Zbornik konference
Jezikovne tehnologije in digitalna humanistika

Proceedings of the conference on
Language Technologies & Digital Humanities

20. september – 21. september 2018
Fakulteta za elektrotehniko, Univerza v Ljubljani
Ljubljana, Slovenija

September 20th – 21st
Faculty of Electrical Engineering, University of Ljubljana
Ljubljana, Slovenia

Uredila / Edited by:
Darja Fišer, Andrej Pančur

ZBORNIK KONFERENCE
JEZIKOVNE TEHNOLOGIJE IN DIGITALNA HUMANISTIKA

PROCEEDINGS OF THE CONFERENCE ON
LANGUAGE TECHNOLOGIES & DIGITAL HUMANITIES

Uredila / Edited by: Darja Fišer, Andrej Pančur

Tehnični uredniki / *Technical editors*: Jakob Lenardič, Mihael Ojsteršek

Založil / *Published by*:
Znanstvena založba Filozofske fakultete v Ljubljani / *Ljubljana University Press, Faculty of Arts*

Izdal / Issued by:
Oddelek za prevajalstvo / *Department of Translation Studies*
Inštitut za novejšo zgodovino / *Institute of Contemporary History*

Za založbo / For the publisher:
Roman Kuhar,
dekan Filozofske fakultete / *Dean of the Faculty of Arts*

Ljubljana, 2018
First edition

Spletno mesto konference / *Conference website:*
http://www.sdjt.si/wp/dogodki/konference/jtdh-2018/

Publikacija je brezplačno dostopna na: / *Publication is available free of charge at:*
http://nl.ijs.si/jtdh18/proceedings-sl.html / *http://nl.ijs.si/jtdh18/proceedings-en.html*

CIP - Kataložni zapis o publikaciji

Narodna in univerzitetna knjižnica, Ljubljana

Predgovor k zborniku konference
“Jezikovne tehnologije in digitalna humanistika”

Z letošnjo konferenco obeležujemo 20 let od prve konference »Jezikovne tehnologije«, ki so jo 1998 v Cankarjevem domu v Ljubljani organizirali Tomaž Erjavec, Vojko Gorjanc, Jerneja Žganec Gros in Anica Rant. Takratna konferenca je bila posvečena uporabi in razvoju jezikovnih tehnologij za slovenščino, možnostim njihove uporabe in smernicam prihodnjega razvoja na tem področju. Na konferenci je bilo predstavljenih 26 prispevkov, ki so obravnavali govorne tehnologije in glasoslovje, uporabo računalnikov v prevajanju in poučevanju jezikov, računalniške korpuse, standarde zapisa jezikovnih podatkov in iskanje informacij na internetu. Konferenci je sledila okrogla miza, katere neposredni rezultat je bil ustanovitev Slovenskega društva za jezikovne tehnologije, ki je bil nato glavni pobudnik in organizator vseh nadaljnjih edicij konference. Skupaj s Centrom za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT), Fakulteto za elektrotehniko Univerze v Ljubljani ter raziskovalnima infrastrukturama CLARIN.SI in DARIAH-SI jo 20. 9. in 21. 9. 2018 na Fakulteti za elektrotehniko prireja tudi letos, že enajstič po vrsti, ki po predlanski uspešni programski širitvi na digitalno humanistiko ohranja povezovalni fokus med disciplinama, hkrati pa si prizadeva postati pomembno srečevališče raziskovalcev v regiji.

Na letošnji konferenci se tako predstavlja 47 prispevkov, od tega dva prispevka vabljenih predavateljev, 36 rednih polnih prispevkov in 5 povzetkov ter 4 študentski prispevki. Vse redne in študentske prispevke so pregledali trije recenzenti. 21 prispevkov je napisanih v slovenskem, 26 pa v angleškem jeziku. Dobra polovica avtorjev prihaja iz Slovenije, iz sosednje Hrvaške 10 %, ostali avtorji pa iz kar 19 različnih držav. Zato smo program sestavili tako, da prvi dan konference poteka v angleškem jeziku, drugi dan pa v slovenščini. Za razliko od prejšnje konference letošnje sekcije niso pararelne, kar udeležencem omogoča spremljanje vseh predavanj in predstavitev. Za ta ukrep smo se odločili prav zaradi želje po še tesnejšem sodelovanju med raziskovalci s področja jezikoslovnih tehnologij in digitalne humanistike. Istočasno smo prvič uvedli predstavitev posterjev, na kateri se jih tokrat predstavlja 9.

Urednika se iz srca zahvaljujeta vsem, ki so prispevali k uspehu konference: vabljenim predavateljem, avtorjem prispevkov, programskemu odboru za predano in natančno recenzentsko delo, organizacijskemu odboru za izvedbo konference, vodjem sekcij, da so predavanja gladko potekala, tehničnim urednikom za pripravo spletnega zbornika in sponzorjem za izkazano podporo.

Ljubljana, september 2018

Darja Fišer, Andrej Pančur

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

With this year's conference we are celebrating the 20th anniversary since the first conference »Language technologies« which took place in 1998 in Cankarjev dom, Ljubljana and was organized by Tomaž Erjavec, Vojko Gorjanc, Jerneja Žganec Gros and Anica Rant. The topics of the first conference were the development and application of language technologies for Slovene and directions for the future. 26 papers were presented, dealing with speech technologies and phonology, computer-assisted translation and teaching, corpora, encoding standards for language data and searching for information on the internet. Following the conference a round table discussion was held, the direct result of which was the establishment of the Slovenian language technologies society which has since been the main initiator and organizer of all the following editions of the conference. Together with the Centre of lagnuage resources and technologies of the University of Ljubljana (CJVT), Faculty of Electrical Engineering of the University in Ljubljana and research infrastructures CLARIN.SI and DARIAH-SI the Society is also organizing this year's conference, held on 20-21 September 2018 at the Faculty of Electrical Engineering. In its 11th installment and after a successful expansion of the conference programme to Digital Humanities in 2016, we have retained the focus on the integration of the two disciplines and at the same time aimed to position the conference as an important meeting hub for fellow researchers in the region.

This year, 47 papers will be presented, including 2 talks by invited lecturers, 36 regular full papers and 5 abstracts, and 4 student papers. All the papers were reviewed by 3 reviewers. 21 papers were submitted in Slovene and 26 in English. Over half of the authors of the accepted papers are Slovene, 10% are from Croatia and the rest of the authors come from as many as 19 different countries. This is why the conference programme was designed in such a way that the first day is international, with the talks in English while talks on the second day will be held in Slovene. As opposed to the previous edition of the conference we have opted for a single track programme so that all the participants can attend all the talks, aiming to promote and foster closer collaboration among the researchers in language technologies and digital humanities. In addition, we have also introduced a poster session with 9 posters.

The editors would like to thank everyone who has contributed to the success of this conference, especially the invited lecturers and the authors of the papers for co-creating an inspiring conference programme, the Programme Committee for their dedicated reviews, the Organizing Committee for all the organizational efforts, the ession Chairs for their smooth and efficient management of the conference programme, the Technical Editors for preparing the online proceedings and the loyal sponsors for their selfless support of our activities.

Ljubljana, September 2018

Darja Fišer, Andrej Pančur

Programski odbor / *Programme committee*

Darja Fišer, predsednica / *Chair*
Filozofska fakulteta, Univerza v Ljubljani / *Faculty of Arts, University of Ljubljana*
Institut "Jožef Stefan" / *"Jožef Stefan" Institute*

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

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

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

Iza Škrjanec
Študentska sekcija / *Student section*

Organizacijski odbor / *Organising committee*

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

Vitomir Štruc (FE)
Fakulteta za elektrotehniko, Univerza v Ljubljani / *Faculty of Electrical Engineering, University of Ljubljana*

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

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

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

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

Špela Arhar Holdt, Center za jezikovne vire in tehnologije, Univerza v Ljubljani / *Centre for language resources and technologies, University of Ljubljana*

Maja Bitenc, 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*

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

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 Philosophy, 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*

Koraljka Kuzman Šlogar, Inštitut za etnologijo in folkloristiko, Hrvaška / *Institute of Ethnology and Folklore Research, Croatia*

Iztok Kosem, Filozofska fakulteta, Univerza v Ljubljani in Institut "Jožef Stefan" / *Faculty of Arts, University of Ljubljana and "Jožef Stefan" Institute*

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, Institut "Jožef Stefan" / *Jožef Stefan Institute*

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

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

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

Maja Miličević, Filološka fakulteta, Univerza v Beogradu / *Filološka fakulteta, Univerza v Beogradu*

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

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

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

Miha Peče, Inštitut za slovensko narodopisje ZRC SAZU / *Institute of Slovenian Ethnology, ZRC SAZU*

Karmen Pižorn, Pedagoška Fakulteta, Univerza v Ljubljani / *Faculty of Education, University of Ljubljana*

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

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

Tanja Samardžić, Univerza v Zurichu / *University of Zurich*

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

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

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

Kristina Štrkalj Despot, Inštitut za hrvaški jezik in jezikoslovje, Hrvaška / *Institute of Croatian Language and Linguistics, Croatia*

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

Jan Šnajder, Fakulteta za elektrotehniko in računalništvo, Univerza v Zagrebu / Faculty of Electrical Engineering and Computing, University of Zagreb

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

Simon Šuster, Raziskovalni center za računalniško jezikoslovje in psiholingvistiko, Univerza v Antwerpnu / *Computational Linguistics & Psycholinguistics Research Center, University of Antwerp*

Toma Tasovac, Beograjski center za digitalno humanistiko in Trinity College Dublin / *Belgrade Center for Digital Humanities and Trinity College Dublin*

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

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

Organizatorji / *Organizers*

SDJT

cjvt

CLARIN.SI

DARIAH-SI

Univerza *v Ljubljani*
Fakulteta *za računalništvo*
in informatiko

Sponzorji / *Sponsors*

Urnik / *Timetable*

Četrtek, 20. 9. 2018 / Thursday 20-9-2018

08.00-08.30	**Registracija / Registration**
08.30-09.00	**Otvoritev / Opening**
09.00-10.00	**Vabljeno predavanje / Invited lecture** Malvina Nissim, ***Too good to be true: Current approaches to author profiling***
10.00-10.30	**Sekcija 1 / Session 1: Strojno prevajanje / Machine translation** Gregor Donaj and Mirjam S. Maučec: *From statistical machine translation to translation with neural networks for the Slovene-English language pair* Mihael Arčan: *A comparison of Statistical and Neural Machine Translation for Slovene, Serbian and Croatian*
10.30-11.00	**Odmor za kavo / Coffee break**
11.00-12.30	**Sekcija 2 / Session 2: Jezikovni viri / Language resources** Darinka Verdonik: *Corpus and database GOS Videolectures* Maria Jose Bocorny Finatto, Paulo Quaresma and Maria Filomena Gonçalves: *Portuguese Corpora of the 18th century: old Medicine texts for teaching and research activities* Nikola Ljubešić, Željko Agić, Filip Klubička, Vuk Batanović and Tomaž Erjavec: *hr500k -- A Reference Training Corpus of Croatian* Nikola Ljubešić, Darja Fišer, Tomaž Erjavec and Filip Dobranić: *The Parlameter corpus of contemporary Slovene parliamentary proceedings* Vuk Batanović, Nikola Ljubešić and Tanja Samardžić: *SETimes.SR - A Reference Training Corpus of Serbian* Polona Gantar, Kristina Štrkalj Despot, Simon Krek and Nikola Ljubešić: *Towards Semantic Role Labeling in Slovene and Croatian*
12.30-13.30	**Odmor za kosilo / Lunch break**
13.30-14.30	**Posterji + odmor za kavo / Poster session + coffee break** Ajda Pretnar and Dan Podjed: *Data Mining Workspace Sensors: A New Approach to Anthropology* Petar Božović, Tomaž Erjavec, Jörg Tiedemann, Nikola Ljubešić and Vojko Gorjanc: *Opus-MontenegrinSubs 1.0: First English - Montenegrin parallel corpus* Sara Ries: *Online database in Research of Correspondence of Franjo Ksaver Kuhač (1834-1911)*

Eneja Osrajnik, Darja Fišer and Vojko Gorjanc:
Korpusna analiza nestandardne vejice po uvajalnih prislovnih zvezah v slovenskih formalnih in neformalnih besedilih

Christof Schöch, Maciej Eder, Carolin Odebrecht, Mike Kestemont, Antonija Primorac, Justin Tonra, Katja Mihurko Poniž and Catherine Kanellopoulou:
Distant Reading for European Literary History. A COST Action

Katja Mihurko Poniž, Amelia Sanz, Marie Nedregotten Sørbø, Suzan van Dijk, Viola Parente-Čapková, Narvika Bovcon and Aleš Vaupotič:
Teaching women writers with NEWW Virtual Research Environment

Nikola Ljubešić, Tomaž Erjavec and Darja Fišer:
KAS-term and KAS-biterm: Datasets and baselines for monolingual and bilingual terminology extraction from academic writing

Benedikt Perak and Filip Rodik:
Building a corpus of the Croatian parliamentary debates using ReLDI API for tokenization, lemmatization, syntactic parsing and Neo4j graph database for creation of social ontology model, text classification and extraction of semantic information

Damjan Popič and Darja Fišer:
Odnosi do jezika v slovenski, hrvaški in srbski računalniško posredovani komunikaciji

14.30-16.00 **Sekcija 3 / Session 3: Digitalna humanistika / Digital humanities**

Ivanka Rajh and Siniša Runjaić:
Crowdsourcing terminology: harnessing the potential of translator’s glossaries

Dan Podjed and Ajda Pretnar:
Self-Promotion on Instagram: A Case of President's Profile

Lana Hudeček and Milica Mihaljević:
One year later - What is different?

Darja Fišer and Monika Kalin Golob:
A corpus analysis of tweets of Slovene corporate users

Tobias Weber and Jeremy Bradley:
Exploring Finno-Ugric linguistics through solving IT problems

Narvika Bovcon and Aleš Vaupotič:
Artistic Visualizations and Beyond: A Study of Materializations of a Digital Database

16.00-17.30 **Sekcija 4 / Session 4: Jezikovne tehnologije / Language technologies**

Milan van Lange and Ralf Futselaar:
Debating Evil: Parliamentary Discussions of Punishment of War Criminals in The Netherlands 1945-1975

Tatjana Marvin, Jure Derganc, Samo Beguš and Saba Battelino:
Word Selection in the Slovenian Sentence Matrix Test for Speech Audiometry
Aleksander Ključevšek, Simon Krek and Marko Robnik-Šikonja:
Efficient calculation of frequency statistics for Slovene language corpora

Kaja Dobrovoljc:
N-gram Frequency Lists for Reference Corpora of Slovenian Language

	Tadej Škvorc, Simon Krek, Senja Pollak, Spela Arhar Holdt and Marko Robnik-Šikonja: *Evaluation of Statistical Readability Measures on Slovene texts* Aniko Kovač and Maja Marković: *A Rule-Based Syllabifier for Serbian*
17.30-19.00	**Občni zbor + kava / General assembly + coffee**
20.00-22.00	**Večerja / Dinner**

Petek, 21. 9. 2018 / Friday 21-9-2018

08.30-09.00	**Registracija / Registration**
09.00-10.00	**Vabljeno predavanje / Invited lecture** Martijn Kleppe, ***Bringing Digital Humanities to the wider public: libraries as incubator for DH research results***
10.00-10.30	**Sekcija 5 / Session 5: Infrastruktura / Infrastructure** Maja Dolinar, Janez Štebe and Sonja Bezjak: *Razvoj smernic za predajo in arhiviranje kvalitativnih podatkov v Arhivu družboslovnih podatkov* Darja Fišer, Jakob Lenardič and Tomaž Erjavec: *Citiranje jezikoslovnih podatkov v slovenskih znanstvenih objavah: stanje in priporočila*
10.30-11.00	**Odmor za kavo / Coffee break**
11.00-12.30	**Sekcija 6 / Session 6: Korpusno jezikoslovje / Corpora linguistics** Nataša Logar and Tomaž Erjavec: *Strokovno-znanstvena slovenščina: besednovrstne in oblikoskladenjske značilnosti* Peter Holozan: *Zbirka primerov rabe vejice Vejica 1.3* Miha Pavlovič and Rena Ito: *Korpusna analiza slovničnih napak v spisih učencev japonščine na osnovni ravni* Helena Dobrovoljc and Urška Vranjek Ošlak: *Zakaj ne z eno poizvedbo po različnih korpusih? (Troje korpusnih preverb pod primerjalnim drobnogledom)* Iztok Kosem, Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej and Cyprian Laskowski: *Kolokacijski slovar sodobne slovenščine in Baza kolokacij sodobne slovenščine* Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Taja Kuzman and Teja Kavčič: *Glagolske večbesedne enote v učnem korpusu ssj500k 2.1*

12.30-13.30	**Odmor za kosilo / Lunch break**
13.30-14.30	**Študentska sekcija + odmor za kavo / Student session + Coffee break** Urška Bratoš: *Gradnja korpusa tvitov slovenskih politikov Janes-TwePo* Isolde Van Dorst: *You, thou and thee: A statistical analysis of Shakespeare's use of pronominal address terms* Gabi Rolih: *K-means Clustering of CMC Data for Tagger Improvement* Klara Eva Kukovičič: *Uporabnost luščilnikov terminologije Sketch Engine in CollTerm z vidika (študenta) prevajalca*
14.30-15.30	**Sekcija 7 / Session 7: Digitalna humanistika / Digital humanities** Andrej Pančur: *Trajnost digitalnih izdaj: Uporaba statičnih spletnih strani na portalu Zgodovina Slovenije - SIstory* Andrej Pančur, Alenka Pirman and Maruša Kocjančič: *Spregledana kulturna dediščina in uporaba digitalne raziskovalne infrastrukture za humanistiko v raziskavi Odlivanje smrti* Alenka Kavčič, Ivan Lovrić and Vera Smole: *Interaktivna karta slovenskih narečnih besedil* Nina Ditmajer, Matija Ogrin and Tomaž Erjavec: *Zapis in prikaz starejših pesniških besedil ter njihovih variant v TEI*
15.30-16.00	**Zaključek / Closing**

Kazalo / Table of Contents

Too good to be true: Current approaches to author profiling

Malvina Nissim*
* Center for Language and Cognition, Faculty of Arts, University of Groningen

Abstract

State-of-the-art performance on author profiling (e.g., gender and age) for English in social media is around 80%, or even higher. I will present systems that yield those results, but I will also question the reliability of such figures. Specifically, I will show that the same reasons why these models work are also the cause of serious performance drops when we change, even moderately, domain or topic. If we create conditions to bypass such powerful but at the same time limiting clues, we might be able to identify features that are indeed more relevant for profiling, and we will build more portable systems. I will describe experiments in this direction, exploring whether portability can be stretched even into a cross-language setting. I will also compare these flexible models to human performance in quite an interesting way!

Bio

Malvina Nissim is Associate Professor in Language Technology at the University of Groningen, The Netherlands. She has extensive experience in sentiment analysis and author identification and profiling, as well as in modelling the interplay of lexical semantics and pragmatics, especially regarding the computational treatment of figurative language and modality. She is the author of 100+ publications in international venues, is member of the main associations in the field, and annually reviews for the major conferences and journals. She has recently co-chaired the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), and was the general chair of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM 2018). She is also active in the field of resource and system evaluation, as both organiser and participant of shared tasks, and is interested in the philosophy behind them. She graduated in Linguistics from the University of Pisa, and obtained her PhD in Linguistics from the University of Pavia. Before joining the University of Groningen, she was a tenured researcher at the University of Bologna (2006-2014), and a post-doc at the Institute for Cognitive Science and Technology of the National Research Council in Rome (2006) and at the University of Edinburgh (2001-2005). In 2017, she was elected as the 2016 University of Groningen Lecturer of the Year.

Bringing Digital Humanities to the wider public: libraries as incubator for DH research results

Martijn Kleppe*
* National Library of the Netherlands

Abstract

Digital Humanities researchers rely on large digital datasets. Since the National Library of the Netherlands (KB) has been digitizing its collection for about ten years, their datasets are popular amongst DH scholars that focus on historical newspapers, periodicals and books. By not only supporting researchers by giving them access to datasets, but also by collaborating with them, the KB aims to incorporate DH research results in their services and products. We do this by sharing our prototype tools and code on our online Lab , invite academics to come and work as researcher-in-residence and are full partner in research projects. In this talk I will describe the challenges and opportunities for libraries and academics when they collaborate. What can researchers gain from collaborating with libraries? And how can libraries bring the affordances of DH research to the wider public?

Bio

Martijn Kleppe is Head of the Research Department of the National Library of the Netherlands (KB). Trained as historian, he wrote a dissertation on photographic iconic images by building and applying computational techniques. Before moving to the KB, he was a researcher in several Digital Humanities that focused on opening up audiovisual and textual archives by using techniques from the National Language Processing Domain, speech recognition and computer vision. At the KB, he now leads the Research Department that covers topics such as digital preservation, copyright, public library research, digital scholarship and improving the usability and discoverability of digital content.

A Comparison of Statistical and Neural Machine Translation for Slovene, Serbian and Croatian

Mihael Arčan

Data Science Institute
National University of Ireland Galway
IDA Business Park, Lower Dangan, Galway
mihael.arcan@insight-centre.org

Abstract

In this paper we present a comparison of translation quality using of Statistical Machine Translation (SMT) and Neural Machine Translation (NMT), considering translation directions between English, Slovene, Serbian and Croatian. Our experiments show that on a reduced training dataset with around two million sentences, SMT outperforms the NMT neural models. Furthermore, we present experiments with enlarged neural architectures, using 1,000 nodes and 4 hidden layers, which shows improved translation quality in terms of the BLEU metric.

1. Introduction

Although automatically generated translations using machine translation approaches are far from perfect, studies have shown significant productivity gains when human translators are supported by machine translation output rather than starting a translation task from scratch (Federico et al., 2012; Läubli et al., 2013; Green et al., 2013).

Due to the large success of NMT in recent years (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2014; Sutskever et al., 2014), we evaluate its translation performance against the usage of SMT, focusing on translation direction between English, Slovene, Serbian and Croatian. Figure 1 illustrates how a sequence-to-sequence neural network used in our experiments can be trained on parallel data. First, a sequence-to-sequence framework reads a source sentence using an encoder to build a dense vector, a sequence of non-zero values that represents the meaning of the source sentence. A decoder processes this vector to predict a translation of the input sentence. In this manner, these encoder-decoder models can capture long-range dependencies in languages, e.g., gender agreements or syntax structures. The challenges involved with the less supported Slavic languages (Krek, 2012) lie in the morphological complexity for all word classes. Furthermore, these languages have rather a free word order and are highly inflected. There are six distinct cases affecting not only common nouns but also proper nouns as well as pronouns, adjectives and some numbers. Some nouns and adjectives have two distinct plural forms depending on the number. There are also three genders for the nouns, pronouns, adjectives and some numbers leading to differences between the cases and also between the verb participles for past tense and passive voice.

Since training a neural translation model is computational expensive, we first limit the data used to train the NMT models to two million parallel sentences. In the next experiment, we then extend the parallel corpus for the English-Slovene language pair to evaluate how the neural architecture, number of nodes and hidden layers, affect the translation quality.

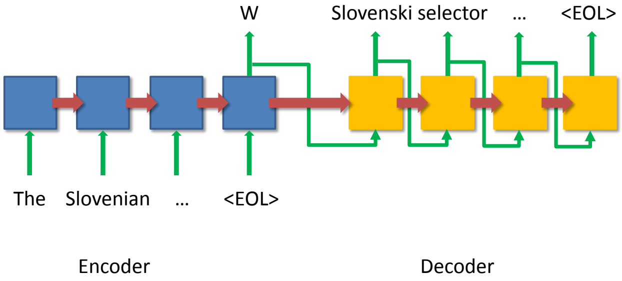


Figure 1: Neural network with the encoder-decoder architecture.

Finally, the neural models trained during this work are publicly accessible through the *Asistent* system[1] (Arčan et al., 2016), an SMT system, which enables automatic translations between English, Slovene, Croatian and Serbian language.

The remainder of the paper is organised as follows: Section 2. gives an overview of the related work on machine translation for the targeted south Slavic languages. Section 3. describes the methods of reducing the number of parallel sentences and subword unit transformation. After this, we give insights on the used parallel resources, translation frameworks and evaluation methods in Section 4. In Section 5. the results of our experiments described in the previous section are illustrated. Finally, we conclude our findings and give an outlook for our further research.

2. Related Work

One of the first results with automatic translations for Slovene was shown in the *Presis* System (Romih and Holozan, 2002). The rule-based translation system annotates each source sentence with grammatical features and uses built-in rules for converting annotated source sentences into the target language.

First publications dealing with SMT systems for

[1] `http://server1.nlp.insight-centre.org/asistent/`

Serbian-English (Popović et al., 2005) and Slovene-English (Maučec et al., 2006) are reporting results using small bilingual corpora. Using morpho-syntactic knowledge for the Slovene-English language pair was shown to be useful for both translation directions in Žganec Gros and Gruden (2007). However, no analysis of results has been carried out in terms of what actual problems were caused by the rich morphology and which of those were solved by the morphological preprocessing. Recent work in SMT also deals with the Croatian language, which is very closely related to Serbian. First results for Croatian-English are reported in Ljubešić et al. (2010) on a small weather forecast corpus, and an SMT system for the tourist domain is presented in Toral et al. (2014). Furthermore, SMT systems for both Serbian and Croatian are described in Popović and Ljubešić (2014) and more recently in Toral et al. (2016) and Sánchez-Cartagena et al. (2016). Work on rule based machine translation between Croatian and Serbian was shown in Klubička et al. (2016).

Different SMT systems for subtitles were developed in the framework of the SUMAT project, including Serbian and Slovene (Etchegoyhen et al., 2014). First effort in the direction of collecting a larger amount of existing parallel datasets for Serbian and Slovene was carried out in Popović and Arcan (2015). The authors built several SMT systems in order to identify the most important language related issues which may help to build better translation systems. However, all the translation systems described were built and used only locally, mainly only on one particular genre and/or domain. In this proposed work, we are building a publicly available mixed-domain SMT system built on existing parallel corpora, which we believe will be useful for the given under-resourced language pairs.

Popovic et al. (2016a) perform a systematic evaluation of MT results between Croatian, Serbian and Slovenian on the differences between the structural properties represent the most prominent issue for all translation directions between the Slavic languages. For translations between Croatian and Serbian, the constructions involving the verb *trebati* (en. should/need) definitely represent the larger obstacle for both translation directions and for both MT approaches, statistical as well as rule-based.

Maučec and Brest (2017) present an overview of numerous relevant works and the main issues on SMT of highly inflectional Slavic languages. The authors give insights on the most difficulties related to inflectional richness and relaxed word order. Furthermore they stress big differences between translation from a highly inflectional language and translation to a highly inflectional language. The research has shown that simple reduction of rich morphology of Slavic language does not improve translation to English, because some important information is also lost. Translation to a highly inflectional language poses a question about morphological features of target words as they are not evident from morphologically less rich source language. In this sense taking source context in account and additional tagging of source text, based on target language, shows promising results. Manojlović et al. (2017) investigate the treatment of idioms in state-of-the art SMT systems involving English and Croatian. The authors construct three short stories abundant with idioms per each language, and translate them into the other language by two state-of-the-art SMT systems. They manually inspect the outputs and present results and devise an error taxonomy for handling idioms. Popović et al. (2016b) demonstrate that a small amount of in-domain training data is very important for the translation quality for the English-to-Croatian SMT of the specific genre of Massive Open Online Courses (MooC), especially for capturing appropriate morpho-syntactic structures. Adding in-domain parallel data containing the closely related Serbian language improves the performance, especially when the Serbian part is translated into Croatian thus producing an artificial English-Croatian in-domain corpus. The improvements consist mainly from reducing the number of lexical errors. Further improvements have been achieved by adding a relatively large out-of-domain news corpus reaching performance comparable with systems trained on much larger (out-of-domain) parallel texts. The authors show that adding this corpus reduces the number of additions and lexical errors, nevertheless it introduces more morphological and ordering errors due to the different nature and structure of the segments.

