejudice against using new technologies. Talented and engaged young students find a surrounding, where their research can thrive and their peers or mentors give them the needed support. Moreover, there are opportunities for the coordination between the subjects. Among other things, the aim of the Teaching-Learning Lab is to prepare students for professional practice, especially with regard to the use of digital tools, as well as research and research-based teaching on all aspects of the teaching profession.

## 2. Utilities

The Teaching-Learning-Lab provides analogue as well as digital media, such as, workbooks, textbooks, teaching games, CDs, radio plays, DVDs or iPads with installed learning apps. All materials — analogue and digital — are available in sufficient quantities, can be borrowed by students or teachers and used directly in class. The available media can also serve as a stimulus to create teaching scenarios and develop ideas for lessons.

### 2.1. Analogue media

The textbooks and workbooks are partly aimed at specific areas of teaching, but there are also materials that are suitable for many teaching subject. For example, suggestions for motivating lesson introductions in the form of movement games or brain and puzzle games.

Apart from textbooks and workbooks, the Teaching-Learning-Lab provides games that promote learning and cognition. These games can be used in a suitable setting in the classroom to practice grammar, reading and writing as well as oral communication in a playful way. The Teaching-Learning-Lab provides material in the form of games for spatial imagination or learning probability distributions. There is a sufficient number of identical games at disposal to enable teachers to borrow enough games for all their students. Most of the games are published products, but can be adapted to suit the needs of the classroom.

For teachers who are looking for ideas, the Teaching-Learning-Lab also offers a wide range of different radio plays and DVDs which can be used in language and literature lessons. For instance, those who would like to test the reading competence of students can have a look at the large selection of different diagnostic tests. The Teaching-Learning-Lab currently offers competence-specific diagnostic tests that focus on reading comprehension, reading fluency or spelling. Among the established tests are SLS, ELFE II, HAMLET 3-4 or LGVT 5-12+[1]. These can also be made available for teaching purposes.

### 2.2. Digital media: working with tablets

Even though the large number of different materials in the Teaching-Learning-Lab is appreciated by teachers and students, the most important media in the Lab are 25 iPads. These iPads are stored in suitable tablet cases, which make it possible to load, synchronize and transport them simultaneously. The set also includes a Mac Book Pro, with which all 25 iPads can be controlled. The iPads also include Bluetooth keyboards, touch-pens and headphones. With the corresponding Apple TV box, which is connected to the beamer, one can project from any tablet chosen. The aim of acquiring the iPads was to promote digital work in the classroom.

[1] This SLS: Salzburger Lese-Screening für die Klassenstufen 1-4 ELFE II: Ein Leseverständnistest für Erst- bis Siebtklässler, Version 2; HAMLET 3-4: Hamburger Lesetest für 3. und 4. Klassen; LGVT: Lesegeschwindigkeit- und Verständnistest für die Klassen 5-12+

## 3. Report on the use of digital media

The TLL has now been in existence for about two years. First, the main focus was to procure and categorize suitable material for all teaching subjects in order to provide an attractive offer for students and teachers from various teaching subjects. Especially for the subjects Mathematics, German and English useful and beneficial materials were found quickly. We have been facing difficulties in obtaining materials for subjects not taught at the Alpen-Adria-University of Klagenfurt. The difficulty is to estimate which materials would be of a valuable use for teaching. The second task was to prepare the iPads to be used for teaching; synchronizing them and installing the updates. Problems of all kinds arose in this process. Due to the differing technical equipment available in the classrooms, a scenario once tried out, did not guarantee that the tablets would work in another room.

When the tablets were technically ready for projecting, researching and watching educational videos, weekly workshops were offered on the use of the tablets in class and on viewing materials in order to make the Teaching-Learning-Lab known to students and teachers. In addition, there were separate workshops for classes held, in which the media for each subject were presented and tried out. These workshops were very popular among students as well as teachers and the offer was accepted in many different courses. There was a low number of registrations for the weekly workshops on tablet use so only a few students participated in these. The low participation of students represented a further challenge. The two main reasons for the lack of participation provided by students were the lack of time and the inability to make it to the Teaching-Learning-Lab during opening hours. To counteract this, the Teaching-Learning-Lab now has changing opening hours and the workshops are also offered at different times. The number of students and teachers interested in the Teaching-Learning-Lab has increased since the winter term 2019/20.