3. Methodology

In this section, we describe the data selection approach of finding relevant sentences within parallel data. Due to the large vocabulary of morphological rich languages, we use subword unit NMT models instead of word-based models and give therefore insights into *Byte Pair Encoding* to minimise the out-of-vocabulary (OOV) issue.

3.1. Relevant Sentence Selection

Due to the computational complexity of training NMT sequence-to-sequence models, we experiment on minimising the set of relevant parallel sentences. Therefore, we start selecting parallel sentences for each language pair from the set of all sentences and select those containing words, which do not appear in the set of previously selected sentences. We repeat this step till we obtain a corpus with the targeted size. With this approach, we plan to exclude duplicate sentences and minimise the set of very similar sentences into the training dataset. For this initial setting, we perform the selection approach for all language pairs limiting the training dataset to two million sentences. Furthermore, focusing only on the English-Slovene language pair, we also generate a parallel corpus with five million sentences.

3.2. Byte Pair Encoding

A common problem in machine translation, in general, are rare and unknown words, e.g. terminological expressions, which the system has rarely or never seen. Therefore, if the training method does not see a specific word or phrase multiple times during training, it will not learn the correct translation. This challenge is even more evident in NMT due to the complexity associated with neural networks. Therefore the vocabulary is often limited only to 50,000 or 100,000 words (in comparison to 200,000 or more words in a two million corpus). To overcome this

Sentence	Tokenised Sentence	BPE Segmented Sentence
procedures for proper handling and disposal of antineoplastic medicinal products should be used.	procedures for proper handling and disposal of antineoplastic medicinal products should be used .	procedures for proper handling and disposal of ant■ ine■ op■ lastic medicinal products should be used ■.
many thanks to mrs van lancker, mr berman, mr lambsdorff, mr hutchinson, mrs scheele, mrs doyle, mrs weber, mr varvitsiotis, mrs hassi and mrs gomes.	many thanks to mrs van lancker , mr berman , mr lambsdorff , mr hutchinson , mrs scheele , mrs doyle , mrs weber , mr varvitsiotis , mrs hassi and mrs gomes .	many thanks to mrs van lanc ■ ker ■, mr berman ■, mr lamb■ s■ dor■ ff ■, mr hut■ chin■ son ■, mrs sche■ ele ■, mrs doyle ■, mrs weber ■, mr var■ vit■ si■ otis ■, mrs hass■ i and mrs g■ omes ■.
kdo vam daje pravico, da invalidom odrekate neomejen dostop do izobraževanja, ali da starejšim ljudem odrekate enako obravnavo pri zavarovanjih in finančnih storitvah?	kdo vam daje pravico , da invalidom odrekate neomejen dostop do izobraževanja , ali da starejšim ljudem odrekate enako obravnavo pri zavarovanjih in finančnih storitvah ?	kdo vam daje pravico ■, da invali■ dom odre■ kate ne■ omejen dostop do izobraževanja ■, ali da stare■ jšim ljudem odre■ kate enako obravnavo pri zavarov■ anjih in finančnih storitvah ■?

Table 1: Examples of tokenised sentences and subword unit (BPE) segmentation.

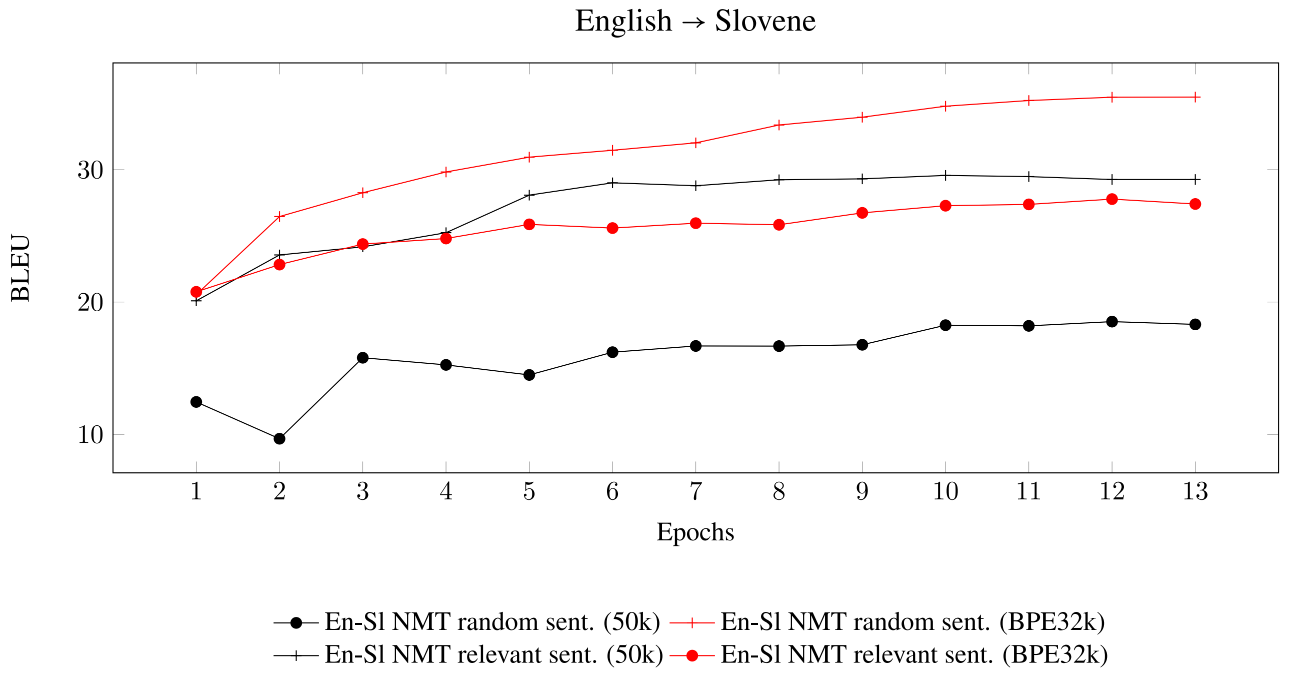


Figure 2: Comparison of the evaluation dataset with NMT systems using word level and subword units.

limitation, different methods were suggested, i.e. character based NMT (Costa-Jussà and Fonollosa, 2016; Ling et al., 2015) or using subword units, e.g. Byte Pair Encoding (BPE). The latter one was successfully adapted for word segmentation specifically for the NMT scenario Sennrich et al. (2015). BPE (Gage, 1994) is a form of data compression that iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte. Instead of merging frequent pairs of bytes as shown in the original algorithm, characters or character sequences are merged for the purposes of NMT. To achieve this, the symbol vocabulary is initialised with the character vocabulary, and each word is represented as a sequence of characters, plus a special end-of-word symbol (■), which allows restoring the original tokenisation after the translation step. This process is repeated as many times as new symbols are created. Table 1 shows the differences between word-based tokenisation and subword unit (BPE) segmentation, while Figure 2 shows the translation quality improvement of subword unit models and the word-based models in terms of the BLEU metric within the span on 13 epochs using a parallel corpus with two million sentences.

4. Experimental Setting

In this Section, we give an overview on the datasets and the translation toolkits used in our experiment. Furthermore, we give insights into the evaluation techniques, considering the translation directions between English and

					L1 Language		L2 Language	
	L1 Lang.	-	L2 Lang.	Sentences	Tokens	Types	Tokens	Types
Training Dataset	English	-	Slovene	2,299,805	37,849,280	631,114	33,379,920	587,018
	English	-	Croatian	2,464,895	34,880,415	626,353	29,744,620	593,672
	English	-	Serbian	2,152,740	27,465,108	658,660	23,536,540	573,241
	Croatian	-	Serbian	2,177,242	22,717,151	1,200,577	22,790,166	1,170,801
	Slovene	-	Croatian	2,004,229	17,759,047	464,392	18,150,733	545,023
	Slovene	-	Serbian	2,131,301	20,515,466	769,019	21,257,242	883,175
Development Dataset	English	-	Slovene	2,017	38,280	32,918	10,092	13,650
	English	-	Croatian	2,114	45,605	40,536	8,264	12,638
	English	-	Serbian	2,092	37,757	35,419	10,373	14,017
	Croatian	-	Serbian	2,000	14,716	14,774	4,601	4,620
	Slovene	-	Croatian	2,000	14,339	14,447	3,924	4,215
	Slovene	-	Serbian	2,000	12,985	13,575	3,890	4,060
Evaluation Dataset	English	-	Slovene	2,015	44,559	39,561	7,414	10,972
	English	-	Croatian	2,113	45,768	40,462	8,218	12,727
	English	-	Serbian	2,036	40,349	37,346	6,833	10,866
	Croatian	-	Serbian	2,000	12,805	13,043	3,909	3,984
	Slovene	-	Croatian	2,000	13,799	14,187	3,794	4,109
	Slovene	-	Serbian	2,000	13,090	13,606	3,900	4,138

Table 2: Statistics on parallel data used for the training, development and evaluation set (tokens = running words; types = unique words).

the targeted Slavic languages.

4.1. Training Datasets

The parallel data used to train the translation systems were mostly obtained from the OPUS web site (Tiedemann, 2012), which contains various corpora, i.e. DGT, ECB, EMEA, Europarl, KDE among others, of different sizes and domains. For the Serbian-English language pair, a small language course corpus of about 3,000 sentence pairs was added as well. Furthermore, a small phrase book with about 1,000 entries was added to the Slovene-Serbian training set. From the set of the available corpora, we select relevant sentences limiting the corpus to a specific size.

Table 2 illustrates the amount of data used to train, tune and evaluate our translation models. The upper part of the table shows the number of parallel entries used to train the translation models, considering the data selection approach (cf. Subsection 3.1.). While corpora for the English-Slavic language pairs consist of different domains, e.g. legal, medical, financial, IT, parallel data between Slavic language pairs consist mostly out of the OpenSubtitles corpus (Lison and Tiedemann, 2016).[2]

4.2. Evaluation Datasets

The dataset used for evaluating the translation performance consists of around 2.000 sentences for each language pair of various domains.[3] When translating from or into English, sentences from different corpora[4] were added to the evaluation dataset (isolated from the training dataset). The data used for evaluating translations between the Slavic languages consist mostly out of the OpenSubtitles corpus since this corpus builds the largest part (≈95%) of the data used to train the translation models.

4.3. Machine Translation tools

For our SMT translation task, we use the statistical translation toolkit **Moses** (Koehn et al., 2007), where the word alignments were built with the GIZA++ toolkit (Och and Ney, 2003). The KenLM toolkit (Heafield, 2011) was used to build a 5-gram language model.

OpenNMT (Klein et al., 2017) is a generic deep learning framework mainly specialised in sequence-to-sequence (seq2seq) models covering a variety of tasks such as machine translation, summarisation, image to text, and speech recognition. We used the default OpenNMT parameters, i.e. 2 layers, 500 hidden bidirectional LSTM[5] units, input feeding enabled, batch size of 64, 0.3 dropout probability and a dynamic learning rate decay. We train the networks for 13 epochs and report the results in Section 5.

In addition to the default setting, we perform for the English-Slovene language pair experiments on larger neural networks, extending LSTM to 1,000 hidden units and increase the network to 4 layers.

4.4. Evaluation Metrics

The automatic translation evaluation is based on the correspondence between the SMT output and reference translation (gold standard). For the automatic evaluation

[2] `http://www.opensubtitles.org/`

[3] The evaluation set can be obtained under: `http://server1.nlp.insight-centre.org/asistent/data/asisten_evaluation_set.tar.gz`

[4] DGT, EMEA, Europarl, KDE and OpenSubtitles for English-Slovene; DGT, hrenWaC, KDE, OpenSubtitles and SETimes for English-Croatian; KDE, OpenSubtitles and SETimes for English-Serbian

[5] LSTM -Long Short Term Memory

	SMT			NMT		
	BLEU	Meteor	chrF	BLEU	Meteor	chrF
English → Slovene	37.35	29.92	60.74	27.41	24.83	53.02
Slovene → English	46.02	37.42	64.61	27.80	29.28	52.76
English → Croatian	32.44	28.03	57.46	20.94	21.69	49.73
Croatian → English	37.49	35.26	60.66	23.73	26.67	50.41
English → Serbian	31.30	27.33	55.28	17.44	20.24	45.89
Serbian → English	32.49	34.18	57.82	20.64	25.00	48.02
Slovene → Serbian	19.37	22.09	41.52	19.95	21.35	40.99
Serbian → Slovene	21.51	22.82	43.46	21.68	22.30	43.10
Slovene → Croatian	21.29	22.93	43.97	19.54	21.01	40.60
Croatian → Slovene	25.94	25.39	47.60	24.77	24.12	45.71
Serbian → Croatian	68.96	48.24	79.96	61.06	42.76	76.85
Croatian → Serbian	68.15	46.70	78.14	64.25	43.84	76.83

Table 3: Automatic evaluation of translation quality for all targeted language pairs using two million sentences, selected based on new vocabulary.

we used the BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and chrF (Popović, 2015) metrics.

BLEU (Bilingual Evaluation Understudy) is calculated for individual translated segments (n-grams) by comparing them with a dataset of reference translations. Those scores, between 0 and 100 (perfect match), are then averaged over the whole *evaluation dataset* to reach an estimate of the translation's overall quality.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is based on the harmonic mean of precision and recall, whereby recall is weighted higher than precision. Along with exact word (or phrase) matching it has additional features, i.e. stemming, paraphrasing and synonymy matching. In contrast to BLEU, the metric produces good correlation with human judgement at the sentence or segment level.

chrF3 is a character n-gram metric, which has shown very good correlations with human judgements on the WMT2015 shared metric task (Stanojević et al., 2015), especially when translating from English into morphologically rich(er) languages.

The approximate randomization approach in MultEval (Clark et al., 2011) is used to test whether differences among system performances are statistically significant with a p-value < 0.05.

5. Evaluation

In this Section, we report the translation quality based on the evaluation datasets generated with the SMT and NMT models. Additionally, we perform experiments extending the training data to five millions entries as well as extending the neural architecture for the English-Slovene language pair.

5.1. Translation Evaluation Based on Two Million Relevant Sentences

As a first evaluation, we automatically compare the translations generated by SMT and subword NMT models, trained on two million selected relevant sentences. As seen in Table 3, the results show a better performance of the SMT system in almost all translation directions. Only when translating from Slovene into Serbian, the NMT system performs statistically significantly ($p < 0.05$) better in comparison to the SMT generated sentences. The lower translation quality generated by the NMT system can be explained due to the relevant sentence and vocabulary selection (see Table 5), while SMT can better handle the high vocabulary density in the parallel corpus of two million sentence used for training. On the other hand, we observed that data selection can be beneficial within the SMT approach. If a corpus of two million random sentences was selected to train an SMT model, the BLEU score drops from 37.35 to 33.43, when translating from English to Slovene and from 46.02 to 40.71, when translating from Slovene into English.

	BLEU	Meteor	chrF
Slovene-Serbian	5.30	11.48	25.61
Serbian-Slovene	5.24	11.97	26.61
Slovene-Croatian	4.80	11.51	26.34
Croatian-Slovene	4.76	11.98	27.80
Serbian-Croatian	66.78	46.70	78.53
Croatian-Serbian	67.20	46.08	77.85

Table 4: Language similarities based on the BLEU, Meteor and chrF metric.

The high evaluation scores in Table 3 between Croatian and Serbian can be explained due to the language similarities between these two languages. Table 4 illustrates the similarity based on the vocabulary, as well as on the character level for the three south Slavic languages. The scores were calculated in a manner that the source text of the evaluation dataset was treated as the generated translation of the translation system and compared with the target side of the evaluation set. We observed that Slovene is expected less similar to Serbian and Croatian, whereby Serbian and Croatian show high similarity even without any translation approaches. The SMT generated translations are, neverthe-

	2M			5M		
	Sentences	Tokens	Types	Sentences	Tokens	Types
random (English)	2,000,000	25,496,017	243,512	4,980,012	61,802,915	386,260
random (Slovene)	2,000,000	21,175,388	415,965	4,980,012	51,219,305	640,471
compressed (English)	2,299,805	37,849,280	631,114	5,003,508	85,675,760	618,112
compressed (Slovene)	2,299,805	33,379,920	587,018	5,003,508	72,868,553	964,763

Table 5: Statistics on the two and five million parallel corpus used to train the English-Slovene translation systems (tokens = running words; types = unique words).

	2M			2M++			5M			5M++		
English → Slovene	BLEU	Meteor	chrF	BLEU	Meteor	chrF	BLEU	Meteor	chrF	BLEU	Meteor	chrF
random	35.49	29.47	60.52	35.26	29.27	59.52	36.60	29.94	60.33	38.01	30.75	61.54
compressed	27.41	24.83	53.02	29.30	25.75	54.51	37.56	30.57	61.75	38.75	31.19	62.58
Slovene → English	BLEU	Meteor	chrF	BLEU	Meteor	chrF	BLEU	Meteor	chrF	BLEU	Meteor	chrF
random	38.77	35.29	61.15	36.86	33.90	59.31	36.72	33.90	59.58	38.02	34.08	59.70
compressed	27.80	29.28	52.76	28.76	29.53	53.19	37.71	34.38	60.09	40.30	35.61	62.16

Table 6: Automatic translation evaluation based on different parallel corpora and network architecture.

less, statistically significantly better than a direct comparison.

5.2. Translation Evaluation on Extended Neural Networks

Due to the low performance of the subword unit NMT models, we experimented with extending the training data for the English-Slovene language pair to five million sentences. Within this experiment, we first randomly selected five million sentences and secondly identified five million relevant sentences with a high vocabulary density, as described in Section 3.1. Furthermore, we extended the LSTM neural network architecture to 1,000 nodes and use 4 layers in the network. Table 5 illustrates the vocabulary change of the datasets based on randomly selected sentences and the identification of relevant sentences based on newly seen vocabulary. As seen, the approach increases the vocabulary of the two million corpus (2M in Table 5) from around 200,000 to more than 600,000 unique words (types) for English, and from 400,000 to almost 590,000 unique words for Slovene. Similarly, we observed a vocabulary increase for the five million entries corpus (5M).

Table 6 shows the results of the automatic translation evaluation between the different corpora described in Table 5 and the network architecture. In summary, we observed that with the usage of the parallel corpus of two million sentences, the network architecture does not have a large impact on the translation quality improvement. In the case when translating from Slovene into English, the translation quality even decreases for the randomly selected parallel corpus. A comparison between randomly selected sentences and identified relevant sentences, we learned that due to the high density of the vocabulary, the network cannot store all the provided information, therefore the neural models trained on the random sample performed better than the relevant sentence corpus (relevant sentences). Due to this, we increase the corpus to five million sentences. In this setting the larger network architecture (5M++)[6] allows translation quality improvement in terms of the BLEU scores. Furthermore, the relevant sentence corpus outperforms the neural model trained on randomly selected sentences, since the neural network is large enough to handle the dense vocabulary within the training dataset.

6. Conclusion

In this paper, we compared the performance of SMT and NMT approaches between English and the morphological rich south Slavic languages, Slovene, Serbian and Croatian. Although SMT performs better on the reduced training dataset, we observed translation quality improvement can be achieved with a parallel corpus containing selected sentences (in comparison to randomly selected sentences) for the NMT approaches, if the networks are enlarged in terms of the LSTM nodes and the number of hidden neural layers.

Our ongoing work focuses further on the selection techniques to reduce the parallel resources while preserving the translation quality. Furthermore, we continue focusing on the subword unit segmentation for terminological expressions and named entities.

Acknowledgement

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight).

[6] 1,000 LSTM nodes, 4 hidden layers

7. References

Mihael Arčan, Maja Popović, and Paul Buitelaar. 2016. Asistent – A Machine Translation System for Slovene, Serbian and Croatian. In Tomaž Erjavec and Darja Fišer, editors, *Proceedings of the Conference on Language Technologies and Digital Humanities*, pages 13–20, Ljubljana, Slovenia. Academic Publishing Division of the Faculty of Arts.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability . In *Proceedings of the Association for Computational Lingustics*.

Marta R. Costa-Jussà and José A. R. Fonollosa. 2016. Character-based neural machine translation. *CoRR*, abs/1603.00810.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard Van Loenhout, Arantza Del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine Translation for Subtitling: A Large-Scale Evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC14)*, Reykjavik, Iceland, May.

Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38, February.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448. ACM.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. Seattle, October. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *CoRR*, abs/1701.02810.

Filip Klubička, Gema Ramírez-Sánchez, and Nikola Ljubešić. 2016. Collaborative development of a rule-based machine translator between Croatian and Serbian. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*, volume 4, Riga, Latvia. Baltic Journal of Modern Computing.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Stroudsburg, PA, USA.

Simon Krek. 2012. *Slovenski jezik v digitalni dobi – The Slovene Language in the Digital Age.* META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at `http://www.meta-net.eu/whitepapers`.

Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. In *Machine Translation Summit XIV*.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W. Black. 2015. Character-based neural machine translation. *CoRR*, abs/1511.04586.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Nikola Ljubešić, Petra Bago, and Damir Boras. 2010. Statistical machine translation of Croatian weather forecast: How much data do we need? In Vesna Lužar-Stiffler, Iva Jarec, and Zoran Bekić, editors, *Proceedings of the ITI 2010 32nd International Conference on Information Technology Interfaces*, pages 91–96, Zagreb. SRCE University Computing Centre.

M. Manojlović, L. Dajak, and M. B. Bakarić. 2017. Idioms in state-of-the-art croatian-english and english-croatian smt systems. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1546–1550, May.

M. Sepesy Maučec and J. Brest. 2017. Slavic languages in phrase-based statistical machine translation: a survey. *Artificial intelligence review*, ??:1–41.

Mirjam Sepesy Maučec, Janez Brest, and Zdravko Kačič. 2006. Slovenian to English Machine Translation using Corpora of Different Sizes and Morpho-syntactic Information. In *Proceedings of the 5th Language Technologies Conference*, pages 222–225, Ljubljana, Slovenia, October.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing

Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Maja Popović and Mihael Arcan. 2015. Identifying main obstacles for statistical machine translation of morphologically rich south slavic languages. In *18th Annual Conference of the European Association for Machine Translation (EAMT)*.

Maja Popović and Nikola Ljubešić. 2014. Exploring cross-language statistical machine translation for closely related South Slavic languages. In *Proceedings of the EMNLP14 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 76–84, Doha, Qatar, October.

Maja Popović, David Vilar, Hermann Ney, Slobodan Jovičić, and Zoran Šarić. 2005. Augmenting a Small Parallel Text with Morpho-syntactic Language Resources for Serbian–English Statistical Machine Translation. pages 41–48, Ann Arbor, MI, June.

Maja Popovic, Mihael Arcan, and Filip Klubicka. 2016a. Language related issues for machine translation between closely related south slavic languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@COLING 2016, Osaka, Japan, December 12, 2016*, pages 43–52.

Maja Popović, Kostadin Cholakov, Valia Kordoni, and Nikola Ljubešić. 2016b. Enlarging scarce in-domain english-croatian corpus for smt of moocs using serbian. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 97–105. The COLING 2016 Organizing Committee.

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, September. Association for Computational Linguistics.

Miro Romih and Peter Holozan. 2002. Slovensko-angleški prevajalni sistem (a Slovene-English translation system). In *Proceedings of the 3rd Language Technologies Conference (in Slovenian)*, Ljubljana, Slovenia.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, pages 256–273, Lisbon, Portugal, September.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Víctor M. Sánchez-Cartagena, Nikola Ljubešić, and Filip Klubička. 2016. Dealing with data sparseness in SMT with factored models and morphological expansion: a Case Study on Croatian. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*, volume 4, Riga, Latvia. Baltic Journal of Modern Computing.

Jörg Tiedemann. 2012. Character-based pivot translations for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 141–151, Avignon, France, April.

Antonio Toral, Raphael Rubino, Miquel Esplà-Gomis, Tommi Pirinen, Andy Way, and Gema Ramirez-Sanchez. 2014. Extrinsic Evaluation of Web-Crawlers in Machine Translation: a Case Study on Croatian–English for the Tourism Domain. In *Proceedings of the 17th Conference of the European Association for Machine Translation (EAMT)*, pages 221–224, Dubrovnik, Croatia, June.

Antonio Toral, Raphael Rubino, and Gema Ramírez-Sánchez. 2016. Re-assessing the Impact of SMT Techniques with Human Evaluation: a Case Study on English-Croatian. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)*, volume 4, Riga, Latvia. Baltic Journal of Modern Computing.

Jerneja Žganec Gros and Stanislav Gruden. 2007. The voiceTRAN machine translation system. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 07)*, pages 1521–1524, Antwerp, Belgium, August. ISCA.

SETimes.SR – A Reference Training Corpus of Serbian

Vuk Batanović,[*] Nikola Ljubešić,[†] Tanja Samardžić[‡]

[*]School of Electrical Engineering, University of Belgrade
Innovation Center, School of Electrical Engineering, University of Belgrade
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
vuk.batanovic@ic.etf.bg.ac.rs

[†]Department of Knowledge Technologies
"Jožef Stefan" Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ijs.si

[‡]Language and Space Lab
University of Zürich
Freiestrasse 16, 8032 Zürich, Switzerland
tanja.samardzic@uzh.ch

Abstract

In this paper we present SETimes.SR – a gold standard dataset for Serbian, annotated with regard to document, sentence, and token segmentation, morphosyntax, lemmas, dependency syntax, and named entities. We describe the annotation layers and provide a basic statistical overview of them, and we discuss the method of encoding them in the CoNLL and the TEI format. In addition, we compare the SETimes.SR corpus with the older SETimes.HR dataset in Croatian.

1. Introduction

Annotated corpora of Serbian are still extremely scarce, despite the fact that various linguistic resources for Serbian have been under development since the early nineties. On the other hand, considerable advancements have recently been made in NLP technologies for Croatian, a language closely related to Serbian, thanks to a series of projects that resulted in a number of richly annotated data sets. The availability of the parallel Croatian-Serbian SETimes corpus, initially compiled by Tyers and Alperen (2010) and distributed through the OPUS platform (Tiedemann, 2009), and later improved by Agić and Ljubešić (2014), presents a good opportunity for cross-linguistic annotation transfer from Croatian to Serbian.

In this paper, we present SETimes.SR, a richly annotated gold standard dataset for Serbian, developed via an extensive use of the existing Croatian data and models. The SETimes.SR corpus is annotated on the following levels: document, sentence, and token segmentation, morphosyntax, lemmas, dependency syntax, and named entities. To the best of our knowledge, this is the first publicly available corpus in Serbian that contains all the annotation layers required for a full natural language processing pipeline.

The remainder of this paper is structured as follows: in Section 2, we describe the layers of annotation included in the SETimes.SR corpus and we present a statistical overview of label distributions in each layer. In Section 3, we present the method used to encode the corpus data, and in Section 4 we compare the new SETimes.SR dataset with the older SETimes.HR corpus in Croatian (Agić and Ljubešić, 2014). Finally, in Section 5, we present our conclusions and discuss some directions of future work.

2. Corpus Description

The SETimes.SR corpus contains news stories collected from the now defunct Southeast European Times news portal and written in Serbian using the Ekavian pronunciation and the Serbian Latin script. The SETimes portal provided news in English and languages spoken in southeast Europe, and was also the source for the SETimes.HR annotated corpus in Croatian, whose content is parallel to SETimes.SR on the document level and, for the most part, on the sentence level as well.

2.1. Segmentation

SETimes.SR is segmented into 163 documents, close to four thousand sentences, and almost 87 thousand tokens. Hence, the average document length is around 24 sentences or 532 tokens, while the average sentence length is around 22 tokens. A statistical overview of the corpus is given in Table 1.

All documents are preceded by a tag indicating their name. Tokenized sentences are preceded by a tag stating their original, untokenized text, as well as a tag contain-

Item	Count
Documents	163
Sentences	3 891
Tokens	86 726
Types	17 586
Lemmas	8 619
MSDs	557

Table 1: A statistical overview of the SETimes.SR corpus

ing their numerical ID. The form of all these tags is compliant with the Universal Dependencies v2 specification[1]. Each token is also annotated with a tag indicating its spacing from the following token in the original text, making it possible to reconstruct the original texts from their tokenized forms.