Teachers were more interested in the workshops and some committed teachers used the tablets for their courses after they were presented in the workshops. Nevertheless, they also struggled with unexpected problems when using the iPads for teaching. For example, it does not seem possible to work with several tablets on one presentation without technical conflicts. If a lecturer shows a presentation from the Mac Book, students can follow the presentation on the tablet. If the task is set to edit in the presentation through the tablets, any editing from another device will change the original version. This results in many different versions that do not match. Therefore, it is not yet possible to edit a presentation together without technical conflicts arising.

The potential for the use of tablets has been continuously expanded and more and more university teachers are now focusing on teaching with digital media. However, the problem has always remained the same. The enormous effort necessary for preparing, wrapping up and during the lecture stands in contrast to an equal or even lower learning growth among the learners. In addition, the actual time on task for the students is decreased on the cost of unpredicted technical difficulties. These turn out as time consuming since the group has to move to another room. Furthermore, sometimes it turns out that the booked classroom does not dispose appropriate technical infrastructure, which requires an enormous flexibility of the teacher who is forced to make a quick adjustment of the lesson plan.

The tablets offer new and interesting possibilities, including the creation of new learning scenarios, and hence, they can improve learning enormously. However, these possibilities cannot yet be fully exploited because a lot of technical knowledge and time would be required for the preparation. Consequently, it can be stated that the amount of time invested can be seen as a negative side effect. Nevertheless, by constantly working with the tablets, a number of possibilities have been found to use tablets in class in a meaningful and beneficial way, thus creating a balance between preparation time and learning gains. Digital tools, which can be created online in a simple way, make learning and teaching with digital media fast and easy, and learning friendly. A positive aspect is that these digital tools are usually free of charge and easily accessible for everyone.

The use of digital media should not be limited to accelerating and more efficiently achieving teaching goals, but should "significantly expand the target dimensions of teaching" (Krommer, 2020, 21). Teachers often complain that preparing lessons with digital media/tools takes more time than preparing lessons with traditional teaching media. The additional positive effect that digital media and tools can have on learners is not recognized. For the Teaching-Learning-Lab, this means first to offer constant technical support to teachers and students in order to keep the additional effort - caused by the use of tools and digital end devices - as low as possible. Second, it means to test innovative and relevant tools in advance in order to detect possible difficulties in certain environments and to determine the teaching suitability of these tools. Third, it means to point out methodologic possibilities offered by tools. Often, their functions are presented and analysed on a superficial level, whereas a link to the methodology of the subjects is missing. This link is essential, however, as there are subject-specific differences in the use of the tools which opens up further possibilities for use. Fourth, it means to encourage teachers and students to critically reflect on the fact that the main advantage of digital tools does not always include making teaching more efficient and quick. Fifth, it means to motivate, support and encourage teachers to use digital devices.

The Teaching-Learning-Lab offers — after a few small teething troubles — a wide range of services that can be used by students, teachers and schools. One goal is now to familiarize students and teachers with the available materials. They should get an impression of the benefits and value that can be created by using the materials. The main task of the Teaching-Learning-Lab will be to create technical and digital framework conditions that enable people with different levels of previous knowledge to use digital media in the best possible way. There will also be an increased focus on assistance in preparation, implementation and subsequent reflection in working with digital media. Consequently, the Teaching-Learning-Lab will offer a wide range of support to draw even more attention to the diverse range of digital media. The goal is to reduce the additional effort in the preparation and implementation of teaching units with digital media and to optimize the resulting value in learning outcomes.