2.2. Morphosyntax and Lemmas

Morphosyntax in the SETimes.SR corpus is encoded using 13 part-of-speech categories and numerous morphosyntactic attributes particular to each category. This annotation scheme was proposed in the MULTEXT East (MTE) v5 guideline draft for Bosnian[2]. The choice of this scheme set was motivated by our goal to keep the tagset as close as possible to the one applied in SETimes.HR. At the time of our morphosyntactic annotation, the most up-to-date version of the Croatian tagset was the one used for Bosnian, another of the closely related languages originating, together with Croatian and Serbian, from the former Serbo-Croatian. The only major difference between the tags used in our corpus and the Croatian specification is a tag for the synthetic future tense[3]. A list of MTE POS categories and their frequencies in the SETimes.SR corpus is given in Table 2.

The process of morphosyntactic annotation of SETimes.SR is already briefly described in (Samardžić et al., 2017) in relation to syntactic annotation. Following previous findings that models trained on Croatian data achieve very similar tagging accuracies on both Croatian and Serbian texts (Agić and Ljubešić, 2014; Ljubešić et al., 2016), we first processed the Serbian corpus with the best performing model for Croatian (Ljubešić et al., 2016). The output was then manually corrected by two expert annotators. The training set for the Croatian model included, among others, the parallel SETimes.HR data, which made the automatic annotations already very accurate.

With the rising popularity of cross-linguistic Universal Dependencies annotations, we decided to also generate POS tags in accordance with the Universal Dependencies version 2 encoding system, which consists of 17 part-of-speech categories. The UD POS tags were, for the most part, created via automatic mapping from the MTE morphosyntactic descriptors. A notable exception that had to be manually converted were the abbreviations (MTE tag Y), since the UD standard does not provide a separate POS tag for this category. The MTE-UD mapping table and code are available on the SETimes.SR GitHub repository[4]. The frequency distribution of UD POS tags in the SETimes.SR corpus is shown in Table 3.

2.3. Dependency Syntax

Syntactic dependencies are annotated according to the Universal Dependencies version 2 standard, which describes 37 syntactic relations. Among them, 33 are present in the SETimes.SR corpus, and their distribution in it is given in Table 4. As in the case of morphosyntax, the first step in syntactic annotation was processing the corpus with the most up-to-date Croatian model (Agić and Ljubešić, 2015). Again, the training data for the Croatian model consisted of the parallel SETimes.HR data, which made these initial annotations rather accurate. Manual correction was made in 14% of all syntactic edges.

MTE POS gloss	POS tag	Count	Percentage
Nouns	N	28 322	32.66%
Verbs	V	12 990	14.98%
Punctuation	Z	10 790	12.44%
Adjectives	A	9 372	10.81%
Adpositions	S	8 460	9.75%
Conjunctions	C	6 032	6.96%
Pronouns	P	4 921	5.67%
Adverbs	R	2 847	3.28%
Numerals	M	2 217	2.56%
Particles	Q	410	0.47%
Residuals	X	350	0.40%
Abbreviations	Y	15	0.02%
Interjections	I	0	0%

Table 2: MTEv5 part-of-speech tag distribution in the SETimes.SR corpus

UD POS gloss	UD POS tag	Count	Percentage
Nouns	NOUN	21 144	24.38%
Punctuation	PUNCT	10 787	12.44%
Adjectives	ADJ	10 392	11.98%
Adpositions	ADP	8 460	9.75%
Verbs	VERB	7 439	8.58%
Proper nouns	PROPN	7 188	8.29%
Auxiliary	AUX	5 551	6.40%
Subord. conj.	SCONJ	3 179	3.67%
Determiners	DET	2 901	3.34%
Coord. conj.	CCONJ	2 853	3.29%
Adverbs	ADV	2 847	3.28%
Pronouns	PRON	2 020	2.33%
Numerals	NUM	1 202	1.39%
Particles	PART	410	0.47%
Other	X	350	0.40%
Symbols	SYM	3	0.01% <
Interjections	INTJ	0	0%

Table 3: UD part-of-speech tag distribution in the SETimes.SR corpus

The process of annotation transfer and correction was described in more detail in (Samardžić et al., 2017). Since the time of that publication, the annotation was completed, validated and shared through the Universal Dependencies infrastructure.[5] The same annotation that can be downloaded as a UD treebank is included in the corpus described here.

To assess the reliability of the annotation, we have measured the inter-annotator agreement on a sample of 300 sen-

[1] http://universaldependencies.org/format.html

[2] http://nl.ijs.si/ME/V5/msd/html/

[3] Our plan is to define a single tagset for the Serbo-Croatian macro language (ISO 639-3 code *hbs*).

[4] http://github.com/vukbatanovic/SETimes.SR/

[5] http://hdl.handle.net/11234/1-2837

UD syntactic tag	Count	Percentage
punct	10 783	12.43%
case	8 537	9.84%
nmod	7 914	9.13%
amod	7 542	8.70%
obl	6 894	7.95%
nsubj	6 730	7.76%
aux	4 101	4.73%
root	3 891	4.49%
conj	3 250	3.75%
obj	3 064	3.53%
mark	2 850	3.29%
flat	2 841	3.28%
cc	2 666	3.07%
advmod	2 391	2.76%
nummod	1 798	2.07%
acl	1 692	1.95%
det	1 570	1.81%
cop	1 329	1.53%
ccomp	1 210	1.40%
compound	1 197	1.38%
parataxis	1 187	1.37%
xcomp	973	1.12%
appos	672	0.77%
advcl	636	0.73%
fixed	414	0.48%
discourse	315	0.36%
csubj	164	0.19%
orphan	79	0.09%
goeswith	19	0.02%
list	11	0.01%
dep	3	0.01% <
iobj	2	0.01% <
vocative	1	0.01% <

Table 4: UD syntactic relation distribution in the SETimes.SR corpus

s1-s100		HR1	HR2	HR3	SR
N=2275	HR1	-	93%	93%	91%
	HR2	156	-	94%	92%
Agr=92%	HR3	159	126	-	92%
	SR	194	174	179	-
s101-s200		HR1	HR2	HR3	SR
N=2194	HR1	-	94%	94%	92%
	HR2	132	-	94%	92%
Agr=93%	HR3	114	140	-	91%
	SR	168	169	187	-
s201-s300		HR1	HR2	HR3	SR
N=2246	HR1	-	94%	94%	92%
	HR2	128	-	93%	92%
Agr=93%	HR3	142	153	-	91%
	SR	178	190	197	-

Table 5: UD annotation agreement between three Croatian native speakers (HR) and one Serbian (SR). The lower sides show the number of disagreements, the upper sides the agreement scores; N=number of tokens; Agr=average agreement scores.

tences, split into three groups of 100 sentences annotated at different time points. Each of these groups was annotated by four annotators: three Croatian native speakers and one Serbian.

The agreement scores between each pair of annotators are shown in Table 5. The agreement measure we use is the proportion of identically annotated tokens (same morphosyntactic label, dependency link, and dependency label) out of all annotated tokens (the upper sides in Table 5). The overall average agreement is slightly below 93%. It is a bit higher within the group of Croatian annotators, and a bit lower between the Serbian annotator and the Croatian group. This distinction, however, is not necessarily due to linguistic differences, but rather due to the fact that the Croatian team was trained together and separately from the Serbian annotator.

2.4. Named Entities

Named entity annotations are encoded in the IOB2 format and include the following five types of entities:

- Person (PER)
- Person derivative (DERIV-PER)
- Location (LOC)
- Organization (ORG)
- Miscellaneous (MISC)

The PER, LOC, ORG, and MISC categories are standard, while the DERIV-PER tag was introduced in order to mark personal possessive adjectives, e.g. ***Darvinova*** *teorija* 'Darwin's theory'. This addition is intended to potentially improve personal data anonymization methods in Serbian. This annotation scheme was originally developed during the annotation of the Slovene ssj500k and Janes-Tag datasets[6].

Almost seven thousand named entities were encountered in SETimes.SR, or around 42 per document, which is high, but not surprising given the journalistic nature of the texts within the corpus. The distribution of named entity types in SETimes.SR is shown in Table 6, while Table 7 contains the distribution of tokens belonging to a named entity.

Named entity type	Count	Percentage
Person	1 884	27.35%
Person derivative	75	1.09%
Location	2 678	38.88%
Organization	1 953	28.35%
Miscellaneous	298	4.33%
Total	6 888	100%

Table 6: Distribution of named entities in the SETimes.SR corpus

[6] http://nl.ijs.si/janes/wp-content/uploads/2017/09/SlovenianNER-eng-v1.1.pdf

Named entity type	Token count	Percentage
Person	3 045	3.51%
Person derivative	75	0.09%
Location	3 137	3.62%
Organization	3 369	3.88%
Miscellaneous	788	0.91%
Total	10 414	12.01%

Table 7: Distribution of named entity tokens with regard to the entire SETimes.SR corpus

The annotation of named entities was performed in the online tool WebAnno. Two annotators performed independent annotations, while a third annotator curated the collisions between them.

3. Corpus Encoding and Publishing

The working version of the SETimes.SR corpus was encoded in a modified variant of the tabular CoNLL-X format (Buchholz and Marsi, 2006), which consists of the following columns:

1. ID, token index in a sentence
2. FORM, token surface form
3. LEMMA, token lemma
4. POS, part of speech according to the MULTEXT East v5 standard
5. MSD, morphosyntactic description according to the MULTEXT East v5 standard
6. MSDFEAT, morphosyntactic features according to the MULTEXT East v5 standard
7. ___, a column left blank in order to preserve formatting equivalence with the hr500k corpus (Ljubešić et al., 2018), which contains older, non-UD dependency relation tags in this position
8. UDDEPREL, dependency relation (head, label) according to the UDv2 standard
9. UPOS+FEATS, part of speech and morphological features according to the UDv2 standard
10. UDSPEC, UDv2 language specific feature tag, used to encode the spacing between tokens in the original sentence texts
11. NER, named entity annotations encoded through IOB2

The CoNLL-type format was then converted to XML according to the TEI, (Guidelines for Electronic Text Encoding and Interchange (TEI Consortium, 2017)), in order to ensure (meta-)data persistence. Apart from the automatic conversion of the text and its annotations, this also involved writing the `teiHeader` element, which gives the metadata of the corpus, containing its name, authors, license, source description, annotation vocabulary, tag usage, revision history etc.

Each sentence in the TEI encoding (`s`), as well as each token (words (`w`) and punctuation symbols (`pc`)), is assigned a unique ID, as illustrated in Figure 1. White space in the sentence is also marked-up, with `c`. The `@lemma` attribute contains the lemma of the words, while the MULTEXT-East MSD is given in the `@ana` attribute. The UD parts of speech and features are placed within the `@msd` attribute, which is an attribute newly introduced into the TEI. Note that the double pipe symbol is used to separate the universal features from the (Serbian) language specific ones. The reason why the MULTEXT-East MSDs are not given in the `@msd` attribute, as might be expected, is that while `@msd` can contain any string, the `@ana` is defined as a pointer, which MULTEXT-East MSDs can be, but UD features cannot. We explain below in more detail the functioning of TEI pointers for linguistic labels as used in the SETimes.SR corpus. Named entities are encoded in-line, by simply using the standard TEI `name` element. Within it, the `@type` attribute contains the type of the named entity.

The final layer of annotation are the UD dependencies, which are encoded in a stand-off format, using the link group (`linkGrp`) element. `linkGrp` is an element of `s` and has attributes specifying its type (here used for the

```
<s xml:id="s2">
  <name type="loc">
    <w xml:id="s2.1" lemma="Kosovo" ana="mte:Npnsn"
       msd="UposTag=PROPN|Case=Nom|Gender=Neut|Number=Sing">Kosovo</w>
  </name>
  <c> </c>
  ...
  <w xml:id="s2.10" lemma="pritužba" ana="mte:Ncfpg"
     msd="UposTag=NOUN|Case=Gen|Gender=Fem|Number=Plur||SpaceAfter=No">pritužbi</w>
  <pc xml:id="s2.11" ana="mte:Z" msd="UposTag=PUNCT">.</pc>
  <linkGrp targFunc="head argument" type="UD-SYN">
    <link ana="ud-syn:nsubj" target="#s2.3 #s2.1"/>
    <link ana="ud-syn:advmod" target="#s2.3 #s2.2"/>
    ...
  </linkGrp>
</s>
```

Figure 1: TEI encoding of a corpus sentence

Serbian: Mogu samo ***da*** *zaključim* ***da*** *nećemo postići napredak* pred Savetom.
Croatian: Mogu samo *zaključiti kako nećemo ostvariti napredak* u Vijeću.
English: I can only *conclude that we are not going to progress* in the Council.

Figure 2: Differences between Serbian and Croatian in the usage of the *da* subordinating conjunction

annotation layer label) and the ordering of the arguments of the links. It also contains the links themselves. Each link is comprised of a link label and pointers to the IDs of the link head and argument. In cases where a syntactic dependency has the (virtual) root as its head, the sentence ID is used as the ID of the head (in the example in Figure 1 that would be `#s2`).

As mentioned, the `@ana` attribute is a pointer, which usually contains a local reference to an ID (e.g. `#s2.1`) or a fully qualified URI. TEI has another option for its pointers, namely using a prefix before the ID and separated from it by a colon (e.g. `mte:Npnsn`). Such pointers are then resolved using the `prefixDef` element in the TEI header, which defines the prefixing schema used, showing how abbreviated URIs using the scheme may be expanded into full URIs. In the case of the SETimes.SR corpus all the prefixes are simply expanded to local references, which are given in the TEI header. The only exception are the MULTEXT-East MSDs, which are defined in the `back` element of the TEI document as a feature-structure giving the decomposition of the MSD into its features. It is therefore quite simple, using just the TEI encoded corpus, to move, for example, from `mte:Mdo` to `Category = Numeral, Form = digit, Type = ordinal`.

The TEI encoded corpus, which is to be regarded as the canonical version of SETimes.SR, was then automatically converted to the so-called vertical format, which is used by CQP-based concordancers, in particular by the (no)Sketch Engine (Rychlý, 2007). The vertical format is able to encode hierarchical structures (e.g. sentences and names), and token annotations (e.g. lemmas and MSDs), but not links between tokens (e.g. dependencies). To nevertheless preserve as much of this information as possible, the dependencies are annotated next to tokens, so that the argument token is annotated with the dependency label and head lemma.

Finally, the TEI, vertical and CoNLL encodings of SETimes.SR were deposited to the CLARIN.SI repository[7], where the data is available under a Creative Commons license. The corpus is also available for exploration via the CLARIN.SI noSketch Engine and KonText concordancers, to which the links are included on the CLARIN.SI repository page.

4. Comparison with SETimes.HR

Since the SETimes.HR corpus in Croatian preceded SETimes.SR and was instrumental in its creation, it is interesting to compare the two corpora and identify the similarities and differences between them. Instead of the original SETimes.HR corpus (Agić and Ljubešić, 2014), we consider the SETimes.HR portion of the hr500k corpus, since it contains both the new annotation layers, as well as updates and corrections within the original annotation layers (Ljubešić et al., 2018).

Item	SR	HR
Documents	163	163
Sentences	3 891	3 757
Tokens	86 726	83 630

Table 8: A comparison between the SETimes.SR corpus and the SETimes.HR part of the hr500k corpus

Both corpora consist of the same number of documents gathered from the same source, but the Croatian one contains fewer sentences and tokens, as shown in Table 8. Work is currently under way to insert any missing sentences from the original SETimes parallel corpus (Tyers and Alperen, 2010) into both the Serbian and the Croatian dataset, thereby reducing the sentence and token count differential between them to a minimum, enabling maximal parallelism.

Tables 9 and 10 contain comparisons of part-of-speech tag frequencies, according to the MTEv5 and the UDv2 standard, respectively. The relationship between the Serbian and the Croatian corpus frequencies for each tag is analyzed using the $\tilde{\chi}^2$ test, quantifying the probability that the difference between the observed and the expected frequencies is due to chance. We use the Phi (Φ) coefficient to measure the effect size.

The largest differences exist with regard to the conjunction category or, more specifically, subordinating conjunctions. This difference is chiefly due to the ”*da*” subordinating conjunction, which is used much more frequently in Serbian than in Croatian (SETimes.SR: 2302, SETimes.HR: 507, $\tilde{\chi}^2$ = 1099.97, p = **3.3E-241**, Φ = 0.08035). In Serbian, unlike Croatian, ”*da*” is used in complex predicates involving modal and phase verbs, as well as within a complex form of the future tense. Figure 2 presents an example of these differences between Serbian and Croatian. We also detected a significant stylistic difference regarding the frequency of pronouns, which stems from the fact that the texts in SETimes.HR employ zero anaphora more often than those in SETimes.SR.

We did not compare the frequencies of dependency relation tags between SETimes.SR and SETimes.HR since somewhat different dependency annotation guidelines were used for each corpus (e.g. the UD syntactic tag *expl* appears 971 times in SETimes.HR but is not used at all in the Serbian corpus). On the other hand, we did perform a comparison regarding the named entity annotation frequencies, but they were very similar and no statistically significant differences could be found.

[7] `http://hdl.handle.net/11356/1200`

MTE POS gloss	POS tag	SR	HR	$\tilde{\chi}^2$	p-value	Φ
Nouns	N	28 322	**28 009**	13.36	**0.00026**	0.00886
Verbs	V	12 990	**12 605**	0.29	0.59139	0.00130
Punctuation	Z	10 790	**10 778**	7.63	**0.00575**	0.00669
Adjectives	A	9 372	**9 322**	5.01	**0.02519**	0.00542
Adpositions	S	8 460	**8 282**	1.04	0.30789	0.00247
Conjunctions	C	**6 032**	4 752	116.15	**4.4E-27**	0.02611
Pronouns	P	**4 921**	4 306	22.83	**1.8E-06**	0.01158
Adverbs	R	2 847	**2 785**	0.28	0.59377	0.00129
Numerals	M	**2 217**	2 081	0.77	0.37937	0.00213
Particles	Q	410	**481**	8.38	**0.00378**	0.00702
Residuals	X	**350**	205	32.43	**1.2E-08**	0.01380
Abbreviations	Y	15	**24**	1.95	0.16304	0.00338
Interjections	I	0	0	—	—	—
Total		86 726	83 630	—	—	—

Table 9: MTEv5 part-of-speech frequency comparison between SETimes.SR and the SETimes.HR part of the hr500k corpus. Frequencies that are larger than expected and p-values below the 0.05 level are in bold.

UD POS gloss	UD POS tag	SR	HR	$\tilde{\chi}^2$	p-value	Φ
Nouns	NOUN	21 144	**20 913**	8.95	**0.00278**	0.00725
Punctuation	PUNCT	10 787	**10 774**	7.58	**0.00589**	0.00667
Adjectives	ADJ	10 392	**10 210**	2.02	0.15485	0.00345
Adpositions	ADP	8 460	**8 282**	1.04	0.30789	0.00247
Verbs	VERB	**7 439**	6 988	2.67	0.10213	0.00396
Proper nouns	PROPN	7 188	**7 119**	2.76	0.09690	0.00402
Auxiliary	AUX	5 551	**5 617**	6.88	**0.00870**	0.00636
Subordinating conjunctions	SCONJ	**3 179**	2 017	225.89	**4.7E-51**	0.03641
Determiners	DET	**2 901**	2 699	1.82	0.17748	0.00327
Coordinating conjunctions	CCONJ	**2 853**	2 735	0.04	0.83357	0.00051
Adverbs	ADV	2 847	**2 785**	0.28	0.59377	0.00129
Pronouns	PRON	**2 020**	1 607	33.75	**6.3E-09**	0.01408
Numerals	NUM	1 202	**1 173**	0.07	0.78559	0.00066
Particles	PART	410	**481**	8.38	**0.00378**	0.00702
Other	X	**350**	205	32.43	**1.2E-08**	0.01380
Symbols	SYM	3	**25**	16.53	**4.8E-05**	0.00985
Interjections	INTJ	0	0	—	—	—
Total		86 726	83 630	—	—	—

Table 10: UD part-of-speech frequency comparison between SETimes.SR and the SETimes.HR part of the hr500k corpus. Frequencies that are larger than expected and p-values below the 0.05 level are in bold.

5. Conclusion

In this paper we have presented SETimes.SR - the first publicly available gold standard corpus of Serbian annotated on the level of document, sentence, and token segmentation, morphosyntax, lemmas, dependency syntax, and named entities. We have described and given a statistical overview of each annotation layer, and presented the way in which the annotations are encoded. We have also compared the new SETimes.SR corpus with the older SETimes.HR dataset in Croatian.

We believe that the creation of SETimes.SR is an important first step in bridging the gap between Serbian and other Slavic languages, such as Czech or Slovene, for which numerous linguistic resources and tools are available. We also hope that the introduction of the SETimes.SR corpus will promote and accelerate the development of other NLP resources and tools for Serbian. In the future, we plan to continue working on the corpus by expanding it with new kinds of annotations, such as a coreference layer. We will also consider enlarging the corpus with additional data.

6. Acknowledgements

This work is supported by the Regional Linguistic Data Initiative (ReLDI) via the Swiss National Science Foundation grant 160501, and the Slovenian Research Infrastructure CLARIN.SI. Vuk Batanović is also partially supported by the III44009 research grant of the Ministry of Education, Science and Technological Development of the Republic of Serbia.

7. References

Željko Agić and Nikola Ljubešić. 2014. The SETIMES.HR Linguistically Annotated Corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC*

2014), pages 1724–1727, Reykjavik, Iceland. European Language Resources Association (ELRA).

Željko Agić and Nikola Ljubešić. 2015. Universal Dependencies for Croatian (that work for Serbian, too). In *Proceedings of the Fifth Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*, pages 1–8, Hissar, Bulgaria. Association for Computational Linguistics.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on Multilingual Dependency Parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X 2006)*, pages 149–164, New York City, NY, USA. Association for Computational Linguistics.

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 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Nikola Ljubešić, Željko Agić, Filip Klubička, Vuk Batanović, and Tomaž Erjavec. 2018. hr500k – A Reference Training Corpus of Croatian. In *Proceedings of the 2018 Language Technologies and Digital Humanities Conference (JT-DH 2018)*. Ljubljana, Slovenia.

Pavel Rychlý. 2007. Manatee/Bonito - A Modular Corpus Manager. In *Proceedings of the First Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007)*, pages 65–70, Brno, Czech Republic. Masaryk University.

Tanja Samardžić, Mirjana Starović, Željko Agić, and Nikola Ljubešić. 2017. Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, pages 39–44, Valencia, Spain. Association for Computational Linguistics.

TEI Consortium. 2017. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Jörg Tiedemann. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing (RANLP)*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Francis M. Tyers and Murat Serdar Alperen. 2010. South-East European Times: A parallel corpus of Balkan languages. In *Proceedings of the LREC 2010 Workshop on Exploitation of Multilingual Resources and Tools for Central and (South) Eastern European Languages*, pages 49–53. European Language Resources Association (ELRA), Valletta, Malta.

Artistic Visualizations and Beyond: A Study of Materializations of a Digital Database

Narvika Bovcon,*[*] Aleš Vaupotič[†]

[*] Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, SI-1000 Ljubljana
narvika.bovcon@fri.uni-lj.si

[†]School of Humanities, University of Nova Gorica
Vipavska 13, SI-5000 Nova Gorica
ales.vaupotic@ung.si

Abstract

The paper discusses a case study of the integration of artistic practice in the digital humanities research. The methodology of coding meaning in visual form is explained for four different sculptures and an artist book. The approach is tied to the allegoric thinking in building emblems and miniatures in the history of art (16th century and later). The use of diagrams is integrated in the experimental projections of a future archeology. The data of digital humanities research can achieve greater visibility and address a wider cultural context when presented in collaboration with art institutions.

Introduction

The virtual research environment (VRE) of the NEWW WomenWriters Database (http://resources.huygens.knaw.nl/womenwriters) has been developed in 2016 (on the basis of the first version, http://www.womenwriters.nl/index.php/Database_WomenWriters) in the frameworks of a European HERA *Travelling TexTs* project. The VRE gathers data about European women writers, focusing on the long 19th century. There are three main entry points into the database: by women authors, by their publications and by receptions. The network graph shows the connections in the literary field by linking each author with all her written works and their receptions by other authors. The principal researcher and coordinator of the project in Slovenia is Katja Mihurko Poniž (2017) from the University of Nova Gorica.

Before (i. e. already for the first version of the database) and alongside the development of this tool (the NEWW VRE) the authors of this paper have been exploring the visualization of humanist data and the methodology of integrating the research process in the curricula of undergraduate study, promoting inter-disciplinary collaboration (2017).

However, the field of visualization is vast and extends beyond the graphs built with standard computer visualization techniques, which are based in statistics. Johanna Drucker argues for graphesis (2014), an idea that encompasses the whole of the visually produced knowledge, throughout the history and in different civilizations. The authors of this paper further their research of visualization in this direction, borrowing certain historic image meaning-making techniques and appropriating them for the contemporary and future needs and sensibilities of human communication.

Similar projects

The artistic exploration of the representational, and sculptural in particular, technical of course, possibilities of computer-controlled 3D printing and fabrication is already a well established practice. The extensive exhibition by Damien Hirst in Venice *Treasures from the Wreck of the Unbelievable* (Palazzo Grassi, Punta della Dogana, 9 Apr. – 3 Dec. 2017) has realized a comprehensive mystification about a lost ancient treasure, a mass of objects recovered from the shipwreck of *Apistos* – including »museum copies […] which imagine the works in their original, undamaged forms« – in different sizes, jewellery too. Here, the scale and the high technical standards of production are astonishing. The visitor of this museum-like exhibition is faced with a reading of history, in the same way as in any museum of e.g. ancient Mediterranean artefacts, and has to read the historical narrative, which is the art-piece. The narrative is condensed into the objects that perplex and stimulate interpretation.

Nataša Skušek has also created a jewellery like cast *Yajdess* (2014), using traditional sculptural techniques, of the Jaydess intrauterine delivery system. Here the sculptural aspect of the shape in silver clashes with the function of this object in the female productive system and in the female body. In the *Art + Science Now* monograph by Stephen Wilson (2010) which is an overview of the intersecting domain between the two »cultures« from the famous C. P. Snow lecture form 1959, the 3D printing is an information visualization technique; the pendent *Future Skeleton in the Closet* (sterling silver, patina, 2006) by Karin Beaumont, who is a researcher working on plankton in Antarctica and also a silversmith, moves visual features from the scientific research into the artistic realm, in this case the shape of marine algae *Asteromphalus*, but also »how they move, how they interact, helps« the author to create the artwork (Wilson, 2010). As in the case of *A Lake* sculpture presented later in this paper, Beaumont has put a literary fragment, a haiku, on the back of the pendant. Carlos Garaicoa approaches the 3D printing techniques more literally. A valuable 3D print is made in gold, therefore it has to be in a safe box: *Saving the Safe – HSBC Building* (21 Kt, safe box, rotating base, LED light, 2017). The HSBC bank is on the list of 50 world's safest banks, which is an ironic emphasis pointing to the use of valuable materials and their potential malleability (gold) of their 3D shape in conceptual art. The computer based prints do not have to be small or »clean«, i.e. polished and without noise: Garaicoa's project at the *documenta 14* (the EMST collection) was a *Photo-Topography* (nine digital inkjet prints on acrylic, nine black-and-white photographs, and pigment and gesso on wood mounted on aluminum, 2011), in essence a translation of grey values from photographs into a 2D area using elevation – dark is a high, bright is a low surface area.

Narvika Bovcon and Aleš Vaupotič have used computer-aided 3D prototyping a number of times. In 2003 at the *50th International Art Exhibition Venice Biennale* together with Gašper Jemec the *Friedhof Laguna Racing Team* was realized; the installation consisted of two boat-like casts produced by Seaway Yachts company, that had inside them a view into an interactive 3D virtual space. The professional execution of the shapes from the design to production established a dialogue with real yachts on the Venetian waterfront. In 2007 the *Dragonfly* 3D print was included (as a digital mesh) in the *Data Dune* virtual space in the project *If you look back, it won't be there anymore* (with Barak Reiser) and in the real gallery space (Bovcon and Vaupotič, 2009) In the 3D print *Atlas Air Tagging* (2011) the miniature silver cast held emblematic objects representing all the works from a group exhibition (Bovcon et al., 2013). In this case, the silver jewellery-like object was additionally remediated in the augmented reality dimension of the video installations at the exhibition.

Transposition of data into artistic form

The paper will present the case study of designing digital sculptures and an artist book. For the invited artists (Narvika Bovcon, Vanja Mervič, Aleš Vaupotič) the initial idea was simply: make a visualization of the NEWW Women Writers database as an art object. The confirmations of a successful translation of the digital humanities database into artistic form were the exhibitions of the products in art galleries (Layerjeva hiša in Kranj, Grad Kromberk in Nova Gorica) and in the National Museum of Slovenia, Metelkova in Ljubljana at which occasion the sculptures were awarded the jury prize of the *Salon ZDSLU 2017*. In the following subsections the production of the art pieces will be discussed.

Data sculptures

The data in the digital form, no matter how much effort and funding is invested into building these complex platforms of digital archives (e. g. at the Huygens institute in Hague, where also Europeana resides), appear vulnerable and unstable. The software and hardware for data storage are prone to obsoleteness. In the archeological perspective, only the hardest materials endured the extinguishment of civilizations and deterioration in time. The idea emerged to build sculptures that will encode the meaning from the NEWW database and become artifacts for the future archeologists. This is the same motivation, why today digitization is going on as an effort to preserve the artifacts from our cultural heritage. The artists are at the same time worried that the digital archives, so painstakingly filled with data, have no alternative copy in another material and are therefor present only in the virtual world and absent in the real world.

We decided to use the technology for building 3D digital models (Maya software) and printing them with 3D printers and finally using the prints for casting silver. These steps were realized in collaboration with Matic Močnik from the goldsmiths Zlatarstvo Močnik in Ljubljana.

The main dilemma about the humanities data stems from their complex existence. What humanities are dealing with are not data or measurable objects but phenomena that resist any kind of simplification and are therefor only imperfectly transformed into data. Interpretation is already a part of all phenomena. Furthermore, they are multidimensional in their meaning and function in inter-human exchange, variable in time, part of any individual's perception. All of these characteristics speak against the principle of reduction that is necessary for data analysis.

The basic question was, how to make meaningful visualization of the NEWW database? The material that is organized in the database underwent the reduction phase of meaning prior to our visualization, converting the phenomena into distinct units of categories, however, the visualization has to reverse this process. This can be done by individual narativizations of the archive proposed by the artists.

Slovene women writers

First we checked the categories that the NEWW database employs to describe the personal data about the authors from the history of literature. These refer to social class, education, financial aspects, marital status, whether they had children, date and place of birth and death. Although very basic, even these categories remain unknown for many women authors. The first sculpture was conceptualized to contain the available data in the categories of personal life of Slovene literary authors.

There are 16 Slovene literary authors in the database: Fanny Hausmann, Lavoslava Kersnik, Luiza Pesjak, Josipina Urbančič Toman, Leopoldina Rott Kersnik, Pavlina Pajk, Marica Nadlišek Bartol, Elvira Dolinar, Ljudmila Poljanec, Minka Govekar, Vida Jeraj, Ivanka Anžič Klemenčič, Ljudmila Prunk, Zofka Kveder, Manica Koman, Lili Novy. Each of them is represented on the timeline (horizontal) with a straight line connecting the years of her birth and death. The writers are organized chronologically along the vertical axes, together with their timelines a surface of a particular shape is formed.

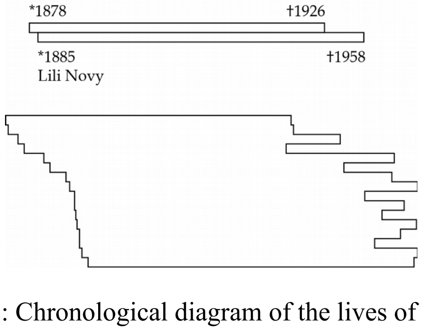


Image 1: Chronological diagram of the lives of Slovene women writers.

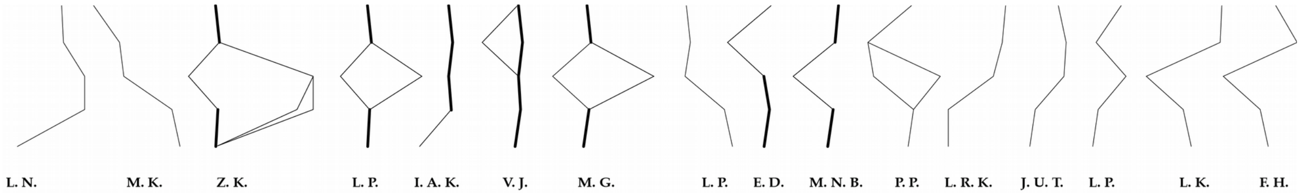


Image 4: Lifelines of Slovene women writers.

The categories are arranged horizontally as columns of different characteristics for each. The characteristics that are most common among the authors are represented with a proportionally bigger circle. The predominant characteristics in each category are arranged in a horizontal line. Thus, if an author shares these common characteristics, the line that connects them is straight and the lifeline of that author is straight. On the other hand, an author with less common characteristics has a more zig-zag lifeline, which is a reflection of her more uncommon lifestyle. Most common characteristics for Slovene women writers are: middle class, school education, unknown financial situation, married, has children.

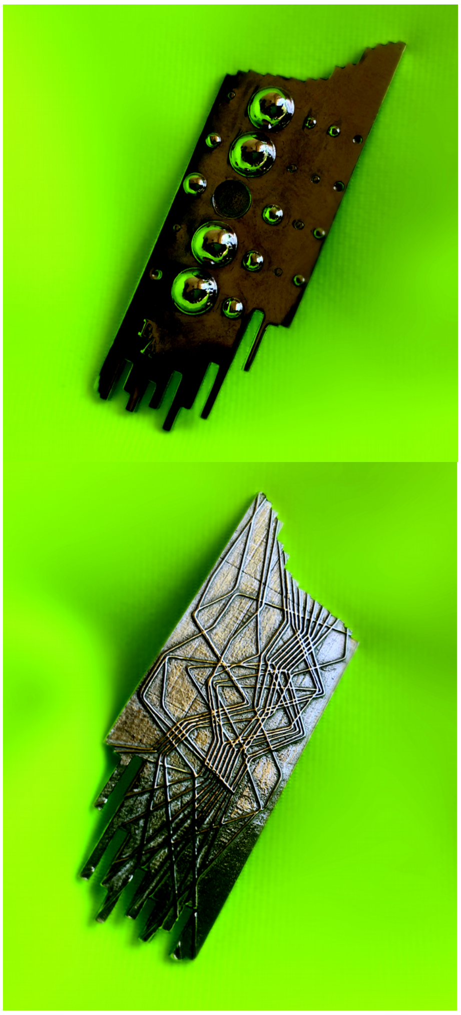


Image 5: *Slovene women writers*. The diagrams are arranged on both sides of the sculpture.
Size: 6 x 2 x 0,3 cm.

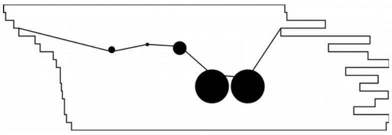


Image 2: Representation of one author:
a line connects her characteristics in each category.

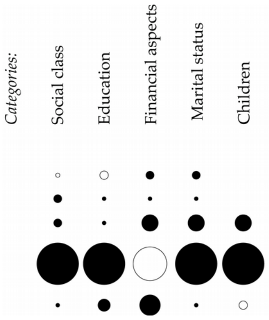


Image 3: Categories of personal and professional situation of the women writers. Circle ø represents the number of Slovene women writers with that characteristic. White circle represents unknown data for that category.

Cloth

The second sculpture is shaped according to the data about personal and professional situation of Spanish women writers. The categories are the same as presented in the previous section for Slovene women writers, however there are some characteristics that in the categories of Slovene authors don't appear, such as royal rank, or are less common, such as home education.

This time the sculpture is designed by arranging the categories of the characteristics in a square grid with the number of authors for every characteristic arranged vertically. For the 3D model of the sculpture first an appropriate 3D bar chart was made. Then, a cloth simulation was virtually dropped over it. To fit well, the cloth was tailored to roughly match the shape of the bar chart, otherwise it wasn't visible enough under the cloth. The simulation was tweaked by changing the cloth material and its deformation properties.

Here a new element enters the narration about the data on personal lives of Spanish women writers. A cloth is added by the artists as an attribute that defines women culturally and socially: it represents the fashions, the restrictions, the covering of women's bodies and faces, their domestic life and their work as housewives.

Image 6: *Cloth*. The sculpture is shaped according to the personal data about Spanish women writers.

Quotation

Another aspect of the database about women writers are their works. The artists have chosen a quotation from one of the most representative and renown novels, *Middlemarch*, written by George Eliot in 1874. The quotation summarizes the status of women at the time, their access to knowledge and emancipation: "I cannot image myself living without some opinions, but should wish to have good reasons for them, and a wise man could help me to see, which opinions had the best foundation, and would help me to live according to them."

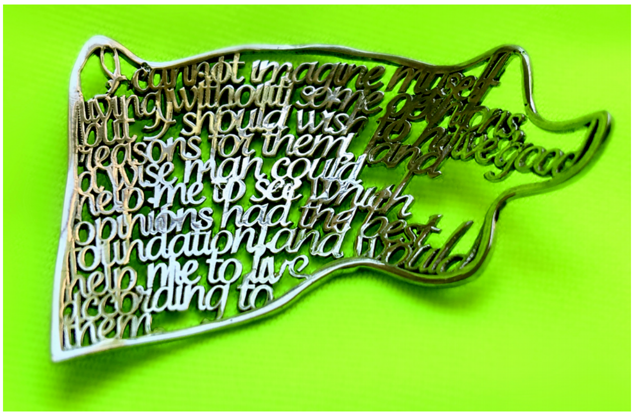


Image 7: *The Quotation*. A quote from *Middlemarch* on education for women in the Victorian time.

The text is designed (in Adobe Illustrator and Maya) as a lace merely by using typography. All the letters have to be connected to each other in order to be physically possible to 3D print this complex and very fine texture and later to cast it in silver. The final shape is curved in space to evoke a shape of a flower petal. The embroideries, the calligraphy, the lace and flowers are all symbols of women and their occupation at the time, when Dorothea from *Middlemarch* planned her marriage to get education and become able/informed how to do good in the world, as she tells us in the quotation.

A Lake

A more contemporary representative author was chosen to speak in the fourth sculpture: Sylvia Plath. Her poems and her life describe the attitude of women who are overshadowed by their unfaithful husbands. The sculpture is designed as a two-sided mirror: on the front the reflective surface is in the form of a woman's face, distorted as if reflected on water, whereas on the back the surface is a perfect mirror, in which the viewer sees her/his own portrait while reading the inscription of the Plath's verses from the poem *Mirror*. In the poem, a woman looks at her reflection in a lake and contemplates the passing of time and the changes.

Thus the four sculptures narrate the women writers database in two perspectives: from the point of view of statistical data on their lives, and from the points of view they described in their works.

Image 8: *A Lake*. Inspired by the poem *Mirror* by Sylvia Plath.

Augmented artist book À mon seul désir

Vanja Mervič conceptualized an artist book inspired by the women writers from the NEWW database. As already the title suggests, he elaborated on the idea encoded in the images of the six tapestries titled *The Lady and the Unicorn* (cca. 1500). Five of the tapestries from the series depict the five senses: sight, touch, taste, smell and hearing, while the sixth carries the inscription "À mon seul désir". George Sand wrote on the medieval series of tapestries in her novel *Jeanne*, and helped preserve them, now they are in the Cluny Museum, National museum of middle ages in Paris.

Image 9: *The Lady and the Unicorn* (cca. 1500).

Likewise, the artist book contains pages dedicated to the five senses: some pages carry images, others are apparently white. The images stimulate the sight with vivid colours, however the content of the images suggests also other senses, e. g. the photograph of a merry-go-round evokes the sound that usually accompanies it. The photograph of a chewing gum evokes the taste. Even the most colourful images are photographs of frozen pigments and the viewer looking at them senses the ice and the heat melting the ice cylinders. The images taken with the Kirlian camera make visible the aura of the artist.

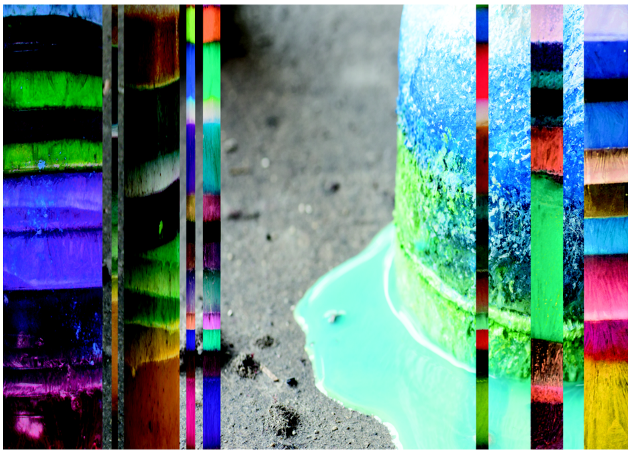


Image 10: Excerpt from an image in the artist book *À mon seul désir* (by Vanja Mervič).

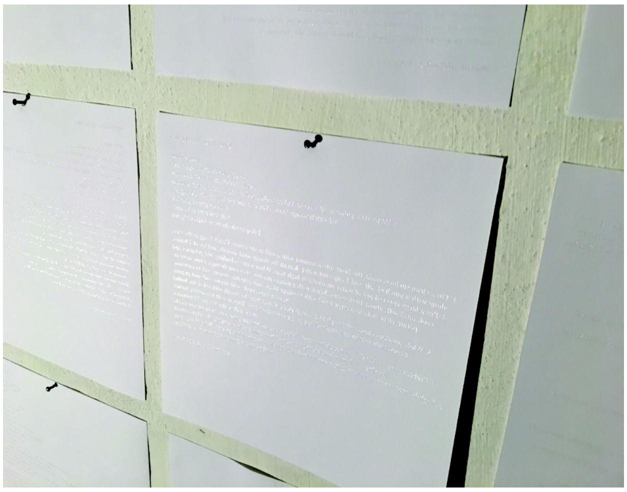


Image 11: Tactile pages from the artist book *À mon seul désir* (by Vanja Mervič).

One page is perfumed.

On the white pages the reader can touch the thick layers of transparent spot UV print insinuating the braille script, which is in fact present on one page. Looking at the UV printed pages under a certain angle of the incident ray of light, the text becomes more or less visible: cooking recipes for different dishes written by the famous women writers are printed, such as Omelette Aurore and Gruyère Tartines by George Sand, Cottage Loaf by Virginia Woolf, Chicken Pie by Charlotte Brontë, Pigeon Pie by Jane Austen, Pasta Bolognese by Anaïs Nin, Classic Tomato Soup Cake with Cream Cheese Frosting by Sylvia Plath, Ninon De L'Enclos recipe for a cream for youthful skin.

The sound is represented with a notation of the *Circus Polka: for a Young Elephant* (1942) by Igor Stravinsky, it is printed on the page adjacent to the photograph of the merry-go-round.

The artist book is carefully executed, with tactile layers of spot UV print and compiled of a variety of papers, smooth and textured, light and heavy, that feel different when the reader touches and turns the pages.

However, the pages in the book are also augmented with the augmented reality technology. If the reader scans the pages with her/his smart phone or tablet using the Layar application, on the screen of the smart handheld device additional videos, sounds and web pages appear and can be examined or browsed further. The videos were shot by the author during his artist in residence stay at the Cite internationale des arts in Paris.

The artist book with virtual reality augmented pages can contain moving media and sound, which in traditional print is not possible. New technologies can thus further extend the idea of engaging all five senses and its actual realization. Especially important for the artists is the possibility to enclose videos in a book.

Every art work has its presentation at an exhibition, which is at the same time a setting and a performance. The *À mon seul désir* artist book and the four silver 3D printed sculptures were accompanied with a culinary performance when a menu of the dishes by the famous women writers that are printed in the artist book was served. In the Mahlerca Gallery at the Layer House in Kranj *The Supper* *of This-Time Reality* was cooked by Sara Hostnik, a painter, cook and confectioner. At the Kromberk castle the catering was contributed by Domačija Lisjak.

Image 12: Exhibition at the Kromberk castle. Artist book and skin cream from the recipe of Ninon De L'Enclos.

Image 13: Classic Tomato Soup Cake with Cream Cheese Frosting by Sylvia Plath served at the Mahlerca Gallery at the Layer House in Kranj at *The Supper of This-Time Reality* cooked by Sara Hostnik.

Conclusion

The case study of the visualization of the NEWW database shows, how artists can contribute innovative visualizations and additional extensions of a digital humanities research. Data that exist in a digital form in a digital archive can be brought back into the real world in a variety of materializations. However, this is not a straightforward procedure. Not anyone, but an artist is needed who has the adequate training and skills to envision the materializations of ideas in such a way that these are not banal or boring, but carry a network of encoded meanings in elaborate artistic forms.

This is a method how to build outreach of scientific results to a wider audience and to other disciplines and fields of research or creativity. Our case study proved this method to be effective, since the visualizations of the NEWW Women Writers database have successfully entered the gallery system and even won the jury prize at the *Salon ZDSLU 2017,* which is the annual curated show of the Association of Slovene Fine Artists Societies. Consequently the artists won also the title for the person of the week at the Val 202 programme of the Slovene national radio, which has the widest possible reach and thus really many people learned about the database.

The visitors to the galleries and the listeners to the Val 202 learned about the women writers and the research on them, which otherwise would remain restricted to the literary scientists specializing on the topic and actually working with the database. At this point it is necessary to mention that the history of literature is predominantly focusing on male authors and it is an important and difficult quest in itself to establish the recognition of the women writers and their in many cases forgotten or unrecognized works. With help of the artistic objects the contents of the database were presented in a more communicative, fascinating and engaging manner and people reacted with interest both, about the artworks and about the women writers.

References

Karin Beaumont. 2006. *Oceanides – Art of the Ocean* http://www.oceanides.com.au.

Narvika Bovcon and Aleš Vaupotič. 2009. New Media Spaces and New Media Objects. *ELMAR 09.* http://eprints.fri.uni-lj.si/1392.

Narvika Bovcon, Aleš Vaupotič, Bojan Klemenc, Franc Solina. 2013. "Atlas 2012" augmented reality : a case study in the domain of fine arts. In: A. Holzinger, M. Ziefle, M. Hitz, M. Debevc, eds. *Human Factors in Computing and Informatics: First International Conference, SouthCHI.* Maribor, Slovenia.

Johanna Drucker. 2014. *Graphesis: Visual Forms of Knowledge Production.* Harward UP.

Carlos Garaicoa. 2012. Photo-topography (Cuatro Caminos). *Artsy*. https://www.artsy.net/artwork/carlos-garaicoa-foto-topografia-slash-photo-topography-cuatro-caminos.

Carlos Garaicoa. 2017. www.documenta14.de/en/artists /22260/carlos-garaicoa.

Damien Hirst. 2017. *Treasures from the Wreck of the Unbelievable. The Guide.* https://www.palazzograssi. it/site/assets/files/6176/guida_damien_hirst_eng.pdf

Katja Mihurko Poniž. 2017. *Literarna ustvarjalka v očeh druge_ga: študije o recepciji, literarnih stikih in biografskem diskurzu.* Nova Gorica: Založba Univerze v Novi Gorici.

Charles Percy Snow. 2001 [1959]. *The Two Cultures*. London: Cambridge University Press.

Aleš Vaupotič and Narvika Bovcon. 2017. Visualization of the Women Writers Database: Interdisciplinary Collaboration Experiments 2012-2015. In: Katja Mihurko Poniž, ed. *Reception of Foreign Women Writers in the Slovenian Literary Systen of the Long 19th Century*, pages 72–116. Nova Gorica, University of Nova Gorica Press.

Stephen Wilson. 2010. *Art + Science Now.* London: Thames and Hudson.

Opus-MontenegrinSubs 1.0: First electronic corpus of the Montenegrin language

Petar Božović,* Tomaž Erjavec,† Jörg Tiedemann,‡ Nikola Ljubešić,§ Vojko Gorjanc‡

* English Language and Literature Department, Faculty of Philology, University of Montenegro
Danila Bojovića bb, 81400 Nikšić
petarb@ac.me

†Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
tomaz.erjavec@ijs.si
‡Department of Linguistics and Philology, Uppsala University
BOX 635, SE-751 26 Uppsala, Sweden
jorg.tiedemann@lingfil.uu.se
§Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ffzg.hr
‡Department of Translation Studies, Faculty of Arts, University of Ljubljana Aškerčeva 2, 1000 Ljubljana
vojko.gorjanc@ff.uni-lj.si

Abstract

Although recent years have witnessed a growth in the number of computational language resources and tools, a lot still needs to be done, especially with low-density languages. This is the case with all South Slavic languages and especially Montenegrin, the fourth standard of the once Serbo-Croatian language that has been re-codified only recently. Even though it became the official language of Montenegro in 2007, there still isn't any publicly available electronic corpus that would be available for empirical research of linguistic, translatological or any other inquiry. This paper introduces the first publicly available English – Montenegrin parallel corpus of subtitles. It describes the process of corpus compilation, presents linguistic annotation and accessibility of the corpus through web concordancers. Furthermore, it gives a brief overview of linguistic situation in Montenegro with some of the most important recent developments especially in the light of the recent official international recognition of the language which took place in December 2017.

1. Introduction

Recent years have witnessed a growth in the number of machine-readable corpora and language tools for a number of world languages. It is currently estimated that there are 7,097 languages in the world, an updated number of officially recognized languages listed by Ethnologue[1] which is to be taken arbitrarily. Out of this number, only in 2006 there were corpora available for less than 1% of all world languages, and 20-30 of these fall into the category of high-density and medium-density languages, where "density" is understood to represent the number of computational resources available (Maxwell and Hughes, 2006). The first group would include a handful of languages only, including English, German, Arabic, etc. Today, the number of available resources has increased to app. 90 languages, which means app. 1.2% of all world languages having any kind of publicly available computational resources. The majority of these, however, are lower-density languages as resources are rather scarce.

Being a multilingual and integrative society, Europe is estimated to cover more than 80 languages, of which 23 are official and the rest are either minority or immigration languages. However, a number of these languages is technologically not supported sufficiently and run the risk of being marginalized or digitally extinct. Thus, a number of initiatives have been introduced, such as the META-NET Strategic Research Agenda for Multilingual Europe 2020, with the aim of using various language technologies for overcoming language barriers, enabling free flow of information, goods, and innovation, thus creating a single digital space and marketplace[2].

When it comes to electronic language resources and corpora for some of the major official languages of former Socialist Federal Republic of Yugoslavia, Bosnian, Croatian, Montenegrin, Serbian, Slovenian (BCMSS), the majority of resources are available for Slovenian, followed by Croatian, Serbian, and Bosnian. The last language standard based of the once Serbo-Croatian or Croato-Serbian language that has recently been re-codified and internationally recognized in December 2017 is Montenegrin. So far, there has not been any electronic public corpora of any kind available for the study of this standard. This all testifies to the fact that there is still a lot of work to be done.

This paper presents Opus-MontenegrinSubs 1.0, the first parallel English-Montenegrin electronic corpus developed as a joint effort of researchers from the University of Montenegro, Jožef Stefan Institute, University of Helsinki, and University of Ljubljana. First we will briefly discuss the potentials and possible applications of parallel corpora, the specifics of subtitle corpora as a sub-type of parallel corpora, followed by an outline of the current available parallel corpora for BCMSS. Since this is the first electronic resource of Montenegrin language that is being presented, we will give a brief overview of the linguistic situation in Montenegro followed by the description of the corpus itself and the first study based on it.

[1] http://www.ethnologue.com, accessed: March 15th, 2018, at 11:30 pm.

[2] http://www.meta-net.eu/sra/key-messages, accessed: March 30th, 2018, at 1:00 pm.

2. Parallel subtitle corpora for BCMSS: the potential, current state and specifics

Parallel corpora have found a number of applications in linguistics, translatology, translation practice, and beyond. They proved to be an indispensable tool for a number of contrastive linguistic studies, word sense disambiguation and construction of lexicons, as an input for parallel concordancing systems. Furthermore, the use of parallel corpora has especially become a trend in translation and interpreting studies for developing and training statistical machine translation systems (which require large amounts of parallel language data), for the study of regularities of translations and translators, translation teaching and learning, translation practice including terminology extraction, identifying translation equivalents and correspondents, translation quality assessment, etc. (Bywood et al 2013; Hu, 2016). However, as noted by Tiedemann (2007), most of the existing parallel corpora cover the high-density languages and the domains of legislation, administration and technical documentation.

With regards the parallel subtitle corpora for BCMSS, most of them were developed as subcorpora within the Opus2 project (Tiedemann, 2009) and for each of the western South Slavic languages they include:

- Bosnian: subcorpus OpenSubtitles 2011 (tokens 26,491,099, words ~ 20,906,596),
- Croatian: subcorpora OpenSubtitles 2011 (tokens 111,981,881, words ~ 86,600,021), TedTalks (tokens 1,285,011, words ~ 993,749),
- Serbian: subcorpus OpenSubtitles 2011 (tokens 154,063,822, words ~ 119,149,120),
- Slovenian: subcorpus OpenSubtitles 2011 (tokens 109,690,961, words ~ 81,500,854).

Another project that involved the creation of subtitle corpora for Slovenian and Serbian was SUMAT (tokens 1,250,000/ 1,500,000) (Bywood et al., 2013, Fishel et al., 2012).

Apart from the above mentioned possible applications of corpora of this type, subtitles can be used for the study of text compression and summarization. The reason for this are unique features of subtitles that make them a specific language resource in many ways. Subtitles are usually transcriptions of spontaneous speech with a diversified language (genre, slang, colloquialisms) and they can be classified into several categories: interlingual and intralingual (depending whether they represent a translation from a source to a target (foreign) language, or they are in the same language as the source audiovisual text); monolingual or multilingual (depending on the number of translations into different languages which are shown on the screen); pre-recorded and live. Interlingual subtitling, which we refer to when we use the term in this paper, is a specific form of translation practice since subtitles per se are “a vulnerable modality” for various reasons (Diaz Cintas and Ramael, 2007). This is primarily the case because viewers are exposed both to the source and target text, and there are specific time and space constraints: they are usually shown in one or two lines with 30 – 40 characters, cca. 3 – 7 seconds only with no room for annotation. This calls for specific translation strategies among which condensation (it is estimated that subtitles are 40 – 75% shorter than spoken version), omission is seen as a legitimate strategy (especially in cases of redundancies and spoken discourse markers such as exclamations, false starts, repetitions, hesitations, question tags, etc.), cultural substitution, generalization and specification. Moreover, standardization is also used frequently (especially in cases of slang, regionalisms, grammar mistakes, etc.), and occasional censorship. This shows that subtitles should be approached as a specific, yet an important and unique, resource of translated language.

3. Linguistic situation in Montenegro: a brief overview of recent history

As previously mentioned, Montenegrin language is the last out of four re-codified standards that stem from the same linguistic base of the polycentric Serbo-Croatian language. The remaining three include Bosnian, Croatian and Serbian. Similarly to these standards, it is based on the Eastern Herzegovian Shtokavian dialect. As it has been the case with most neighboring countries, the situation in language policy in Montenegro has reflected a rather turbulent political situation in the Socialist Federal Republic of Yugoslavia (SFRY), and later the Socialist Republic of Yugoslavia (SRY) and the union of Serbia and Montenegro (SM), only to reach its current state in the post-2006 period when Montenegro became independent.

During the pre-1991 period, Serbo-Croatian was one of the official languages of the SFRY, together with Slovenian, Macedonian and other languages which were constitutionally of equal status, but the reality was somewhat different as they seem to have been in a position of a “competitive coexistence” (Gorjanc, 2013; Požgaj Hadži et al., 2013). In the light of the above mentioned historical events, the constitutions of the Socialist Republic of Montenegro of 1963 and 1974 define Serbo-Croatian as the language in official use in Montenegro. This polycentric language, as its very name suggests, stemmed from two main standards, the eastern (with its center in Belgrade) and the western (with its center in Zagreb), while other language forms with their center in Podgorica (i.e. Titograd, as the capital of Montenegro was called then) and Sarajevo were marginalized as regional variants, and re-codified as provincialisms, a situation which would later be seen as having significant political implications especially regarding the politics of assimilation and hegemony.