## 4. Conclusion

To conclude, after about two years of Teaching-Learning-Lab at the University of Klagenfurt, it can be said that despite initial difficulties and enormous efforts to inform teachers and students about the offer, the work is now beginning to bear fruit. The technical obstacles, especially in the work with the tablets, are being overcome. The number of users of the premises is constantly rising. Additionally, the offers are being adapted to suit the needs. As said at the beginning, it is important to convince teachers of benefit of digital media and their benefit (especially in language learning) in order to be able to offer students a wide range of methods. However, some offers and teaching-learning workshops remain unexploited. For example, there is room for improvement in the use of technical conditions of class observation in the premises. Here, it is conceivable to analyse one's own teaching sequences with the help of the recording in order to subsequently adapt one's own pedagogical and didactic behaviour. Nevertheless, students began to spread the news on the campus about materials and the possibility to borrow these for different purposes. Additionally, the teachers have started to include tablets in their lessons. Students and teachers alike evaluate online tools as well as applications for their classrooms.

In future, the Teaching-Learning-Lab, its technical equipment and especially digital tools and applications may face even more interest and usage due to the experienced importance of those tools during the lockdown and learning in an online environment in the COVID-19 crisis in spring 2020. Even more, the crisis showed the importance of an adaptation of humans in the educational system to digital environments and the combination of analogue and digital material.

## 5. References

Jon Altuna and Arkaitz Lareki. 2015. Analysis of the Use of Digital Technologies in Schools That Implement Different Learning Theories. *Journal of Educational Computing Research*, 53(2): 205-227. https://doi: 10.1177/0735633115597869.

Lauren M. Anstey and Gavan P.L. Watson. 2018. Rubric for eLearning Tool Evaluation. *EDUCAUSE Review*, September 10. Available online: https://er.educause.edu/articles/2018/9/a-rubric-for-evaluating-e-learning-tools-in-higher-education. Accesed on 15.09.2020

Gerhard Brandhofer, Peter Baumgartner, Martin Ebner, Nina Köberer, Christine Trültzsch-Wijnen, and Christian Wiesner. 2018. Bildung im Zeitalter der Digitalisierung. In: S. Breit; F Eder; K. Krainer; C. Schreiner; A. Seel, & C. Spiel. *Nationaler Bildungsbericht Band 2: Fokussierte Analysen und Zukunftsperspektiven für das Bildungswesen:* pages 307 – 362. https://doi:10.17888/nbb2018-2-8.

Vernon Curran, Diana L. Gustafson, Karla Simmons, Heather Lannon, Chenfang Wang, Mahyar Garmsiri, Lisa Fleet, and Lyle Wetsch. 2019. Adult learners' perceptions of selfdirected learning and digital technology usage in continuing professional education: An update for the digital age. *Journal of Adult and Continuing Education*, 25(1): pages 74 – 93. https://doi.org/10.1177/1477971419827318.

Axel Krommer. 2020. Zum ‚Mehrwert' digitaler Medien. Friedrich Jahresheft, 38: 20–21.

Wolfgang Lenhard, Alexandra Lenhard, and Wolfgang Schneider. 2020. ELFE II. Ein Leseverständnistest für Erst- und Siebtklässler – Version II. Göttingen: hogrefe (4. Auflage).

Rainer H. Lehmann, Rainer Peek, and Jan Poerschke. 2006. HAMLET 3 – 4. Hamburger Lesetest für 3. und 4. Klassen. Göttingen: hogrefe (2. Auflage).

Heinz Mayringer and Heinz Wimmer. 2003. SLS 1 – 4. Salzburger Lese-Screening für die Klassenstufen 1 – 4. Mannheim: Huber.

Wolfgang Schneider, Matthias Schlagmüller, and Marco Ennemoser. 2017. LGVT 5 – 12. Lesegeschwindigkeits- und Verständnistest für die Klassen 5 – 12+. Göttingen: hogrefe

Yao-Ting Sung, Kuo-En Chang, and Je-Ming Yang. 2015. How effective are mobile devices for language learning? A meta-analysis. *Educational Research Review*, 16:68–84. https://dx.doi.org/10.1016/j.edurev.2015.09.001.