With the disintegration of Yugoslavia in 1991, Serbo-Croatian was re-codified into 4 separate standards, starting with Serbian and Croatian, and later followed by Bosnian and Montenegrin. Due to the official state policy of the day, the official language of Montenegro in the constitution of 1992 was designated as Serbian of the Ijekavian standard, and this remained the case until 2007. Shortly after the independence which took place in 2006, and upon the ratification of the new Constitution which took place on 22 October 2007, Montenegrin became the official language in Montenegro. After much controversy arising from two different approaches to the process of standardization, the first Montenegrin grammar and orthography were adopted in 2010 by the Council for General Education. The question of language standardization still remains an ongoing issue and it’s highly debatable whether some of their solutions will fully integrate into language practice.

Census data from 2011 shows an increase in the number of speakers who designate their mother tongue as Montenegrin. One of the most significant events was certainly the international recognition of Montenegrin language and the assignment of the international code. This was approved on 8 December 2017 and the ISO 639-2 and 3 code [cnr] was assigned. Needless to say, much still needs to be done. The first volume of the Dictionary of the Montenegrin Literary and Vernacular Language published by the Montenegrin Academy of Sciences and Arts in 2016 was soon withdrawn due to reactions of part of political public to some entries that seemed to be controversial. Regarding the electronic language resources, up to this date, there haven't been any officially published electronic corpora of Montenegrin language that would be available to researchers for various linguistic and translatological studies. That is why the electronic corpus which we will present in this paper is of high importance as it is the first ever electronic corpus of Montenegrin language.

4. Corpus compilation and accessibility

The corpus Opus-MontenegrinSubs 1.0 contains parallel English-Montenegrin subtitles. The data and copyrights were obtained from the Radio and Television of Montenegro, the public service broadcaster of Montenegro. The corpus consists of English and Montenegrin subtitles of three series: House of Cards, Damages, and Tudors. The corpus contains 10 seasons, and 110 episodes, which are cca. 5,563 minutes in length. A detailed breakdown is given in Table 1.

Series	No. of seasons	No. of episodes	Length
House of Cards	1	13	686 mins.
Damages	5	59	2878 mins.
Tudors	4	38	1999 mins.

Table 1: Corpus breakdown

4.1. Processing the corpus

Sentence alignment and basic encoding was performed inside the OPUS project3. The original subtitle files were converted to Unicode UTF-8 using *iconv* and the Unix tool *file* for automatic detection of the character set in the original file. After that the OPUS subtitle tools (Lison & Tiedemann, 2018) were applied to convert the files to standalone XML with sentence markup; the remaining XML-well formedness problems were fixed with the program *tidy*. Finally, all translated subtitles were aligned using the time-based alignment method described in Tiedemann (2007) and the standard OPUS import pipeline was used to integrate the data in OPUS with download formats in XML, plain text and TMX.

In the second stage, the source XML data was converted to the latest version of TEI (TEI, 2018), so that the subtitles for each language are stored in a separate <text> element, with the sentence alignments being maintained by cross-links as well as separately, in a <linkGrp> element. An important part of this conversion was also the encoding of the <teiHeader> element, which contains the meta-data of the corpus, explicating its authors, license etc. but also listing all the used XML elements in the corpus, together with a short explanation, and how the MSD annotation prefixes are to be interpreted.

Then, the English and Montenegrin texts were tokenized, sentence segmented and tagged with morphosyntactic descriptions (MSDs) and lemmas. To perform this annotation for Montenegrin, we used the ReLDI tokeniser[4] and tagger[5] (Ljubešić & Erjavec, 2016) with its model for Serbian. The MSD tagset used follows the MULTEXT-East specifications (Erjavec, 2012), in particular, the version 5 specifications for Bosnian[6].

For English, we used Tree Tagger (Schmidt 1994, 1995) with its model for English, which uses the Penn Tree Bank tagset. In order to make the English tagset harmonized with the one for Montenegrin, we converted it to the SPOOK tagset for English[7], i.e. performed a 1-1 mapping between the original PTB tagset to MULTEXT-East compatible SPOOK tagset.

Figure 1 illustrates the TEI encoding of the linguistically annotated corpus, giving the first translation unit (annotated as anonymous block, <ab>) for both languages. As can be seen, each language text contains the divisions marking the structure of the corpus, while the translation units are given IDs and the alignment via their @corresp attribute. Each translation unit is then divided into sentences, and these into words, punctuation symbols and whitespace. The tokens are lemmatized and MSD tagged, where the value prefix *mte* resolves to the MULTEXT-East MSD definition (i.e. its decomposition into features), while the *spook* one resolves to the SPOOK decomposition. It should be noted that the original PTB tag is retained as the value of the *@function* attribute.

4.2. Corpus distribution and use

The TEI corpus was converted to the so called vertical format, used by (no)Sketch Engine and mounted on the CLARIN.SI concordancers,[8] namely noSketch Engine and KonText, as well as Sketch Engine, so that it is available on-line for searching and exploration; both concordancers also allow displaying the aligned translation units.

The complete corpus in TEI, as well as vertical format, was also made available for download in the CLARIN.SI repository (Božović et al., 2018) under the Creative Commons - Attribution-ShareAlike license.

5. First corpus studies

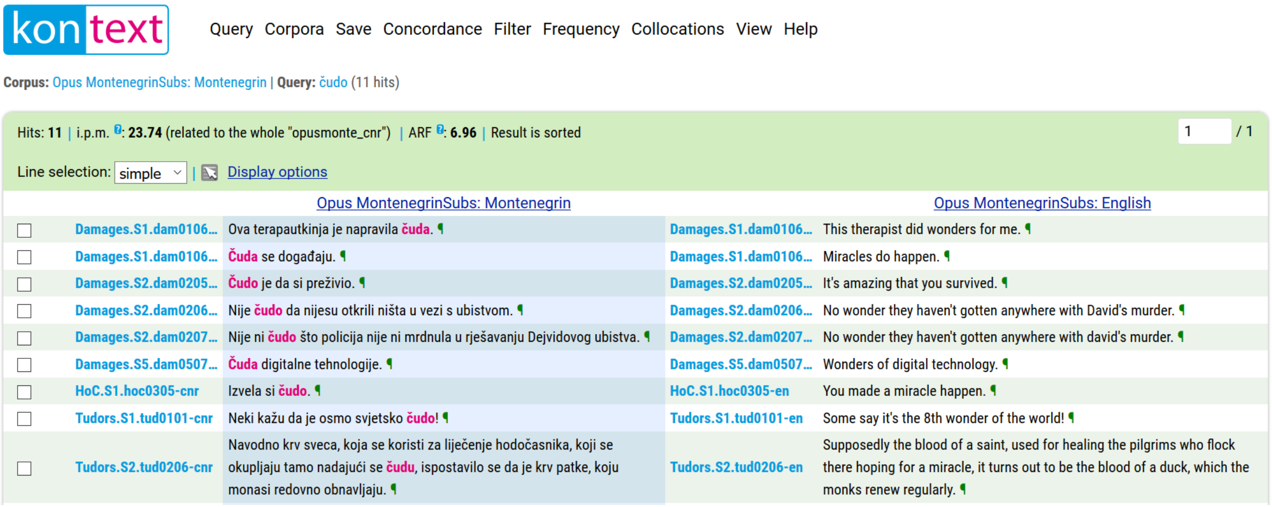
The first study based on this corpus is the one conducted for the purpose of the Ph.D. studies by Petar Božović with the thesis topic *Audiovisual Translation and Elements of Culture: A Comparative Analysis of Transfer with Reception Study in Montenegro*, which is in the field of translation studies and corpus linguistics. Corpus-based translation studies are becoming increasingly relevant for the industry ever since the methodology has been introduced from an allochthonous field of corpus linguistics in the seminal paper by Baker (1993). It wasn't long after this that it became evident that using corpora in translation research was to have a great potential for scholarly empirical research, but also for terminologists and practitioners.

[3] http://opus.nlpl.eu/MontenegrinSubs.php
[4] https://github.com/clarinsi/reldi-tokeniser
[5] https://github.com/clarinsi/reldi-tagger
[6] http://nl.ijs.si/ME/V5/msd/html/msd-bs.html
[7] http://nl.ijs.si/spook/msd/html-en/msd-en.html
[8] http://www.clarin.si/info/concordances/

```
<text xmlns="http://www.tei-c.org/ns/1.0" xml:lang="cnr">
  <body>
    <div type="series" xml:id="Damages-cnr">
      <div type="season" xml:id="Damages.S1-cnr">
        <div type="episode" xml:id="Damages.S1.dam0101-cnr">
          <ab n="1" xml:id="Damages.S1.dam0101.SL1-cnr" corresp="#Damages.S1.dam0101.SL1-en">
            <s>
              <w ana="mte:Agpfpny" lemma="opasni">OPASNE</w><c> </c>
              <w ana="mte:Ncfpn" lemma="igra">IGRE</w><c> </c>
              <w ana="mte:Ncmsn" lemma="pilot">Pilot</w><c> </c>
              <w ana="mte:Ncfpg" lemma="epizoda">epizoda</w>
            </s>
          </ab>
 …

<text xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
  <body>
    <div type="series" xml:id="Damages-en">
      <div type="season" xml:id="Damages.S1-en">
        <div type="episode" xml:id="Damages.S1.dam0101-en">
          <ab n="1" xml:id="Damages.S1.dam0101.SL1-en" corresp="#Damages.S1.dam0101.SL1-cnr">
            <s>
              <w lemma="Season" function="NP" ana="spook:Np-s">Season</w><c> </c>
              <w lemma="1" function="CD" ana="spook:M-c">1</w><c> </c>
              <w lemma="episode" function="NN" ana="spook:Nc-s">Episode</w><c> </c>
              <w lemma="1" function="CD" ana="spook:M-c">1</w><c> </c>
              <w lemma="pilot" function="NN" ana="spook:Nc-s">Pilot</w><c> </c>
              <pc function="(" ana="spook:Z">(</pc>
              <w lemma="Dimension" function="NP" ana="spook:Np-s">Dimension</w>
              <pc function=")" ana="spook:Z">)</pc>
            </s>
          </ab>
  …
```

Figure 1: TEI encoding of the corpus texts



Corpus: Opus MontenegrinSubs: Montenegrin | **Query:** čudo (11 hits)

Hits: **11** | i.p.m.: **23.74** (related to the whole "opusmonte_cnr") | ARF: **6.96** | Result is sorted   1 / 1

Line selection: simple | Display options

		Opus MontenegrinSubs: Montenegrin		Opus MontenegrinSubs: English
☐	Damages.S1.dam0106…	Ova terapautkinja je napravila **čuda**. ¶	Damages.S1.dam0106…	This therapist did wonders for me. ¶
☐	Damages.S1.dam0106…	**Čuda** se događaju. ¶	Damages.S1.dam0106…	Miracles do happen. ¶
☐	Damages.S2.dam0205…	**Čudo** je da si preživio. ¶	Damages.S2.dam0205…	It's amazing that you survived. ¶
☐	Damages.S2.dam0206…	Nije **čudo** da nijesu otkrili ništa u vezi s ubistvom. ¶	Damages.S2.dam0206…	No wonder they haven't gotten anywhere with David's murder. ¶
☐	Damages.S2.dam0207…	Nije ni **čudo** što policija nije ni mrdnula u rješavanju Dejvidovog ubistva. ¶	Damages.S2.dam0207…	No wonder they haven't gotten anywhere with david's murder. ¶
☐	Damages.S5.dam0507…	**Čuda** digitalne tehnologije. ¶	Damages.S5.dam0507…	Wonders of digital technology. ¶
☐	HoC.S1.hoc0305-cnr	Izvela si **čudo**. ¶	HoC.S1.hoc0305-en	You made a miracle happen. ¶
☐	Tudors.S1.tud0101-cnr	Neki kažu da je osmo svjetsko **čudo**! ¶	Tudors.S1.tud0101-en	Some say it's the 8th wonder of the world! ¶
☐	Tudors.S2.tud0206-cnr	Navodno krv sveca, koja se koristi za liječenje hodočasnika, koji se okupljaju tamo nadajući se **čudu**, ispostavilo se da je krv patke, koju monasi redovno obnavljaju. ¶	Tudors.S2.tud0206-en	Supposedly the blood of a saint, used for healing the pilgrims who flock there hoping for a miracle, it turns out to be the blood of a duck, which the monks renew regularly. ¶

Figure 2: Searching the corpus in the KonText concordander

Hence, the research which is based on the corpus is focused on the highly-influential and fast-growing audiovisual translation field with the aim of mapping the different translation strategies for rendering the extralinguistic elements of culture in subtitling. This is an issue that is at the core of some of the major challenges in the industry as transfer of elements of culture proves to be one of the “crisis points” in translation process, especially in subtitling due to the time and spatial constraints of the modality and it can have an important influence on the reception and placement of the audiovisual product on the target market and for the target audience (Pedersen, 2011). The extralinguistic elements will be extracted as types, not tokens, by using one of the concordancers. After that, these elements will be categorized according to the level of transculturality, and translation strategy for rendering that element will be defined. The goal is to gain a better understanding and map how culture is rendered in subtitling from English into Montenegrin and to supplement this with the reception study which is also part of this research. It is hoped that this will provide an important empirical feedback for translators and broadcasting companies who could tailor the translation policy better to meet the needs and expectations of the real, not ideal or intuitive, target audience.

6. Conclusions

The Opus-MontenegrinSubs 1.0 is the first publicly available parallel electronic corpus of Montenegrin language the appearance of which is timely considering the recent sociolinguistic developments, especially constitutional and international acknowledgement that this language has received. Needless to say, a lot still remains to be done in order to provide the computational resources and tools necessary for state-of-the-art linguistic approaches and analyses. It is hoped that this corpus will encourage other researches and contribute to the affirmation and development of corpus linguistics and corpus-based translation studies in the region. Moreover, it is hoped that it will encourage the development of other corpora of Montenegrin language, first and foremost of the reference corpus, which would be of a pivotal importance for the process of restandardization and without which a modern linguistic description of Montenegrin will not be possible.

Acknowledgments

The authors acknowledge the support of the Radio and Television of Montenegro and Vuk Perović for the corpus materials and are most grateful for the permissions granted.

7. References

Jan Pedersen. 2011. *Subtitling Norms for Television: An Exploration Focusing on Extralinguistic Cultural References*. John Benjamin Publishing.

Jörg Tiedemann. 2007. Improved Sentence Alignment for Movie Subtitles. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'07)*.

Jörg Tiedemann. 2007. Building a Multilingual Parallel Subtitle Corpus. In *Proceedings of the 17th Conference on Computational Linguistics in the Netherlands (CLIN 17)*. Leuven, Belgium.

Kaibao Hu. 2016. *Introducing Corpus-based Translation Studies*. Springer.

Lindsay Bywood, Martin Volk, Mark Fishel, and Panayota Georgakopoulou. 2013. Parallel subtitling corpora and their applications in machine translation and translatology. *Perspectives: Studies in Translatology*, 21(4): 595–610.

Mike Maxwell and Baden Hughes. 2006. Frontiers in linguistic annotation for lower-density languages. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 29–37, Sydney, Australia. Association for Computational Linguistics.

Mona Baker. 1993. Corpus Linguistics and Translation Studies – Implications and Applications. In: Mona Baker et al., ed., *Text and Technology: In Honor of John Sinclair*, pages 233–252, John Benjamins Publishing.

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*, Portorož. http://www.lrec-conf.org/proceedings/lrec2016/pdf/811_Paper.pdf

Petar Božović, Tomaž Erjavec, Jörg Tiedemann, Nikola Ljubešić, and Vojko Gorjanc. 2018. *English-Montenegrin parallel corpus of subtitles Opus-MontenegrinSubs 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1176.

Pierr Lison and Jörg Tiedemann. 2018. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. *LREC 2018*.

Ranko Bugarski. 2009. *Nova lica jezika*. Biblioteka XX vek.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*. Manchester.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin.

Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language resources and evaluation*, 46(1): 131–142, doi: 10.1007/s10579-011-9174-8

Vojko Gorjanc. 2013. Slovenačka jezička politika i odnosi društvene moći. In: Vesna Požgaj Hadži, ed., J*ezik između lingvistike i politike*, pp.12–36. Biblioteka XX vek.

Vesna Požgaj Hadži, Tatjana Balažic Bulc, and Vlado Miheljak. 2013. Srpskohrvatski jezik iz slovenske perspektive. In: Vesna Požgaj Hadži, ed., *Jezik između lingvistike i politike*, pages 12–36. Biblioteka XX vek, Beograd.

TEI Consortium, eds. 2018. *TEI P5: Guidelines for Electronic Text Encoding and Interchange 3.3.0*. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Zapis in prikaz starejših pesniških besedil ter njihovih variant v TEI

Nina Ditmajer[*], Matija Ogrin[*], Tomaž Erjavec[‡]

[*] Inštitut za slovensko literaturo in literarne vede ZRC SAZU
ZRC SAZU, Novi trg 2, Ljubljana
nina.ditmajer@zrc-sazu.si, matija.ogrin@zrc-sazu.si
‡Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

Na primeru diplomatičnega prepisa baročne Foglarjeve pesmarice (1757–1762), ki je v literarnozgodovinski vedi prepoznana kot najstarejši ohranjeni slovenski štajerski rokopis pesmi, v prispevku prikažemo problematiko zapisa verza in variantnih mest po Smernicah za kodiranje besedil TEI. Za znanstvenokritično izdajo besedila je pomembno, da pri prepisovanju dosledno označimo verzno strukturo vsake kitice v pesmi, kot tudi, da izdelamo kritični aparat variantnih mest glede na druge ohranjene verzije iste pesmi. Za praktično uporabnost izdaje je zelo pomembno, da so struktura in variantna mesta v spletni izdaji v formatu HTML nazorno izpisana. V prispevku najprej predstavimo Foglarjevo pesmarico in prikažemo diplomatični zapis verza v problematičnih primerih. V nadaljevanju se osredotočamo na zapis variantnih mest, določen z metodološko koncepcijo Foglarjevega rokopisa kot glavnega besedila, in na problematiko njihovega prikaza v HTML. Na koncu orišemo več možnosti prikaza elektronskega diplomatičnega besedila v izpisu HTML, npr. prikaz aparata v obliki opomb in večbarvni sinoptični prikaz glavnega in variantnega teksta. Ob orisanih prednostih in pomanjkljivostih novejših spletnih orodij za prikaz in analizo TEI kodiranih znanstvenokritičnih besedil nakažemo razlike med zahodnoevropsko in slovensko literarno tradicijo, ki pomembno določajo potrebe in koncepte znanstvenega izdajanja slovenskih slovstvenih tekstov.

Encoding and rendering early modern Slovenian poetry texts and their variants in TEI

On the example of the diplomatic transcription of the Foglar hymnal (1758–1762), which is in literary studies considered to be the oldest extant manuscript of Slovenian poetry from the Styria region, we show the challenges of encoding a verse and its variants, using the Text Encoding Initiative Guidelines. For a scholarly edition of the text it crucial that the transcription consistently marks up the structure of the verses of every stanza of the poem, as well as the variant readings in the other extant versions of the poem. For presenting the edition it is also very important that the structure and the readings in the on-line HTML edition are rendered clearly. The paper first introduces the Foglar hymnal and presents the difficult cases in the diplomatic annotation of the verse. Next, the encoding of the variant readings is presented, based on the methodological decision to treat the Foglar manuscript as the base text, followed by the challenges of their presentation in HTML. Finally, we sketch several possibilities of how to display the critical apparatus in HTML, e.g. the display of the apparatus as notes and using colours in a synoptic rendering of the base and variant texts. On the basis of described advantages and disadvantages of current web-based tools for the display and analysis of TEI encoded scholarly texts we sketch the differences between the Western European and Slovenian literary tradition, which define the requirements and concepts of scholarly publishing of Slovenian literary texts.

1. Uvod

Za slovenski jezikovni prostor 18. stoletja je značilno soobstajanje rokopisnih in natisnjenih literarnih besedil. Še posebej pogoste so bile pesmarice, ki so jih pisci bodisi prepisovali iz predhodnih natisnjenih ali rokopisnih pesmaric, letakov ob posebnih priložnostih (npr. romanje, posvetitev cerkve), lekcionarjev, katekizmov in molitvenikov bodisi so jih pisali po nareku ali posluhu. Glede na žanrsko pojavnost v obravnavanem obdobju je mogoče govoriti o skupnem repertoarju pesniških besedil v različnih pokrajinah v slovenskem jezikovnem prostoru. Izobraženi pisci, ki so bili v tem času večinoma duhovniki, so se praviloma držali knjižne literarne tradicije predhodnih cerkvenih besedil, predvsem Dalmatinove Biblije in Schönlebnovega lekcionarja, vnašali pa so tudi nekatere jezikovne inovacije, skladne s sočasno govorno podobo in pokrajinsko pripadnostjo pisca. Manj izobraženim piscem, ki so se večinoma naučili osnov branja in pisanja v domači župniji, je bilo bolj pomembno, da bo pesniško besedilo razumljivo ljudem, ki ga bodo peli v cerkvi, pri opravljanju katere od pobožnosti ali na romanju. Zato je v slovenskem jezikovnem prostoru vse do sredine 19. stoletja in do oblikovanja vseslovenskega knjižnega jezika obstajalo več pokrajinskih različic, ki so pomembno pripomogle k nastanku variantnih literarnih besedil.

V članku želimo na primeru diplomatičnega prepisa izbranega starejšega pesniškega besedila prikazati problematiko zapisa verza in variantnih mest istega besedila v preostalih rokopisnih in tiskanih verzijah po Smernicah TEI (TEI Consortium, 2018). Priporočila TEI sestavljajo smernice (prozni opisi oznak) in vrste posameznih modulov, ki jih je mogoče kombinirati, da bi ustvarili želeno shemo XML za določen projekt. S tem je nato mogoče formalno validirati zapis XML posameznih elektronskih izdaj, dostopna pa je tudi dokumentacija za vse uporabljene elemente XML (Ogrin in Erjavec, 2009). Zato so Smernice TEI splošno uporaben, *de facto* standard za pripravo raznolikih elektronskih besedil, od preprostih bralnih izdaj do znanstvenokritičnih edicij, slovarjev in jezikoslovnih korpusov.

Za znanstvenokritično izdajo pesniškega besedila je pomembno, da pri prepisovanju dosledno označimo verzno strukturo vsake kitice v pesmi; sem sodi tudi primeren zapis refrena in posebnih znakov, npr. znaka za ponovitev refrena »&«, medtem ko drugih grafičnih simbolov, rabe rdečega črnila in podobnih grafičnih posebnosti načeloma ne podajamo, saj nimajo besedilnega pomena. V prispevku prikažemo tudi način diplomatičnega zapisa verza v tistih pesmih, kjer verzna struktura ni razvidna. V nadaljevanju prispevka se osredotočamo na zapis variantnih mest ter problematiko njihovega prikaza v izpisu HTML. Prikaz variantnih mest je zelo pomemben in koristen, če želimo primerjati pravopisno, glasoslovno, oblikoslovno ali besedno raven posameznega besedila. Prav tako lahko z uporabo primernih orodij preučujemo kitične oblike, verz in metrum. Na koncu prikažemo več možnosti prikaza elektronskega diplomatičnega besedila v izpisu HTML, npr. prikaz aparata v obliki opomb, večbarvni hkratni prikaz osnovnega in variantnega teksta, ter prednosti in pomanjkljivosti novejših spletnih orodij za prikaz in analizo takih besedil.

2. Opis izbranega rokopisa in struktura izdaje

Foglarjeva pesmarica (1757–1762) je v literarnozgodovinski vedi prepoznana kot najstarejši ohranjeni slovenski štajerski rokopis pesmi (prim. Ditmajer, 2017). Vsebuje najstarejše slovenske marijinoceljske pesmi,[1] štiri svetniške pesmi, praznično pesem v čast Sveti trojici, dve pesmi z eshatološko vsebino, pesem o caščenju Jezusovega imena, postno spokorno pesem in pesem o ljubezni do Boga. Po jezikovnih značilnostih sodeč izvira rokopis iz vzhodnoštajerskega prostora, omenjeni so tudi mesto Maribor, reka Drava in cerkev svetega Martina v Kamnici. Pesmarico danes hranijo v Narodni in univerzitetni knjižnici pod signaturo R 281102. Rokopis je vezan v zbornik, ki vsebuje tudi tri tiskane pesmarice: Primož Lavrenčič, *Missionske catholish karshanske pejssme* (1752); Ahacij Stržinar, *Peissem Od teh velikih odpuſtikov* (1744); in *Andohtlive pejsme* (1756). Platnice so iz lesenih deščic, prevlečenih s temnorjavim usnjem. Zapirača sta že odpadla, v obeh platnicah so vidni njuni sledovi. Prvotnega številčenja ni, s svinčnikom je vpisana komaj vidna sekundarna paginacija. Rokopis sestavlja osem leg. Prvih šest snopičev je vsebovalo po 8 folijev (kvaternij), zadnja dva po 6 (ternij). Zadnji folij zadnje lege (i) je prilepljen na zadnjo platnico knjige. Manjkata dva folija, po eden v prvi in v sedmi legi, kar predstavlja kodikološka formula tako: 7(-1) + 8 + 8 + 8 + 8 + 8 + 5(-1) + 5 + i (Ditmajer in Ogrin, 2018).

Rokopisne pesmi so napisane v osmih različnih pisavah. Prvih deset pesmi v rokopisu je zapisal Lovrenc Foglar [Voglar]. Bil je nevešč pisanja. Pisal je v malih tiskanih črkah, pisava je zelo oglata, kot da bi pisec sledil pisavi natisnjenih pesmaric, verzi nimajo ločil, med posameznimi zlogi besed pa so vidni razmiki, česar v natisnjenih pesmaricah ne zasledimo, npr. *vei ko ma* (vekomaj). To lahko kaže na to, da je pisal po nareku ali po posluhu, saj so takšni zapisi besed po navadi pod notnim črtovjem. Pesmi imajo na koncu kitice znak za ponovitev refrena »&«, ime Jezus pa se zapisuje z velikimi tiskanimi črkami, kar je prav tako značilnost predhodnih natisnjenih pesmaric. Rabo predlog nakazuje tudi kristogram IHS na koncu desete pesmi, posnemanje natisnjenih pesmaric ali pripravo rokopisa za tisk pa nakazujejo kustode na dnu vsake strani.

Naslednje štiri pesmi je istega leta napisal Matija Vezjak/Bezjak/Bizjak (*Matiaſ Weſiak*). Na koncu sledi podoben sklepni zapis kot pri Foglarju s simbolom Družbe Jezusove, Marijinim kronogramom in križem. Tudi zanj je značilno zapisovanje verzov brez ločil, razmiki med zlogi, velike tiskane črke, ponovitev refrena je zapisana z latinsko ligaturo »&c«.

Tretji pisec se ni podpisal, nehal je tudi z zaporednim štetjem pesmi. Pisal je s pisanimi črkami, vstavljal ločila (vejica, pika, deljaj, dvopičje) in znak za ponovitev refrena »&«. Z veliko začetnico piše tako osebna kot stvarna lastna imena (*Ioseph*, *Terplenje*), kar nakazuje na poznavanje vsaj nemškega jezika.

Prav tako neznani četrti pisec je pesmi zapisal s pisanimi črkami, uporabljal je ločila (pika, dvopičje), pri pisanju je viden vpliv nemškega in latinskega jezika, prav tako je večina besed zapisanih z veliko začetnico.

Peti pisec izkazuje podobno oglato pisavo kot prva dva, prav tako ni rabil ločil, viden je jezuitski simbol IHS.

Naslednje pesmi je napisal Jožef Glazer, učitelj (*ludimagister*) pri Gornji sveti Kungoti, ki je bila podružnična cerkev župnije sv. Martina v Kamnici. Pesmi so napisane s pisanimi črkami, vendar pisava ni lepopis, od ločil je rabil pike in vejice. Župnijski učitelji v tistem času prav tako niso imeli višje izobrazbe, saj so se prva učiteljišča na našem ozemlju pojavila šele v začetku 19. stoletja.

Zadnji dve pesmi sta pisali dve različni roki brez podpisa. Rokopis se konča z letnico 1762, latinskim zapisom *Domine deus pater* (Gospod Bog Oče), abecedo in števili od ena do dvajset ter letnico 1758. V besedilu je viden latinski vpliv, zadnji pisec je uporabljal vejice in pike, na začetku verza pa je vedno velika začetnica.

Ta rokopis, ki je z opisom in digitalnim faksimilom predstavljen (Ditmajer in Ogrin, 2018) v *Registru slovenskih rokopisov 17. in 18. stoletja*, bo elektronska znanstvena izdaja predstavila z naslednjimi glavnimi elementi:

- digitalni faksimile predstavlja dokumentarni vir celotne izdaje: jamči za verodostojnost tako prepisov kakor kritičnega aparata ter interpretacij, podanih v opombah in študijah;
- diplomatični prepis predstavi besedilo rokopisa z vsemi historičnimi besedilnimi posebnostmi, vključno z napakami; dodan mu je kritični aparat variantnih mest;
- kritični prepis predstavi besedilo rokopisa v prilagojeni, branju namenjeni podobi, kar najbolj zvesti domnevni historični izreki; je jezikoslovna interpretacija, pripravljena po pretehtanih jezikovnih načelih;
- dodatek vsebuje na prvem mestu seznam rokopisov ter tiskov, po katerih je narejen kritični aparat variantnih mest; sledi mu znanstveni komentar z razlago okoliščin nastanka rokopisa, jezikovnih in literarnih posebnosti ter uredniških načel prepisa.

2.1. Diplomatični prepis glavnega besedila Foglarjeve pesmarice

V starejših slovenskih pesmaricah ena grafična vrstica ne predstavlja vedno tudi enega metričnega verza. Večkrat

[1] Pesmarica je na prvi pogled namenjena slovenskim romarjem, ki so hodili s Spodnje Štajerske na romarsko pot v Marijino Celje (Maria Zell), ki leži na Zgornjem Štajerskem v današnji Avstriji. Iz imena Mariazell je že kmalu nastal slovenski historični eksonim Marijino Celje, ki ga je v tej obliki Prešeren že uporabljal skupaj s Trsatom idr. znanimi romarskimi središči. Od tod pridevnik marijinoceljski.

so pisci zaradi pomanjkanja prostora na papirju naslednjo besedo ali besedno zvezo zapisali v drugi grafični vrstici. V diplomatičnem prepisu smo v zapisu TEI uporabili za številčenje kitic oznako <label>, verzi <l> (*line*) so gnezdeni v oznako za kitico <lg> (*line group*), prelom verzne vrstice pa smo preprosto zaznamovali z oznako <lb> (*line break*), kot prikazuje spodnji primer zapisa:

```
<label>2</label>
  <lg>
     <l>Od Boga ozhe ta jeſt le zhem <lb/>Sa zhe ti</l>
     <l>Ker je mira kel ne dela na <lb/>Tem ſvei ti</l>
     <l>Tou na to ti Semli</l>
     <l>Hva la bo di nie mi</l>
     <lg type="refrain">
       <l>Sa hva le na do vei ko ma</l>
       <l>Bo di Sveta tro y za</l>
     </lg>
  </lg>
```

Težavo povzročajo predvsem pesmi, kjer se pisec ni oziral na verzno vrstico in je pesem zapisal v prozni obliki. Tako smo pesmi, kjer se drugi verz prične v isti grafični vrstici, kjer se je končal prvi verz, zapisali z uporabo oznake <ab> (*anonymous block*); da gre za kitico, pa smo zaznamovali z uporabo atributa @type, in označili le prelome vrstic, kot prikazuje naslednji primer:

```
<ab type="lg"><label>1.</label>
<lb/>Vsak Brat inu Sestra <lb/>Serze Posdigni, Iesusa
<lb/>Mario  Josepha hvali: <lb/>Klizi Jesus Maria mojo
<lb/>Serze moj glas, ô Jo-  <lb/>seph moj varih sdajna
<lb/>Posledni zhass.</ab>
```

3. Zapis variantnih mest v diplomatičnem prepisu

V knjižnici *Elektronskih znanstvenokritičnih izdaj slovenskega slovstva* (eZISS)[2] je vsaka izdaja izoblikovana tako, da je zaradi zgodovinsko-jezikoslovne problematike osredinjena na možnost vizualizacije oz. primerjave diplomatičnega in kritičnega prepisa (Ogrin in Erjavec, 2009). Prvi primer, ko smo v katero od izdaj eZISS z izrecnim označevanjem v zapisu TEI vključili problem obstoja istega besedila v dveh verzijah, pa je bila izdaja romana *S poti* Izidorja Cankarja (Cankar, 2007). Tu smo izrecno označili razlike med prvotno verzijo romana, objavljeno v *Domu in svetu* leta 1913, in poznejšo, predelano knjižno izdajo iz leta 1919, ki je obveljala kot avtorjeva dokončna verzija.

Že tu smo izmed treh načinov kodiranja variantnih mest, ki jih kodificirajo smernice TEI (TEI Consortium 2018, 12.2), izbrali t. i. metodo vzporednega segmentiranja, saj je ta idealna za zajem manj kompleksne besedilne tradicije, ko želimo primerjati variantna mesta le iz dveh ali nekaj verzij besedila in po možnosti tudi strojno ekstrapolirati celotno besedilo posamezne verzije. Primerjava, ki je bila še v celoti narejena ročno, je našla 478 razlik, izrecno prikazane v enotah kritičnega aparata <app> (*apparatus entry*). Nadaljnja analiza je odkrila tri večje skupine ali tipe popravkov, ki so segali od jezikovnih in slogovnih do pomembnejših vsebinskih posegov, s katerimi je pisatelj niansiral problematiko obeh svojih protagonistov (Cankar 2007, § 5–26).

Pozneje smo se posvečali problemu besedilne variantnosti v proznem besedilu ob pripravljalnem delu za kritično izdajo *Poljanskega rokopisa.* Primerjali smo ohranjeni vzporedni besedili *Poljanskega rokopisa* in njegovega, žal fragmentarno ohranjenega starejšega protografa (Ogrin in Žejn, 2016). V raziskavi smo uporabili dve orodji za kolacijo besedil, Juxta[3] in CollateX.[4] Obe orodji sta usmerjeni najprej k vizualizaciji variantnih mest, obe pa lahko strojno generirata tudi eksterni kritični aparat in ga prikažeta v obliki opomb. Nam sta služili predvsem za analizo in klasificiranje variantnih mest po njihovih lastnostih v tipične skupine, kar je odprlo nadaljnja pomembna spoznanja in perspektive raziskovanja (Ogrin in Žejn, 2016: 131).

Z označevanjem besedilne variantnosti v pesniških besedilih smo se prvič spoprijeli pri pripravi elektronske znanstvenokritične izdaje pesmi Antona Martina Slomška, zasnovane v letih 2006–2011, ki še ni dokončana. Za evidenco variantnih mest v Slomškovih pesmih je bil pripravljen kritični aparat s čez 2.000 elementi <app> z metodo vzporednega segmentiranja. Za njihovo vizualizacijo smo pripravili pretvorbo XSLT, ki v izpisu HTML prikaže variantna mesta v zavitih oklepajih in z obarvanimi črkami, in ta prikaz smo uporabili tudi pri pripravi Foglarjeve pesmarice.

Diplomatičnemu prepisu Foglarjeve pesmarice smo dodali kritični aparat (*apparatus criticus*), ki zajema variantna mesta nekaterih verzij teh pesmi, ohranjenih v rokopisnih in tiskanih slovenskih pesmaricah 18. stoletja. Doslej smo v prepis vključili šest pesmaric: *Pesmarico iz Gorij* (1761–92),[5] Paglovčevo pesmarico (1733),[6] *Cerkvene pesmi in molitve*,[7] Maurerjevo (1754)[8] in Krebsovo pesmarico[9] ter natisnjene Lavrenčičeve

[2] Prim. http://nl.ijs.si/e-zrc/

[3] Prim. http://www.juxtasoftware.org/

[4] Prim. https://collatex.net/

[5] V rokopisu je ohranjena romarska pesmarica iz Gornjih Gorij (prim. NRSS, Ms 113), ki vsebuje 22 pesmi, zapisanih v sedmih različnih pisavah med letoma 1761 in 1792 s podpisom Jestin Amroshizh. Oba rokopisa, štajerski in gorenjski, imata skupnih kar pet marijinoceljskih pesmi, ki se tudi jezikovno in pravopisno ne oddaljujejo preveč, kar pomeni, da je gorenjski pisec bil prepisovalec prej nastale štajerske pesmarice, ki je bila tudi izvirna, saj sta v eni pesmi omenjena reka Drava in Maribor (Ditmajer, 2017).

[6] Svetniška pesem o sveti Notburgi je zapisana tudi v Paglovčevi rokopisni pesmarici, vendar manjkata dve strani, ki sta vključevali drugo polovico četrte kitice, peto, šesto in sedmo kitico ter šest verzov osme. Nad Paglovčevim zapisom pesmi je omenjen tudi tiskan letak s to pesmijo (v Ljubljani, 1738), ki je do danes neznan. Pesem od vernih duš prav tako najdemo pri Paglovcu, kasneje je bila zapisana tudi v Redeskinijevi pesmarici (Od virneh dush v' vizah).

[7] Foglarjeva Pesem od Svete trojice je identično zapisana v pesmarici iz Gornjih Gorij, najdemo pa jo tudi v štajerski cerkveni pesmarici s konca 18. stoletja (NRSS Ms 052).

[8] Pesem o Mariji Magdaleni je zapisana tudi v Maurerjevi koroški pesmarici (1754) pod naslovom Ta drdvga [sic!] pisem od Marie Magdalene (*An Bart Magdalena je vshgana od plamena*).

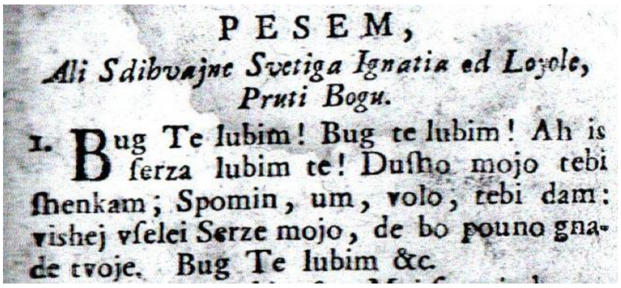
[9] Iz Krebsove rokopisne pesmarice (NRSS Ms 022) iz 18. stoletja je znana pesem Od izgubljene ovčice, pesem z istim naslovom iz

Miſionske Peſme inu molitve (1757).[10] Ker so se omenjena pesniška besedila večinoma pela in se na ta način širila po širšem območju, kjer so takrat živeli Slovenci, ni vedno mogoče ugotoviti, katero besedilo je primarno in katero variantno. Tudi zato smo se odločili, da bomo kot glavno besedilo označili Foglarjev rokopis, kot variantna mesta pa smo označili vse besedilne razlike v zgoraj omenjenih verzijah, ne glede na čas zapisa v pesmarici. Uporabili smo metodo vzporednega segmentiranja variantnih mest in jih prikazali z uporabo <app>, ki vsebuje en vnos glavnega besedila oz. lemo <lem> (*lemma*), navedbo verzije z atributom @witt (*witness*) in eno ali več variantnih mest, označenih z <rdg> (*reading*):

```
<l>
  <app>
     <lem wit="#F">Po ſluſhai kaiti jaſ povem</lem>
     <rdg wit="#P">Poslushei kar ti jest povem</rdg>
  </app>
</l>
```

Pri tem se vrednost atributa @wit sklicuje na identifikator opisa rokopisov in tiskov z omenjenimi verzijami pesmi, kjer denimo vrednost »P« pomeni siglo Paglovčeve pesmarice *Cantilenae variae,* sigla »L« označuje verzijo v Lavrenčičevi tiskani pesmarici itn. Ti rokopisi in tiski so navedeni v elementih <witness> in zbrani v seznamu verzij <listWit> (*witness list*).

Če je bilo variantno besedilo v eni od verzij zapisano v prozni obliki, kar pomeni, da je bilo v rokopisu členjeno na grafične vrstice ne glede na metrične verzne enote, te posebne členitve znotraj elementa za zapis variacije <rdg> doslej nismo prepisovali, čeprav shema TEI tak zapis omogoča in bi znotraj elementa <rdg> lahko uvedli potrebne prelome vrstic. V enote kritičnega aparata smo zajeli čiste besedilne razlike, ne pa členitve besedila na verzno-kitično strukturo variantnega besedila. V spodnjem primeru je potek besedila v Lavrenčičevi pesmarici zaradi grafične razporeditve stavka prilagojen, vendar tega ne prikazujemo, saj bi izpis v HTML izgubil potrebno nazornost in jasnost:



PESEM,
Ali Sdibvajne Svetiga Ignatia od Lojole,
Pruti Bogu.

1. Bug Te lubim! Bug te lubim! Ah is ſerza lubim te! Duſho mojo tebi ſhenkam; Spomin, um, volo, tebi dam: vishej vſelei Serze mojo, de bo pouno gnade tvoje. Bug Te lubim &c.

Slika1: Izsek besedila iz Lavrenčičeve pesmarice

V zapisu XML sta prva verza označena tako:

```
<l>
  <app>
    <lem wit="#F">BOG TE lu bim Bog te lu bim</lem>
    <rdg wit="#L">Bug Te lubim! Bug te lubim!</rdg>
  </app>
</l>
<l>
  <app>
    <lem wit="#F">O hi ſerza lu bim te</lem>
     <rdg wit="#L">Ah is ſerza lubim te!</rdg>
  </app>
</l>
```

Za celoten ustroj diplomatičnega in posledično tudi kritičnega prepisa Foglarjevega rokopisa je bila temeljna odločitev, da ima Foglarjev tekst v naši izdaji status t. i. glavnega ali temeljnega besedila (*base text*). Zato so njegova variantna mesta vsa označena kot leme, tj. z oznako <lem> (*lemma*), vsa variantna mesta v drugih verzijah (*readings*) pa le z oznako <rdg>, kar pomeni, da je vsa ohranjena besedilna preoddaja ali tradicija prikazana oz. v izdaji organizirana tako, da je podrejena glavnemu, tj. Foglarjevemu besedilu.

Ta odločitev je bila metodološka. Zavedamo se, da je v preostali besedilni tradiciji – bodisi pri Paglovcu, Lavrenčiču ali drugod – posamezna pesem morda zapisana bolje, bodisi v jezikovnem, verzološkem ali sploh literarnem pogledu, medtem ko je v Foglarjevem rokopisu ista pesem morda utrpela poškodbe raznih oblik besedilne kontaminacije. Kljub temu ima ta v naši izdaji status glavnega besedila, saj edicija nastaja v okviru širše raziskave baročnega pesništva na slovenskem Štajerskem; zato smo glede na ta posebni raziskovalni namen vsa variantna mesta organizirali glede na leme v Foglarjevem tekstu.

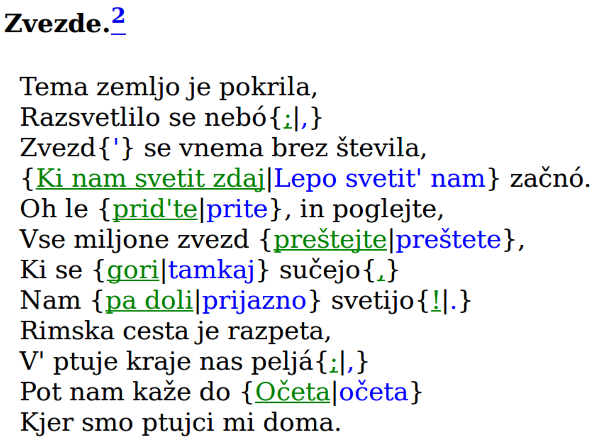
Maurerjeve rokopisne pesmarice (1754) pa se vsebinsko ne ujema z njima. Pisec Vezjak je pri prepisovanju storil očitno napako, saj je v 10. kitici prepisal peti verz iz prejšnje kitice: *»Glei da v nih ne saſspish /…/.«* Glede na varianto iz Krebsove pesmarice bi na tem mestu moral stati verz: *»Nozhem se nigdar uezh /…/.«*

[10] Pesem z naslovom Od Božje ljubezni, kjer manjka šest kitic, je v celoti zapisana v Lavrenčičevi pesmarici *Miſionske Peſme inu molitve* (1757), vendar pod naslovom Sdihvajne Svetiga Ignatia od Loyole, Pruti Bogu. Drugo verzijo Pesmi od Božje ljubezni najdemo prav tako pri Lavrenčiču (1752) in Stržinarju (1752), dve kitici lahko povežemo tudi s pesmijo De amore Dei Super omnia v Paglovčevi pesmarici (1733). Tovrstne pesmi najdemo kasneje tudi pri Redeskiniju (1775), še prej pa pri Maurerju (*To je ena lipa peisem od popovnama grivenege: Moi bug jeſt tebe Lubem / Ne Li sa nabv / Moi bug jeſt tebe slushem / ne li se iſt trahu*).

4. Orodja za prikaz in analizo besedil

Za prikaz variantnosti v besedilni preoddaji[11] pesmi v Foglarjevem rokopisu smo uporabili oziroma preizkusili troje orodij, ki imajo zelo različen nabor funkcionalnosti in izhajajo iz raznih konceptov grafičnega ponazarjanja besedilnih razlik. Poleg teh treh ima verjetno najdaljšo zgodovino orodje *Versioning machine* (VM),[12] ki nudi sicer obilo funkcionalnosti. Zanj se nismo odločili, ker bi bilo treba zapis XML izdatno prilagoditi, da bi ga VM lahko dobro prikazal. Orodja smo presojali glede na to, kako naš dokument, dosledno urejen po smernicah TEI, pretvorijo brez posebnih prilagoditev.

1. **Pretvorba XSLT.** Med samo pripravo znanstvenokritične izdaje Foglarjeve pesmarice smo največ uporabljali že omenjeno pretvorbo XSLT. Deluje po načelu, da generično pretvorbo zapisa TEI v HTML nadgradi s tem, da variantno mesto obda z zavitima oklepajema, v notranjosti pa izpiše najprej lemo, nato varianto, ločeni sta z izmenjalno navpičnico. Lema je izpisana z zelenimi črkami, varianta z modrimi. Ob dotiku miške (*hoover*) se izpiše ime verzije, na katero kaže wit/@witness. Pri pesmih A. M. Slomška iz srede 19. stoletja je šlo za manjše variacije, ki zadevajo predvsem besedne oblike, kot npr. na sliki 2.



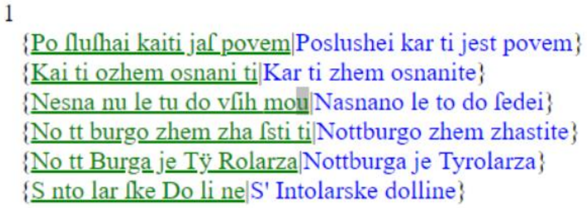
Zvezde.2

Tema zemljo je pokrila,
Razsvetlilo se nebó{;|,}
Zvezd{'} se vnema brez števila,
{Ki nam svetit zdaj|Lepo svetit' nam} začnó.
Oh le {prid'te|prite}, in poglejte,
Vse miljone zvezd {preštejte|preštete},
Ki se {gori|tamkaj} sučejo{,}
Nam {pa doli|prijazno} svetijo{!|.}
Rimska cesta je razpeta,
V' ptuje kraje nas peljá{;|,}
Pot nam kaže do {Očeta|očeta}
Kjer smo ptujci mi doma.

Slika 2: Variantnost v pesmi *Zvezde* A. M. Slomška

Baročno besedilo v Foglarjevem rokopisu je mnogo bolj zaznamovano z nestandardno rabo pravopisa in arhaičnimi ter narečnimi besednimi oblikami v variantnih vejah preoddaje. Razlik je toliko, da v večini primerov en posamezen element za variantno mesto <rdg> vsebuje kar celoten verz. V nasprotnem primeru bi morali v vsak verz uvesti povprečno po tri enote aparata <app>, kar bi lahko poleg koristi za strojno primerjavo imelo tudi slabosti za bralčevo evidenco nad razlikami. Ob dilemi, ki jo ta problem odpira, kaže opozoriti le, da sta obseg kritičnega aparata in stopnja njegove detajliranosti ali granularnosti v filologiji že od začetka kritičnega izdajanja besedil predmet diskusij, še posebej, kar zadeva razlikovanje med ravnjo zgolj ortografskih razlik ali t. i. akcidentalij in ravnjo bolj pomenotvornih razlik ali substancialij, ki segajo vsaj do Gregove teorije predloge (*theory of copy-text*) in še dlje v zgodovino filologije.[13]

Slika 3 tako prikazuje z zelenimi črkami lemo, torej verz Foglarjevega teksta, z modrimi pa verz iz rokopisa Mihaela Paglovca.



1
{Po ſluſhai kaiti jaſ povem|Poslushei kar ti jest povem}
{Kai ti ozhem osnani ti|Kar ti zhem osnanite}
{Nesna nu le tu do vſih mou|Nasnano le to do ſedei}
{No tt burgo zhem zha ſsti ti|Nottburgo zhem zhastite}
{No tt Burga je Tÿ Rolarza|Nottburga je Tyrolarza}
{S nto lar ſke Do li ne|S' Intolarske dolline}

Slika 3: Sinoptični prikaz osnovnega Foglarjevega besedila in Paglovčeve variante v izpisu HTML (*Pesem od svete Notburge*)

To orodje, ki generično pretvorbo konzorcija TEI nadgradi z barvno zaznamovanim sinoptičnim izpisom kritičnega aparata v vrstici glavnega besedila, je namenjeno preprostemu, toda filološko natančnemu prikazu besedilne variantnosti v znanstvenokritični izdaji. Pogoj za njegovo uporabo je dosledna uporaba metode vzporednega segmentiranja v zapisu TEI. Bralcu sicer ne omogoča fleksibilnosti prikaza (npr. da bi skril ali prikazal določeno verzijo teksta), vendar je kot orodje dragoceno v tem, da je dostopno kot spletna storitev,[14] brez posebnih težav pa si ga lahko tudi namestimo na svoj računalnik in ga med pripravo edicije poljubno poganjamo, ko popravljamo svoje besedilo. Njegova uporaba je idealna za prikaz besedil, kjer v posamezni enoti aparata primerjamo le dve ali tri, morda štiri verzije. Za besedilne tradicije, ki so bogatejše oziroma bolj kompleksne, bi bilo potrebno uporabiti drugo orodje; vendar kaže opozoriti, da v slovenskem slovstvenem gradivu, še zlasti starejših obdobij, tako bogatih besedilnih tradicij ne poznamo.

2. **Orodje TEI Critical Apparatus Toolbox** (TEI CAT) je spletna storitev,[15] ki jo razvija skupina pod vodstvom Marjorie Burghart. Izrecno je namenjena urednikom, ki pripravljajo kritično izdajo po Smernicah TEI z metodo vzporednega segmentiranja. Služi torej kot delovni pripomoček, s katerim avtorji znanstvenokritične izdaje med pripravo svojo edicijo preverjajo in vizualizirajo njene pomenske sestavine. Za ta namen omogoča številne funkcionalnosti, mdr. za preverjanje napak in nedoslednosti

[11] Preoddaja (nem. *Überlieferung*, *Textüberlieferung*) je termin, splošno uveljavljen na področjih klasične filologije, biblicistike, medievalistike ter v povezanih disciplinah tekstne kritike in edicijske tehnike. Uporablja se pri raziskovanju posredovanja istega besedila s prepisi iz enega rokopisa (ali tiska) v drugega (prim. Jäger, 1998).

[12] Prim. http://v-machine.org/.

[13] Temeljit zgodovinski pregled stališč, ki so se v tekstni kritiki izoblikovala glede tega vprašanja, gl. v Sahle (2013: 172–173).

[14] Prim. http://nl.ijs.si/tei/convert/, kjer moramo izbrati profil pretvorbe ZRC.

[15] V konzorciju, ki orodje razvija, sodelujejo mdr. CNRS in Univerza v Lyonu, prim. http://ciham-digital.huma-num.fr/teitoolbox/index.php.

označevanja (Burghart, 2016). Osredinili se bomo le na tiste, ki so najbolj relevantne za naše besediloslovne raziskave.

Datoteko XML uporabnik pošlje servisu, da preveri pravilnost označevanja; če je rezultat pozitiven, servis prikaže glavno besedilo ali t. i. kritični tekst izdaje. Ob vsaki enoti aparata se na zaslonu izpiše puščica, ob kliku se odpre okence z vsebino te enote aparata v klasični obliki, ki temelji na rabi desnega oglatega oklepaja: kar je levo od oglatega oklepaja, je lema, desno je varianta, zaznamovana s siglo verzije. Denimo, lema je vselej iz glavnega teksta »F« (*Foglar*), varianta je iz rokopisa »C« (*Cerkvene pesmi* …):

Son ze no lufft rengira F] sonze, semlo zira. C

Pri tem imamo možnost izbrati še številne kontrolnike, denimo, ali naj sistem prikaže prelome strani ali obarva enote aparata, ki ne vsebujejo vseh verzij, ali nasprotno, naj obarva le enote aparata, ki vsebujejo določeno verzijo itn.

Glavna funkcionalnost, ki jo omogoča TEI CAT, je vzporedni prikaz verzij. Ne glede na to, da je po Smernicah TEI priporočeno mesto za seznam verzij <listWit> v t. i. kolofonu z metapodatki <teiHeader>, sistem CAT najde <listWit> tudi drugod v dokumentu TEI (v našem primeru je umeščen v <back>) in njegove informacije smiselno razvrsti glede na sigle. Uporabnik lahko izbere, naj sistem nato vzporedno prikaže vse verzije, ali pa odkljukamo le sigle posameznih verzij, ki jih želimo vzporedno izpisati za primerjavo (prim. sliko 4).

Slabost vzporednega prikaza pri orodju TEI CAT je, da se stolpci med sabo ujemajo le na začetku datoteke, v nadaljevanju pa se razmerje lahko tudi poruši in bralec izgubi referenčno primerjavo. Druga šibka stran je, da orodja (trenutno še) ni moč prenesti na svoj računalnik in poganjati lokalno ter da ni namenjeno pripravi edicije kot javne objave, marveč služi le preverjanju v procesu njene priprave. Vendar ga poleg praktične uporabnosti za prikaz aparata odlikuje tudi to, da naredi osnovno statistiko dokumenta, ne le uporabljenih oznak TEI, temveč tudi besedila: napravi preprost, toda informativen frekvenčni seznam besedja v izdaji, pri tem pa seveda vsako pravopisno posebnost šteje kot novo obliko besede.

3. Odprtokodno orodje EVT – Edition Visualization Technology je namenjeno izdelavi in objavi znanstvenokritičnih izdaj v zapisu TEI; tudi v tem primeru se zahteva označevanje kritičnega aparata z metodo vzporednega segmentiranja.[16] Skupina, ki jo vodi Roberto Rosselli del Turco, je EVT zasnovala z izrecno ambicijo, da premosti vrzel med Smernicami TEI kot prvovrstnim standardom za izdelavo kompleksnih filoloških del, kot so kritične izdaje, in težavami, ki jih filologi imajo, ko želijo svojo edicijo, zapisano v TEI, vizualizirati in objaviti na spletu (Rosselli del Turco, 2015).

EVT odpremo in uporabljamo kot spletno stran v svojem brskalniku, bodisi lokalno ali po spletu. Ker je zasnovana kot dinamično okolje, za nadgraditev možnosti HTML uporablja Javascript. Nudi paleto možnosti za prikaz kritičnega teksta in variant, tudi vzporedni prikaz verzij in razne detajle glede posamezne enote aparata, kar je moč poljubno izbirati s preklapljanjem in sprotnim generiranjem raznih prikazov (prim. Sliko 5). Med opcijami, ki bi bile za tip izdaj, kakršne združuje knjižnica eZISS, dobrodošle, so podpora za dinamičen prikaz digitalnega faksimila, podpora poimenovanih entitet in njihovih seznamov, npr. krajevnih in osebnih imen ipd. (seveda morajo biti prej primerno zapisane v TEI) in visoka raven prilagajanja nastavitev projektnim potrebam.

Med slabostmi orodja EVT z gledišča uporabnosti za izdaje eZISS je na prvem mestu ta, da je EVT koncipirana v klasičnem pojmovnem svetu zahodnoevropske filologije, kjer urednik običajno predstavi tekst enega izbranega rokopisa in ga opremi z manjšo ali večjo skupino verzij tega istega teksta, podano v obliki kritičnega aparata. V slovenskem besedilnem izročilu pa, če izvzamemo redke primere, vse do 18. stoletja *nimamo* istega rokopisnega besedila v dveh ali več verzijah, marveč največkrat le v obliki enega samega preživelega rokopisa, t. i. *codex unicus*. Ta postane edini predmet kritične izdaje, ki ga je potrebno izčrpno predstaviti zlasti z metodo razlikovanja med njegovo diplomatično in kritično besedilno fakturo, kar je posebnost filologije, kakršna je slovenska. Zato bi bilo tudi tako kakovostno in kompleksno orodje, kakor je EVT, potrebno šele prilagoditi ali predelati, da bi prikazovalo vzporedni prikaz diplomatičnega in kritičnega prepisa *istega* besedila (ki sploh ne bo imelo kritičnega aparata, v izdelavo katerega je skupina EVT zastavila največ truda, če se nam nista ohranili vsaj dve verziji besedila).

Foglarjeva pesmarica pa je ravno posebej zahteven primer, ki s skupino šestih doslej evidentiranih verzij besedilne preoddaje zahteva klasičen zahodnoevropski tip kritične edicije, z razlikovanjem med diplomatično in kritično podobo glavnega teksta pa slovenski filološki profil znanstvenokritične izdaje. Tej potrebi kaže v bodoče prilagoditi tudi reševanje njenega bralnega prikaza.

5. Sklep

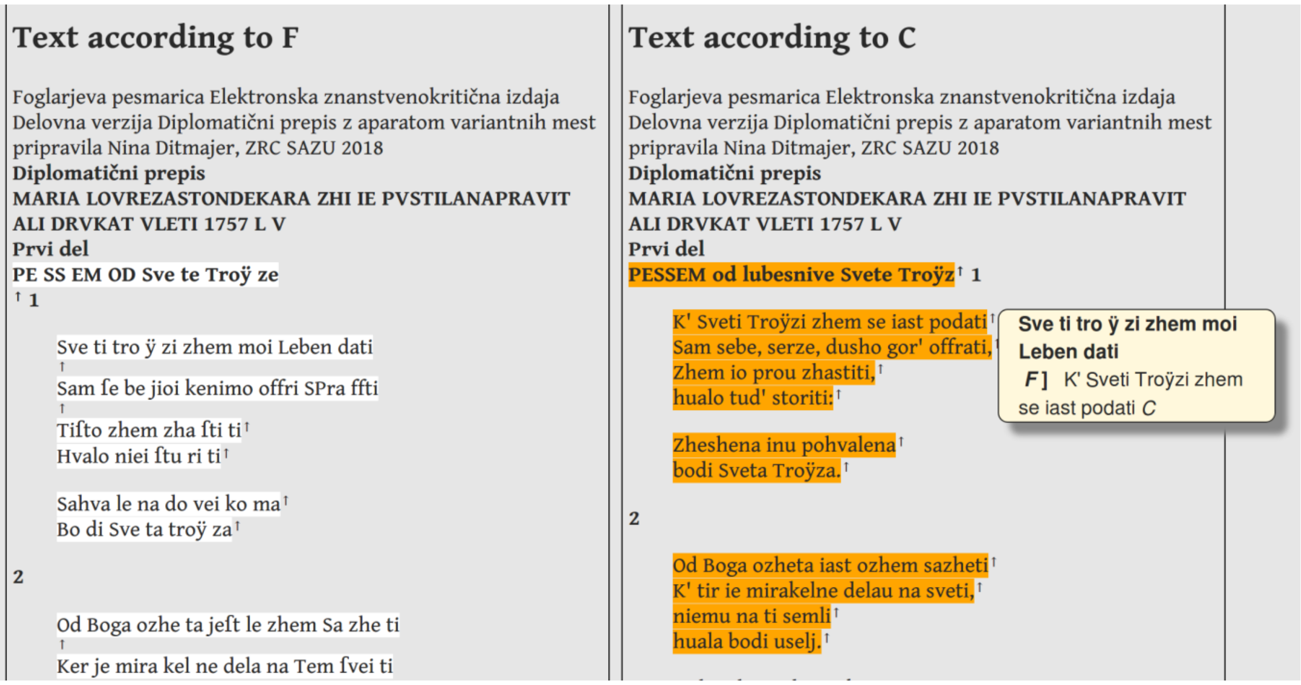
Pesniška forma ali vezana beseda je zaradi verznometrične sestave mnogo bolj razčlenjeno in strukturirano besedilo kakor proza. Zato je tudi problematika verzij in variantnih mest pesniškega teksta lahko bolj kompleksna kakor problematika variantnosti proznih besedil. V primeru besedilne tradicije Foglarjevega rokopisa bi bil večji izziv podrobno zapisovanje in prikaz verznih oz. vrstičnih prelomov v verzijah ohranjene preoddaje, čemur smo se namenoma izognili, da bi pozornost v kritičnem aparatu usmerili na bistvene besedilne pomenske razlike. Poudarimo lahko naslednje.

1. Kritični aparat je zasnovan tako, da je popolnoma orientiran na Foglarjev rokopis kot glavni tekst, njegove enote (773 elementov <app>) so zajete v aparat kot leme z oznakami <lem>. Tako bo kritični aparat izdaje bolje osvetlil načine, kako so bili pesniški teksti iz raznih slovenskih dežel recipirani in spremenjeni na slovenskem Štajerskem, in tudi nasprotno, kako se je štajerska, marijinoceljska pesem spremenila, ko je prispela na Gorenjsko v Gorje.

[16] Orodje EVT je prosto dostopno za prenos na osebni računalnik z naslova in je enostavno za namestitev.

2. Za prikaz aparata na zaslonu smo uporabili troje orodij, od katerih je za objavo elektronske izdaje trenutno primerna le naša pretvorba XSLT. Ta izdela statično spletno stran z obarvanim izpisom variantnih mest v vrstici. Orodje TEI CAT zelo dobro podpira nekatere potrebe preverjanja in testiranja kritične izdaje v času nastajanja, vendar ni namenjeno končnemu publiciranju. Vse razsežnosti uredniškega nadzora nastajajoče izdaje kot tudi njene končne objave na spletu bi lahko adekvatno združilo orodje EVT, vendar bi ga bilo potrebno prilagoditi še za vzporedni prikaz diplomatičnega in kritičnega prepisa, mu dodati kolofon TEI ter modul za prikaz komentarja oziroma študij, verjetno tudi modul za prikaz besedišča izdanega besedila itn.

3. Problematika variantnosti starejših pesniških besedil bolj kot besedila dosedanjih izdaj vzbuja podobo, da razna orodja ustrežejo potrebam po raznih funkcionalnostih, nobeno orodje pa ne odgovori vsem potrebam. S tem se odpira (ne povsem nov) horizont, v katerem se dodatno povečuje vrednost kanoničnega zapisa naše izdaje v XML TEI, saj jo lahko nato procesiramo v raznolikih, tudi razvijajočih se orodjih glede na razne potrebe prezentacije in raziskovanja. Nobeden od teh prikazov nemara ni povsem dokončen in kanoničen. Stremeti sicer kaže k temu, da tudi za slovensko besedilno tradicijo dosežemo idealno metodo prikaza in izdajanja besedil. Temeljno veljavo pa ima slej ko prej le datoteka v zapisu XML TEI, ki jo je moč prikazovati na nove in nove načine, ne nazadnje pa tudi združevati z drugimi v TEI zapisanimi znanstveno-kritičnimi izdajami v enovito kodirano in medsebojno povezljivo digitalno knjižnico.



Text according to F

Foglarjeva pesmarica Elektronska znanstvenokritična izdaja
Delovna verzija Diplomatični prepis z aparatom variantnih mest
pripravila Nina Ditmajer, ZRC SAZU 2018
Diplomatični prepis
MARIA LOVREZASTONDEKARA ZHI IE PVSTILANAPRAVIT ALI DRVKAT VLETI 1757 L V
Prvi del
PE SS EM OD Sve te Troÿ ze
↑ **1**

Sve ti tro ÿ zi zhem moi Leben dati
↑
Sam ſe be jioi kenimo offri SPra ffti
↑
Tiſto zhem zha ſti ti↑
Hvalo niei ſtu ri ti↑

Sahva le na do vei ko ma↑
Bo di Sve ta troÿ za↑

2

Od Boga ozhe ta jeſt le zhem Sa zhe ti
↑
Ker je mira kel ne dela na Tem ſvei ti

Text according to C

Foglarjeva pesmarica Elektronska znanstvenokritična izdaja
Delovna verzija Diplomatični prepis z aparatom variantnih mest
pripravila Nina Ditmajer, ZRC SAZU 2018
Diplomatični prepis
MARIA LOVREZASTONDEKARA ZHI IE PVSTILANAPRAVIT ALI DRVKAT VLETI 1757 L V
Prvi del
PESSEM od lubesnive Svete Troÿz↑ **1**

K' Sveti Troÿzi zhem se iast podati↑
Sam sebe, serze, dusho gor' offrati,↑
Zhem io prou zhastiti,↑
hualo tud' storiti:↑

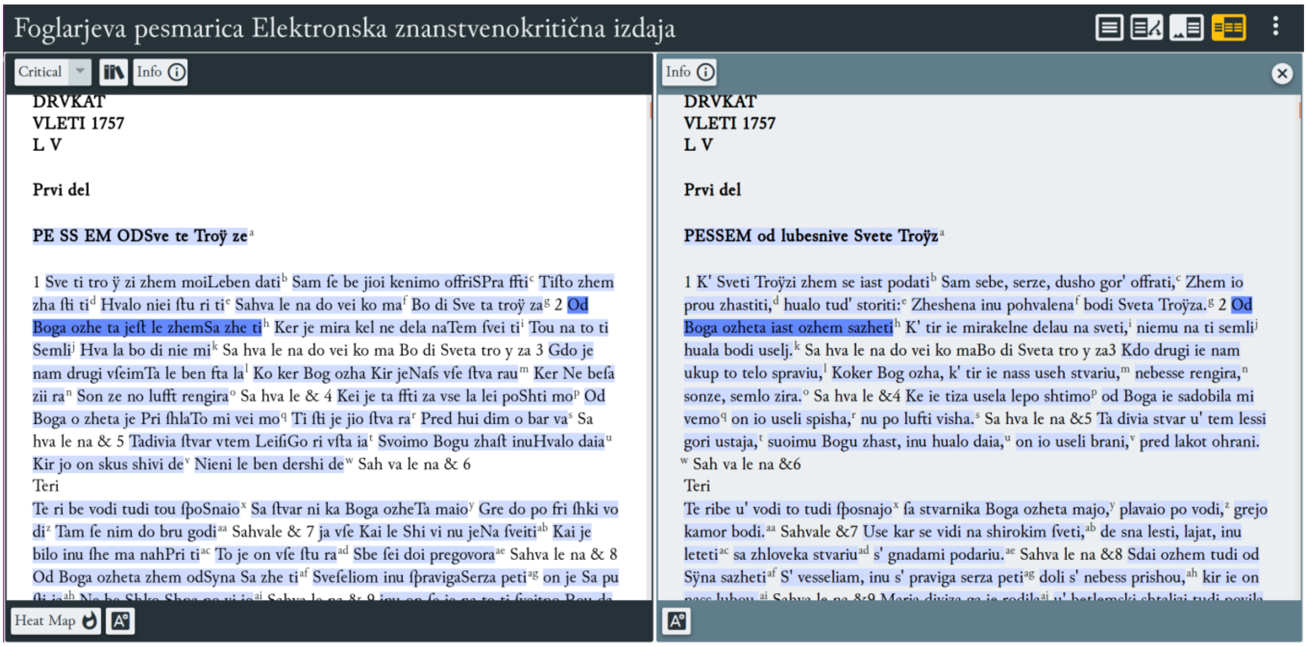
Zheshena inu pohvalena↑
bodi Sveta Troÿza.↑

2

Od Boga ozheta iast ozhem sazheti↑
K' tir ie mirakelne delau na sveti,↑
niemu na ti semli↑
huala bodi uselj.↑

Sve ti tro ÿ zi zhem moi Leben dati
F] K' Sveti Troÿzi zhem se iast podati *C*

Slika 4: Orodje TEI CAT omogoča uredniku izdaje vzporedni izpis glavnega teksta in izbranih verzij.



DRVKAT
VLETI 1757
L V

Prvi del

PE SS EM ODSve te Troÿ ze[a]

1 Sve ti tro ÿ zi zhem moiLeben dati[b] Sam ſe be jioi kenimo offriSPra ſſti[c] Tiſto zhem zha ſti ti[d] Hvalo niei ſtu ri ti[e] Sahva le na do vei ko ma[f] Bo di Sve ta troÿ za[g] 2 Od Boga ozhe ta jeſt le zhemSa zhe ti[h] Ker je mira kel ne dela naTem ſvei ti[i] Tou na to ti Semli[j] Hva la bo di nie mi[k] Sa hva le na do vei ko ma Bo di Sveta tro y za 3 Gdo je nam drugi vſeimTa le ben ſta la[l] Ko ker Bog ozha Kir jeNaſs vſe ſtva rau[m] Ker Ne beſa zii ra[n] Son ze no lufft rengira[o] Sa hva le & 4 Kei je ta ſſti za vse la lei poShti mo[p] Od Boga o zheta je Pri ſhlaTo mi vei mo[q] Ti ſti je jio ſtva ra[r] Pred hui dim o bar va[s] Sa hva le na & 5 Tadivia ſtvar vtem LeiſiGo ri vſta ia[t] Svoimo Bogu zhaſt inuHvalo daia[u] Kir jo on skus shivi de[v] Nieni le ben dershi de[w] Sah va le na & 6
Teri
Te ri be vodi tudi tou ſpoSnaio[x] Sa ſtvar ni ka Boga ozheTa maio[y] Gre do po fri ſhki vo di[z] Tam ſe nim do bru godi[aa] Sahvale & 7 ja vſe Kai le Shi vi nu jeNa ſveiti[ab] Kai je bilo inu ſhe ma nahPri ti[ac] To je on vſe ſtu ra[ad] Sbe ſei doi pregovora[ae] Sahva le na & 8 Od Boga ozheta zhem odSyna Sa zhe ti[af] Sveſeliom inu ſpravigaSerza peti[ag] on je Sa pu

DRVKAT
VLETI 1757
L V

Prvi del

PESSEM od lubesnive Svete Troÿz[a]

1 K' Sveti Troÿzi zhem se iast podati[b] Sam sebe, serze, dusho gor' offrati,[c] Zhem io prou zhastiti,[d] hualo tud' storiti:[e] Zheshena inu pohvalena[f] bodi Sveta Troÿza.[g] 2 Od Boga ozheta iast ozhem sazheti[h] K' tir ie mirakelne delau na sveti,[i] niemu na ti semli[j] huala bodi uselj.[k] Sa hva le na do vei ko maBo di Sveta tro y za3 Kdo drugi ie nam ukup to telo spraviu,[l] Koker Bog ozha, k' tir ie nass useh stvariu,[m] nebesse rengira,[n] sonze, semlo zira.[o] Sa hva le &4 Ke ie tiza usela lepo shtimo[p] od Boga ie sadobila mi vemo[q] on io useli spisha,[r] nu po lufti visha.[s] Sa hva le na &5 Ta divia stvar u' tem lessi gori ustaja,[t] suoimu Bogu zhast, inu hualo daia,[u] on io useli brani,[v] pred lakot ohrani.[w] Sah va le na &6
Teri
Te ribe u' vodi to tudi ſposnajo[x] ſa stvarnika Boga ozheta majo,[y] plavaio po vodi,[z] grejo kamor bodi.[aa] Sahvale &7 Use kar se vidi na shirokim ſveti,[ab] de sna lesti, lajat, inu leteti[ac] sa zhloveka stvariu[ad] s' gnadami podariu.[ae] Sahva le na &8 Sdai ozhem tudi od Sÿna sazheti[af] S' vesseliam, inu s' praviga serza peti[ag] doli s' nebess prishou,[ah] kir ie on

Slika 5: Orodje EVT omogoča razne dinamične načine prikaza kritične izdaje, mdr. prikaz glavnega teksta na levi in izbrane verzije na desni.

6. Literatura

Marjorie Burghart. 2016. The TEI Critical Apparatus Toolbox: Empowering Textual Scholars through Display, Control, and Comparison Features. *Journal of the Text Encoding Initiative.* (vol. 10) https://journals.openedition.org/jtei/1520#article-1520.

Izidor Cankar. 2007. *S poti. Elektronska znanstvenokritična izdaja.* Ur. M. Ogrin, L. Vidmar, T. Erjavec, Elektronske znanstvenokritične izdaje slovenskega slovstva, ZRC SAZU, IJS. http://nl.ijs.si/e-zrc/izidor/.

Nina Ditmajer. 2017. Romarske pesmi v Foglarjevi pesmarici (1757–1762). V: *Rokopisi slovenskega slovstva od srednjega veka do moderne*, str. 75–82. Znanstvena založba Filozofske fakultete. http://centerslo.si/wp-content/uploads/2017/10/Obdobja-36_Ditmajer.pdf.

Nina Ditmajer in Matija Ogrin. 2018. Foglarjeva pesmarica. Ms 123. V: *Register slovenskih rokopisov 17. in 18. stoletja.* http://ezb.ijs.si/nrss/.

Gerhard Jäger. 1998. Uvod v klasično filologijo. Ljubljana: Študentska založba.

Matija Ogrin in Tomaž Erjavec. 2009. Ekdotika in tehnologija: elektronske znanstvenokritične izdaje slovenskega slovstva. *Jezik in slovstvo.* 54/6, str. 57–72.

Matija Ogrin in Tomaž Erjavec. 2009. Elektronske znanstvenokritične izdaje slovenskega slovstva eZISS: metode zapisa in izdaje. V: *Infrastruktura slovenščine in slovenistike* (Obdobja 28), 123–128. Ljubljana, Znanstvena založba Filozofske fakultete. http://www.centerslo.net/files/file/simpozij/simp28/Erjavec_Ogrin.pdf.

Matija Ogrin in Andrejka Žejn. 2016. Strojno podprta kolacija slovenskih rokopisnih besedil: variantna mesta v luči računalniških algoritmov in vizualizacij. V: *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, str. 124–132. ZIFF in IJS.

Patrick Sahle. 2013. *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 1: Das typografische Erbe.* BoD, Norderstedt.

Roberto Rosselli Del Turco, Giancarlo Buomprisco, Chiara Di Pietro, Julia Kenny, Raffaele Masotti, and Jacopo Pugliese. 2014. Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions. *Journal of the Text Encoding Initiative.* (vol. 8) https://journals.openedition.org/jtei/1077.

TEI Consortium, eds. 2018. TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Version 3.3.0]. [31 Jan. 2018]. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Zakaj ne z eno poizvedbo hkrati po različnih korpusih? (Troje korpusnih preverb pod primerjalnim drobnogledom)

Helena Dobrovoljc,[*] Urška Vranjek Ošlak[**]

*Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
Novi trg 4, 1000 Ljubljana
Fakulteta za humanistiko Univerze v Novi Gorici
Vipavska 13, 5000 Nova Gorica
helena.dobrovoljc@zrc-sazu.si

**Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
Novi trg 4, 1000 Ljubljana
urska.vranjek@zrc-sazu.si

Povzetek

Prispevek predstavlja utemeljitev potrebe po skupnem vmesniku za trenutno dostopne korpuse slovenskega jezika. Potreba je orisana s pomočjo preverb izbranih jezikovnih vprašanj: (1) uresničevanja preglasa pri pregibanju samostalnikov moškega spola, ki se končujejo na izglasni govorjeni c, (2) aktualnost normativno-stilističnih napotkov pri rabi predloga *prek* in vezljivosti glagola *upravljati*; (3) socialno- in funkcijskozvrstno prevrednotenje samostalnika *tenisač* in pridevniškega sopomenskega para *duševen – mentalen*. Zaradi različnosti besedil, na katerih temeljijo posamezni korpusi sodobne slovenščine, se tudi rezultati poizvedb pričakovano razlikujejo. Pri pravopisnih in leksikografskih poizvedbah, ki temeljijo na ovrednotenju obstajajočih različic jezikovne rabe, bi integracija korpusov s skupnim iskalnim vmesnikom olajšala poizvedbe in poleg tega ponujala vsebinske in statistične primerjalne analize.

One query to search different corpora? (Comparative Scrutiny of Three Corpora Checks)

This paper presents an argumentation of the need for a common interface of the currently available Slovenian corpora. This need is outlined through corpus queries on selected language issues: (1) realization of inflectional progressive vowel fronting (*o* to *e*) in masculine nouns ending in pronounced final c, (2) relevance of normative and stylistic guidelines in the use of the preposition *prek* and in the valency of the verb *upravljati*; (3) re-valuation of register for the noun *tenisač* and the adjective synonyms *mentalen – duševen*. Due to the diversity of texts on which individual corpora of modern Slovenian are based, the results of these queries differ as expected. For orthographic and lexicographic queries based on the evaluation of existing versions of language use, the integration of corpora with a common search interface would facilitate queries and in addition provide for substantive and statistical comparative analysis.

1. Uvod

Sodobne zadrege jezikovnih uporabnikov jezikoslovje usmerja z napotili, ki se vsebinsko opirajo na jezikoslovna prepričanja zadnjih desetletij prejšnjega stoletja. Pri posodobitvi jezikovnih priročnikov knjižnega jezika in ugotavljanju normativne vrednosti jezikovnih pojavov sodelavci programske skupine »Slovenski jezik v sinhronem in diahronem pogledu« proučujemo aktualne jezikovnosistemske uresničitve v različnih položajih ter izkoriščamo jezikovnotehnološke možnosti pri preverjanju jezikovne rabe po korpusih sodobne slovenščine.

Empirični pristopi pri ugotavljanju rabe jezikovnih prvin so v elektronski dobi dobili nove perspektive, ki jih tradicionalne metode niso omogočale. Korpusno jezikoslovje ne ponuja le obsežnih zbirk raznovrstnih besedil, ki jih je mogoče na različne načine povezovati, temveč prinaša možnosti posploševanja, predvidevanja in napovedovanja jezikovnega vedenja in vzorcev. Od vsega najbolj pomembna je opustitev pristranskosti in subjektivnosti, ki izvira iz favoriziranja jezikovnega čuta posameznika.

Zaradi zvrstne in obsegovne različnosti ter specializiranosti aktualnih slovenskih korpusov se tudi rezultati poizvedb pričakovano razlikujejo, kar je – kot bo razvidno iz nadaljevanja prispevka – relevantno pri vseh tipih poizvedb, ki temeljijo na presojah pogostnosti, predvsem pa pri poizvedbah, ki vplivajo na leksikografsko socialno- in funkcijskozvrstno vrednotenje oz. kategoriziranje izbranih jezikovnih prvin.

Korpusne poizvedbe iste jezikovne prvine v različnih korpusih naj bi odsevale sorazmerno različnost v odvisnosti od različnih gradivskih zaledij. V korpusnem jezikoslovju obstaja precej primerjalnih korpusnih raziskav, npr. raziskovalci so opozarjali predvsem na probleme pri vzporejanju korpusov z različnim obsegom (Górski, 2007: 119–124; Sharoff, 2010) ali na obsegovne razlike med poizvedbami po različnih korpusih (Möhrs idr., 2017). Precej več študij je bilo usmerjenih v utemeljenost jezikovnotehnoloških analiz oz. poudarjanje prednosti računalniško podprte interpretacije jezikovnih vzorcev v različnih ali velikih korpusih, ki omogočajo primerjavo med uresničitvami v zvrstno različnih besedilih, hkrati pa nakazujejo tudi na novosti (Biber, 1998).

Pričujoča raziskava se približuje tematski primerjavi korpusov Gigafida in slWaC (Logar Berginc in Ljubešić, 2013), ki razkriva vrzeli pri primerjavi lem v korpusu Gigafida kot korpusu tiskanih oz. sekundarno digitaliziranih besedil glede na korpus spletnih besedil slWaC. Te seveda nakazujejo tudi potencialne nadaljnje razlike pri slovničnih, pravopisnih in leksikografskih poizvedbah.

2. Namen članka

V prispevku želimo prikazati postopek preverbe treh relevantnih jezikoslovnih pojavov, za katere sodobna jezikovna praksa izkazuje nekoliko drugačno vedenje od tradicionalnega, in sicer v najbolj aktualnih korpusih slovenskega jezika. S tem bo za razliko od dosedanjih študij prikazano, kako izbira korpusa vpliva na poizvedbo in koliko poizvedb je potrebnih za relevantnost podatka o aktualni rabi. Pri tem bo upoštevano dejstvo, da je referenčni korpus Gigafida, v katerem so zadnja besedila iz leta 2011, za raziskave novejše leksike (npr. *brexit*, *finteh*, *slimacid* ...) in premikov v jezikovni praksi zadnjega desetletja vse bolj neuporaben.

V poglavju »Predstavitev korpusov« so orisane temeljne značilnosti korpusov in njihove razlikovalne lastnosti, zaradi katerih je mogoče pričakovati razlike pri rezultatih. V poglavju »Predstavitev preverb« so predstavljena tri področja leksikološko-normativne obravnave, za katera predvidevamo, da bi pregled njihove rabe v naštetih korpusih pokazal na bistvena razhajanja, ki vplivajo na vrednotenje in kategorizacijo jezikovnih prvin. Gre za področja socialne stratifikacije, normativnosti in empiričnih pristopov pri ugotavljanju specifik izrazne ravnine jezika, ki so se v zadnjem času korenito spremenila.

S primerjavo poizvedb (poglavje »Raziskava«) po izbranih gradivskih korpusih želimo pokazati na potrebo po integraciji orodij ter na interpretativne možnosti, ki jih raznolikost korpusov omogoča.

3. Predstavitev korpusov

Pri izbranih preverbah, ki jih predstavljamo v nadaljevanju, bomo izvedli poizvedbe po naslednjih korpusih:[1]

3.1. Referenčni korpus Gigafida

Referenčni korpus Gigafida (v1.1 DeDup – referenčni, dedupliciran) z več kot 918 milijoni pojavnic vključuje besedila iz tiskanih medijev ter le 15 odstotkov spletnih besedil. Besedila segajo časovno v obdobje 1991–2011, kot navajata Logar Berginc in Ljubešič (2013), pa je večina pojavnic iz obdobja po letu 2000.

3.2. Korpus slWaC

Korpus slWaC (v.2.1) vsebuje 749 milijonov pojavnic s spletnih strani, ki nosijo domeno .si (Ljubešić, Erjavec, 2014). Pregledana je bila različica iz leta 2014, ki časovno pokriva obdobje do junija 2014 (Erjavec, Ljubešić, Logar, 2015).

3.3. Korpus KAS

Korpus KAS (v0.2) je korpus z več kot 771 pojavnicami iz zaključnih akademskih del (diplomske, magistreske, doktorske naloge) iz repozitorijev slovenskih univerz iz obdobja 2000–2015. Ker večina slovenskih akademskih ustanov od kandidatov, diplomantov, magistrandov in doktorandov zahteva lektorirana besedila, so to jezikovno korigirana in zgoščena strokovna oz. znanstvena besedila (Erjavec idr. 2016).

3.4. Korpus Janes

Korpus Janes (v.1.0) vsebuje približno 189 milijonov pojavnic oz. 13.000.000 besedil slovenske računalniško posredovane komunikacije (*computer mediated communication* – CMC). Besedila, vključena v korpus, so bila objavljena v obdobju 2001–2017, razlikujejo pa se glede na to, katero obliko CMC zajamemo (najstarejši so forumski viri, najbolj sveži pa so tviti) (Fišer, Erjavec, Ljubešić, 2016).

Prikazati želimo, da bi pri pravopisnih in leksikografskih poizvedbah, ki temeljijo na ovrednotenju obstajajočih različic jezikovne rabe, integracija korpusov s skupnim iskalnim vmesnikom in možnost statistične primerjave rezultatov olajšala poizvedbe in poleg tega ponujala možnost vsebinske analize.

4. Predstavitev preverb

Razlike pri poizvedbah po različnih korpusih bodo prikazane ob:
(1) prikazu izjem pri **preglaševanju** končniškega in priponskega *o* v *e* pri samostalnikih moškega spola z izglasnim [c] (npr. *Leibnitz – Leibnitzov*, *Leibnitzev*, *Leibničev*);
(2) odpravljanju zastarelih **stilističnih** zapovedi (*prek*/*preko*, *upravljati z*/s *čim*);
(3) utemeljevanju sprememb pri **funkcijsko-** in **socialnozvrstni kategorizaciji** besedja v pomenskorazlagalnih slovarjih (*tenisač*, *mentalen – duševen*).

Vse nakazane zadrege so izbrane kot aktualna vprašanja slovenske normativistike, saj se mdr. izkazujejo v vprašanjih jezikovnih uporabnikov, zastavljenih v Jezikovni svetovalnici Inštituta za slovenski jezik Frana Ramovša pri ZRC SAZU, ki jo uporabljamo kot vir za ugotavljanje perečih oz. aktualnih uporabniških zadreg.

[1] Do vseh korpusov smo dostopali s pomočjo orodja NoSketch Engine, https://www.clarin.si/noske/, od koder so povzeti tudi podatki o številu pojavnic.

4.1. Končniško in obrazilno preglaševanje

V slovenščini naj bi samoglasniško preglaševanje oziroma prilikovanje samoglasnikov za funkcijsko mehkimi soglasniki dosledno uresničevali pri vseh samostalnikih moškega spola z izglasnim [c, j, č, ž, š, dž], pa tudi pri tistih, pri katerih navedenemu izglasju sledi o (*Srečo*, *Mišo*, *Braco*). Enako velja za samostalnike srednjega spola (*sonce* – tudi v imenovalniku), za pridevniško sklanjatev srednjega spola in za nedoločniško glagolsko pripono *-ova-* nasproti *-eva-* (*kupovati – bojevati*, *kupčevati*) (Toporišič, [4]2004: 265–266).

V *Slovenskem pravopisu* 2001 (dalje SP 2001) je navedeno slovnično določilo uveljavljeno kot normativna rešitev z eno samo izjemo, in sicer z določilom: »Za *ts* pišemo samo *-ov*: *Smuts – Smutsov*.« (SP 2001: § 957) Že v slovarskem delu istega pravopisa opazimo razhajanje s pravili ob zgledu *Keats* s pisno dvojnico *Keatsev* in *Keatsov* in z enotno fonetično uresničitvijo [kíčev-].[2] Drugi pravopisni odmik od realne prakse (in pravil) pa je povezan s preoblikovanjem izglasja imen, ki tvorijo pridevniško obliko, npr. *Clausewitz – Clausewitzev/Clausewičev*, pri čemer dvojnične oblike *Clausewičev* (§ 957) v rabi ne zasledimo.

Opisana glasovna posebnost se zlasti pri nekaterih tujih imenih moškega spola (*Franz*, *Leibnitz*) in tudi domačih vzdevkih (*Rac*) ter imenih z izglasnim *o* (*Joco*, *Braco*) ne uresničuje.[3]

Pri preverbi rabe se bomo osredinili na tuja imena z veččrkovnim izglasjem, pri katerih naj bi se (v normativnih priročnikih) preglas uresničeval, a raba izkazuje izjeme.

Sodobno usmerjanje jezikovne rabe, ki se odraža v novem pojmovanju normativnosti, ne sledi zgolj sistemskim možnostim, temveč upošteva tudi načela **izročila** (tradicije) in **jezikovne rabe**:

(1) Glede na **izročilo** v normativnih priročnikih do SP 2001 preglaševanje ni bilo nikoli izpeljano brezizjemno. Tudi določilo v *Slovenski slovnici* ([4]2004: 196) Jožeta Toporišiča je glede preglasa zadržano: »če se končuje na *c*, se ta navadno premenjuje s *č* in tudi dobiva *-ev*, če pa se ne premeni (navadno pri prevzetih besedah), pa *-ev* ni obvezen, zato *stric – stričev* proti *Horac – Horačev* in *Horacev*/*Horacov*«. Kljub temu je pravopisni slovar (SP 2001) dvojnice iz slovnice ukinil: *Horác – Horáčev*.

(2) Opazovanje **rabe** pravi, da na neupoštevanje kodifikacije zagotovo vpliva opazna razlika med knjižnim in govorjenim jezikom, pri katerem se premena zadnjejezičnega *o* v sprednjejezični sredinski samoglasnik *e* v slovenskih regijah od govora do govora razlikuje (Smole 1997).

4.2. Normativna in stilistična napotila na (obliko)skladenjskem področju

Z vse pogostejšo rabo skladenjsko prevzetih in stilistično zaznamovanih zgradb – tudi v novih skladenjskih položajih – prihaja do pomenskih in kategorialnih sprememb ter premikov, ki spreminjajo normativna načela, če slednja pojmujemo kot ovrednotenje jezikovnih pojavov glede na rabo v knjižnem jeziku.

4.2.1. *upravljati kaj* oz. *upravljati z/s čim*

Med dolgo »vzdrževanimi« stilističnimi napotki je zavračanje vezave glagola *upravljati* s predmetom v tožilniku (*upravljati kaj*) in iskanje primernejših vzporednic (npr. *upravljati krmilo → usmerjati*). Vezava glagola *upravljati* z orodnikom (*upravljati s/z čim*) pa se vse od SP 1962 uvršča med skladenjske zgradbe, ki so prepovedane oz. od SP 2001 odsvetovane za rabo v knjižnem jeziku.[4] Nekoliko blažjo omejitev prinaša *Slovar slovenskega knjižnega jezika* (dalje SSKJ), ki vezavo s tožilnikom »rehabilitira«, saj namesto »nepravilne« zveze *upravljati s skladom* usmerja k vezavi s tožilnikom (*upravljati sklad*). V drugi izdaji SSKJ (dalje SSKJ2 2014) sestavljavci ne govorijo več o nepravilnosti vezave z orodnikom, le usmerjajo k prednostni vezavi s tožilnikom.

4.2.2. *prek*/*preko*/*po*

Tradicionalna stilistika, ki je upoštevana v SSKJ, je rabo predloga *prek* (*preko*) v pomenu ʽizražanje sredstva, posrednika' zavračala, uporabnike pa so navadno napotili k predlogu *po*: *javnost je obvestil prek radia – javnost je obvestil po radiu*. V SSKJ se neposrednemu normativnemu vrednotenju sicer izognejo z uporabo oznake »neustaljeno« (neustalj.); ta zaznamuje jezikovno prvino, ki se »kljub dosedanjim prepovedim dosti uporablja« (SSKJ, Uvod § 157).

Nenavadnost te slovarske oznake izvira iz normativnega ovinka: avtorji slovarja so ugotovili, da se jezikovna prvina razmeroma pogosto uporablja, pri tem pa ni pojasnjen (niti v uvodnih poglavjih niti v spremljajoči literaturi) značaj te prepovedi, torej kdo in kje[5] obravnavano prvino prepoveduje. Kot nadrejena sopomenka predlogu *prek* je predlagan predlog *po*:

> **prek** [...] neustalj. *za izražanje sredstva*, *posrednika; po:* pismo so poslali prek kurirja; prek radia vplivati na javno mnenje (SSKJ)

SP 2001 to omejitev stopnjuje in rabo predloga *prek* v tem pomenu označuje kot prepovedano, svetuje pa nadomestilo s predlogom *po*:

[2] Pravilo bi zahtevalo nekaj dopolnil, npr. opozorilo glede neenotnega uresničevanja pisnih izglasnih sklopov *-ts* in *-ds* kot [c] oz. kot [ts], saj SP 2001 take rešitve uveljavlja, npr. pri *Massachúsetts* [mesečjusets] in *Yeats* -a [jêjts] nasproti *shórts* [šorc].

[3] Kot je razvidno iz raznovrstnih vprašanj v Jezikovni svetovalnici, imajo uporabniki ob njej občutek negotovosti, npr.: Kako je prav: »Francova« ali »Frančeva« sestra? (Dobrovoljc, 2015); Pregibanje imen dveh avstrijskih politikov: »Kurz« in »Van der Bellen« (Dobrovoljc, Lengar Verovnik, 2018).

[4] V SP 1962 je zgradba *upravljati z/s čim* označena z votlim krožcem, kar pomeni, da gre za »večjo besedno in slogovno spako in najhujši nebodigatreba«, *upravljati kaj* pa je označena kot »nepotrebna ali nelepa ali v nasprotju z duhom slovenskega jezika«.

[5] Vprašanje, ki se zastavlja, je povezano predvsem s provenienco prepovedi: ali gre za pravopis kot uradni kodifikacijski priročnik ali za individualne presoje jezikoslovcev (Gradišnik, pisci brusov, tj. normativnih priročnikov s prepovedovalno črno-belo logiko, npr. Sršen, ipd.)?

prek [...] poslati pismo •~ kurirja *po kurirju* (SP 2001)

V SSKJ2 (2014) je normativna oznaka »neustaljeno« odstranjena, slovarski prikaz še vedno usmerja na predlog *po*, zgledi so nespremenjeni:

prek [...] neustalj. *za izražanje sredstva*, *posrednika; po*^{2}: pismo so poslali prek kurirja; prek radia vplivati na javno mnenje (SSKJ2)

Vendar pa pregled rabe priča, da narašča pogostnost predloga *prek* v zgradbah, ki opisujejo dejavnosti, povezane s spletnimi oz. internetnimi storitvami, npr. *prodaja prek spleta*. V teh kontekstih se pojavljajo zgledi kot *pogovarjati se prek Skypa*; *sporočiti prek interneta*, ki še niso obremenjeni s normativnostjo, uporabniki pa jih pojmujejo kot nevtralne (Dobrovoljc, 2014), zato je take primere smiselno preučiti na novo.

4.3. Socialnozvrstna in funkcijskozvrstna kategorizacija

Leksika v slovenskem jeziku se nenehno in tudi zelo opazno spreminja z vidika pogostnosti, medtem ko njeno stilno prevrednotenje opazujemo redkeje, morda še najbolj očitno v primerih, ko v jeziku soobstajata pomensko prekrivni besedi in se sčasoma začneta specializirati za omejeno zvrstno in stilno uporabo.[6]

4.3.1. *tenisač* (šport. žarg.)

Tako je npr. beseda *tenisač* v prvi in drugi izdaji SSKJ veljala za *žargonsko*, danes pa jo obravnavamo kot nevtralno. Ž argonizmi, ki v SSKJ2 (Uvod § 142) označujejo »strokovno pogovorno leksiko, ki **ni znana širšemu krogu**, temveč predvsem poznavalcem področja« (poud. a.), se danes že zelo razširjeno uporabljajo v knjižnem jeziku, celo v strokovnih besedilih.[7] Malenkostno omiljena je omejitev žargonske rabe v pravopisnem slovarju (SP 2001: § 1060), saj je oznaka »žargonsko« opredeljena kot »jezikovna prvina, ki se v strokovnem izrazju uporablja **namesto navadnega strokovnega** poimenovanja« (poud. a.). Zdi se, da je ta omejitev za besede, kot sta npr. *tenisač* in *radijec*, preozka.

4.3.2. *duševen – mentalen*

Spreminja se tudi funkcijskozvrstna pripadnost leksike, ki jo v prispevku ponazarjamo pri sopomenskem paru *mentalen – duševen*. Medtem ko je prvi pridevnik (*duševen*) prisoten že v protestantskem besedišču (*Besedje*, 2014), drugega (mentalen) poznajo slovenski slovarji šele od prvega povojnega pravopisa (SP 1950) dalje. Prvi slovarji pridevnika *mentalen* normativno ne odsvetujejo, so pa bolj naklonjeni tedaj bolj razširjeni sopomenski različici *miseln*, ki kasneje iz slovarskih napotil ob pridevniku *mentalen* izgine, ohranja pa se pri pojmovni samostalniški izpeljanki *mentalnost – miselnost*, *mentaliteta* (SSKJ2, 2014). Tudi najnovejši slovar sopomenk, *Sinonimni slovar slovenskega jezika* (Snoj idr., 2016), pridevnika *miseln* ne navaja med sopomenkami pridevnikov *duševen* in *mentalen*.

Izhodišče za korpusno primerjavo je torej opredelitev v SSKJ2 (2014). Pridevnik *mentalen* (*mentalni*) v tem slovarju opredeljujejo kot izraz, rabljen »zlasti v leposlovnem ali znanstvenem jeziku«, in zanj predlagajo nevtralno sopomenko *duševen* (*duševni*). Vendar pa se pogostnost uporabe pridevnika *mentalen* nakazuje, da ga v nekaterih kontekstih ni mogoče nadomestiti s slovarsko nadrejenim pridevnikom.

5. Raziskava in interpretacija

5.1. Oblikoslovno-besedotvorno uresničevanje preglasa pri izbranih tujih imenih[8]

5.1.1. *Leibnitz – Franz – Fritz*

Lastno ime *Leibnitz* po veljavni pravopisni normi tvori svojilni pridevnik na *-ov*/*-ev* s pisnima oblikama *Leibničev* in *Leibnitzev*, ki se govorno uresničujeta kot [lájbničev-], v rabi pa zasledimo tudi nestandardno možnost *Leibnitzov*, torej nepreglašeno obliko, za katero predvidevamo, da se bo pojavljala le v korpusih z nelektoriranimi besedili. Primerjalno poizvedbo smo izvedli tudi za imeni *Franz* in *Fritz*, pri katerih smo odkrivali le razmerje med pridevniki s pisno nespremenjenim izglasjem, a z uresničenim končniškim preglasom. Izbira imen *Franz* in *Fritz* se namreč ni izkazala za najbolj primerno, saj podomačenih oblik ***Frančev*** in ***Fričev*** v korpusih ni bilo mogoče ločiti od enakopisnih pridevnikov na *-ov*/*-ev* iz lastnih imen *Franc* in *Fric*. Za izvedbo primerjave smo se odločili, ker smo presodili, da je relevantno tudi razmerje med obema obraziloma, npr. *Franzov – Franzev*.[9]

Besedotvorje

Pravopisna prednostna oblika svojilnega pridevnika ***Leibničev*** in neprednostna dvojnica ***Leibnitzev*** se v rabi potrjujeta samo v korpusu Gigafida. Nestandardna tvorba *Leibnitzov*, ki po rabi izstopa kot najpogostejša, se pojavlja v treh korpusih. (Izjema je korpus Janes.)

[6] O podobnih vprašanjih je ob parih *stegno – bedro*, *čik – ogorek* ipd. razpravljal že Tomo Korošec (*Pet minut za boljši jezik*, 1972), ko tovrstne raziskave še niso bile empirično podkrepljene.

[7] Npr.: *Jeseni 1941 so nekateri slovenski športniki iz Ljubljane (atleti, plavalci, kajakaši, tenisača in pozimi umetnostna drsalka) nastopili na različnih tekmovanjih v Italiji, peščica pa jih je nadaljevala športno pot v italijanskih klubih*. (*Zanimanje za sport je prodrlo med Slovenci že v široke sloje*, 2005).

[8] Za iskanje smo uporabili pogoje npr. [lemma_lc="Franzev" & tag="P.*"] in [lemma="Franz" & lc="Franzem" & tag="S..e.*"], s katerimi smo želeli zmanjšati šum in pri oblikoslovju zajeti le orodniško obliko ednine. Najdene rezultate smo vedno primerjali tudi s preprostim iskanjem. Razlog za tovrstno poizvedbo je v tem, da imena niso vedno zapisana z veliko začetnico (napaka v lematizaciji morda?).

[9] Naslov zgodb o dečku Francu avstrijske pisateljice Christine Nöstlinger se v tiskanih različicah glasi *Francove zgodbe*, medtem ko jih v elektronskih virih najdemo lektorirane, npr. v Wikipediji (https://sl.wikipedia.org/wiki/Christine_Nöstlinger), v *Franceve zgodbe*, ne najdemo pa oblike *Frančeve zgodbe*.

Lema	Pridevnik	GF	KAS	slWaC	Janes	Σ
Leibnitz	*Leibnitzev*	2	0	0	0	2
	Leibnitzov	5	**8**	**14**	0	27
	Leibničev	**7**	0	0	0	7
Franz	*Franzev*	**33**	9	**8**	0	50
	Franzov	13	**23**	3	1	50
Fritz	*Fritzev*	29	3	**45**	1	78
	Fritzov	**35**	**9**	40	2	86

Tabela 1: Pogostnostna primerjava besedotvorne variantnosti imen *Leibnitz*, *Franz* in *Fritz* v štirih izbranih korpusih

Pri imenih *Fritz* in *Franz* bi najbolj zvesto približanje normi knjižnega jezika pričakovali v besedilih korpusa akademskih besedil, ki naj bi bila tudi v največjem odstotku lektorirana, je v tem korpusu največje odstopanje od edine oblike, ki jo pravopis dopušča: *Franzev* in *Fritzev*. Najbolj se normativnim pričakovanjem približata Gigafida (*Franzev*, toda *Fritzov*!) in korpus spletnih besedil slWaC, v katerem se raba nagiba k preglašenim oblikam. (Pojavitev v korpusu Janes je bilo zelo malo ali nič.)

Oblikoslovje

Oblika za orodnik ednine s preglasom (***Leibnitzem***) se potrjuje v korpusih Gigafida, KAS in slWaC, a ima zelo malo pojavitev, nestandardna oblika brez preglasa (*Leibnitzom*) pa je prav tako zelo redka, a se pojavi v korpusih KAS, slWaC in Janes.

Pri orodniku ednine sta obliki s preglasom (*Franzem*, *Fritzem*) pogostejši, a tudi nepreglašeni obliki (*Franzom*, *Fritzom*) sta zastopani: v korpusu slWaC imata skoraj pol toliko pojavitev kot preglašeni obliki. V korpusu Janes ni nobene od njiju.

Lema	Pridevnik	GF	KAS	slWaC	Janes
Leibnitz	*Leibnitzem*	3	1	1	0
	Leibnitzom	0	1	4	1
Franz	*Franzem*	**216**	**28**	**77**	0
	Franzom	72	13	31	2
Fritz	*Fritzem*	**90**	**9**	**56**	0
	Fritzom	34	5	36	0

Tabela 2: Pogostnostna primerjava oblikoslovne variantnosti imen *Leibnitz*, *Franz* in *Fritz* v štirih izbranih korpusih

5.1.2. *Keats – Yeats*

Imeni *Keats* in *Yeats* smo izbrali zaradi očitne zadrege pravopiscev, ki so v SP 2001 pri imenu *Yeats* predvideli obliko pridevnika *Yeatsov*, za *Keats* pa obliki *Keatsev* in *Keatsov*; obe se govorno uresničujeta kot [kíčev-]. Podobno je tudi pri orodniških oblikah, saj imenu *Yeats* pripišejo orodnik *Yeatsov*, imenu *Keats* pa pisno dvojnico: *s Keatsem/Keatsom* z enotno govorno uresničitvijo [s kícem]. Zanimalo nas je, kako se neenotna kodifikacija odraža v rabi oz. v besedilih korpusov 17 let po izidu pravopisa.

Besedotvorje

Lema	Pridevnik	GF	KAS	slWaC	Janes	Σ
Yeats	*Yeatsev*	0	0	0	0	0
	Yeatsov	15	1	13	0	39
Keats	*Keatsev*	0	0	0	0	0
	Keatsov	16	1	6	0	23

Tabela 3: Pogostnostna primerjava besedotvorne variantnosti imen *Yeats* in *Keats* v štirih izbranih korpusih

Korpusi Gigafida, KAS in slWaC (v korpusu Janes ni zadetkov) za svojilni pridevnik na *-ov*/*-ev* iz imen *Yeats* in *Keats* izkazujejo prevlado nepreglašene oblike (*Yeatsov*, *Keatsov*), ki pa je v nasprotju z imeni *Leibnitz*, *Franz* ..., tudi pravopisno dopuščena, pri *Yeats* celo edina.

Oblikoslovje

Tudi pri oblikoslovnem pregibanju v orodniku prevladujejo nepreglašene oblike (*Yeatsom*, *Keatsom*). Edini primer preglašene oblike (*Yeatsem*) najdemo v korpusu KAS.[10]

Lema	Orodnik	GF	KAS	slWaC	Janes	Σ
Yeats	*Yeatsem*	0	1	0	0	1
	Yeatsom	4	1	2	0	6
Keats	*Keatsem*	0	0	0	0	0
	Keatsom	3	1	4	0	8

Tabela 4: Pogostnostna primerjava oblikoslovne variantnosti imen *Yeats* in *Keats* v štirih izbranih korpusih

Imeni *Yeats* in *Keats* se torej približujeta pravopisnemu določilu v členu 957, ki določa, da za izglasnim sklopom »*ts* pišemo samo -ov: *Smuts – Smutsov*«, in oddaljujeta od napotil pravopisnega slovarja za ime *Keats*. Vendar bi za nedvoumno napotilo v kodifikacijskem priročniku potrebovali predvsem podatek o tem, kako zapisa *Keatsov* in *Keatsev* izgovorno uresničevati

5.2. Normativna napotila na (obliko)skladenjskem področju

5.2.1. *upravljati kaj* oz. *upravljati z/s čim*

V izbranih korpusih smo preverili pojavljanje zvez *upravljati kaj* in *upravljati s čim*.[11] Ugotovili smo, da sta v Gigafidi zvezi približno enako pogosti, medtem ko je v drugih treh korpusih pogostejša zveza z orodnikom, ki je

[10] Gre za zgled v jezikoslovnem delu, natančneje v končnem poročilu projekta Sporazumevanje v slovenskem jeziku, Kazalnik 17 – Slogovni priročnik, kjer so nanizane empirično pridobljene normativne zadrege uporabnikov: http://projekt.slovenscina.eu/Media/Kazalniki/Kazalnik17/Kazalnik_17_Slogovni_prirocnik_SSJ.pdf.

[11] Iskalna pogoja: [lemma="upravljati"][tag="S...t"] in [lemma="upravljati"][lemma="z"][tag="S...o"].

bila v normativno-slogovnih priročnikih druge polovice 20. stoletja odsvetovana.

	GF	KAS	slWaC	Janes
upravljati kaj	2597	4258	2230	276
upravljati s čim	2668	6084	3248	497

Tabela 5: Prikaz različne vezljivosti glagola *upravljati* v štirih izbranih korpusih

Težko bi trdili, da je glagol *upravljati* v orodniški vezavi drugače kolokabilen kot v tožilniški, saj se v vseh korpusih najpogosteje pojavljajo tudi sicer prekrivne besedne zveze:

- s tožilnikom: *upravljati družbo*, *upravljati* ***državo***, *upravljati* ***premoženje***;
- z orodnikom: *upravljati z denarjem*, *upravljati z* ***državo***, *upravljati s* ***premoženjem***.

Navedeno je potrdil tudi primerjalni pregled najpogostejših lem, ki se pojavljajo v tožilniku in orodniku v več kot dveh korpusih.[12]

upravljati			
Tožilnik	**Korpusi**	**Orodnik**	**Korpusi**
družbo	GF, KAS, SW, J	*s* ***premoženjem***	GF, KAS, SW, J
državo	GF, KAS, SW, J	*z denarjem*	GF, KAS, SW, J
premoženje	GF, KAS, SW, J	*z* ***državo***	GF, KAS, SW, J
delovanje	GF, KAS, SW	*s časom*	KAS, SW, J
podjetje	GF, KAS, SW	*s sistemom*	GF, KAS, SW
sredstva	GF, KAS, SW	*s* ***sredstvi***	GF, KAS, SW
vozilo	GF, KAS, SW	*z nepremičninami*	GF, KAS, SW

Tabela 6: Primerjava najpogostejših samostalniških lem ob glagolu *upravljati*, ki se pojavljajo v več kot dveh korpusih.

Čeprav bi pri orodniški vezavi pričakovali več samostalnikov, ki lahko nastopajo v vlogi sredstva (*upravljati z denarjem*, *s sredstvi*), pri tožilniški vezavi pa samostalnike, ki imajo abstrakten ali pojmovni pomen neke skupnosti (*družba*, *država*, *podjetje*), se je po razvrstitvi lem izkazalo, da je razdelitev preveč razpršena, da bi jo lahko posplošili. V obeh sklonih se namreč pojavljajo samostalniki iz obeh zgoraj navedenih skupin: *dokument*, *država*, *elektrarna*, *informacija*, *infrastruktura*, *podatek*, *podjetje*, *premoženje*, *sklad*, *smučišče*, *sprememba*, *sredstva*, *tveganja*, *viri*, *vozilo*, *znanje*.

Raziskava vezave glagola *upravljati* s tožilnikom in orodnikom ter primerjava vseh štirih korpusov je bila zamudna, saj iskalni mehanizem ne omogoča primerjave rezultatov različnih korpusov, zato smo primerjave lem izvedli ročno. Z vidika presoje ustreznosti slovarskih napotil tudi gradivska raziskava pritrjuje mnenju, izraženo v Jezikovni svetovalnici (Vranjek Ošlak, 2017), saj je orodniška vezava kljub večdesetletnemu prepovedovanju pogostnejša in med uporabniki bolj razširjena, ne glede na pomenski oz. kolokacijski okvir.

5.2.2. *prek/preko/po*

Pregled izbranih korpusov je pokazal,[13] da pri predlogih *prek* in *preko* v prvih 40 samostalniških kolokacijah v vseh štirih korpusih močno prevladuje pomen ʽsredstvo, posrednik', npr.: *prek/preko interneta*, *prek/preko elektronske pošte*, *prek aplikacije*, *prek/preko javnih medijev*, *prek/preko mobilnega telefona*, *prek/preko omrežja*, *prek/preko računalnika*, *prek/preko vmesnika*, *prek/preko spletne strani*, *prek/preko študentskega servisa* itd.

V drugih pomenih (ʽveč kot', ʽčez') nastopata ta dva predloga v zvezah, kot so npr. *prek/preko meje*, *preko noči*, *preko zime*, *preko mostu*, *preko dneva*, *prek 200 metrov* itd.

Za primerjavo smo preverili še 40 najpogostejših kolokacij predloga *po*, ki naj bi bil primerna oz. normativno priporočljiva zamenjava za predloga *prek* in *preko*. V pomenu ʽsredstvo, posrednik' nastopa ta predlog v le treh kolokacijah, in sicer: *po telefonu*, *po pošti* in *po cesti*.

Očitno uporabniki kljub normativnim napotkom v slovarjih (tako SP 2001 kot obeh izdajah SSKJ) uporabljajo predloga *prek* in *preko* v pomenu ʽsredstvo, posrednik' in ju le izjemoma zamenjujejo s predlogom *po*. Zlasti se nagibajo k rabi predloga *prek* v zvezah s samostalniki, ki označujejo sodobne načine komunikacije, npr. *internet*, *elektronska pošta*, *aplikacije*, *mobilni telefon*, *omrežje*, *vmesnik*.

5.3. Stilno prevrednotenje

5.3.1. *tenisač* (šport. žarg.)

Pri preverjanju socialnozvrstnih oznak je korpusna poizvedba bistvenega pomena. Samostalnik *tenisač* je kot športni žargonski izraz (tako jo opredeljujejo vsi sodobni normativni slovarji) poenobesedena različica »nevtralne« zveze *teniški igralec*. Vendar pa pregled prakse po korpusih pokaže, da uporabniki tudi samostalnik *tenisač* dojemajo kot nevtralen. Navajamo primer rabe iz korpusov KAS (diplomsko delo) in Janes (primer iz spletne klepetalnice):[14]

- *Iz načina preživljanja njegovega prostega časa in izbora vrste športa, lahko precej razberemo o njegovih značajskih karakteristikah. Igralci košarke in nogometa so verjetno bolj timsko naravnani ljudje kot* ***tenisači***. (vir: Ana Jakelić: *Vpliv zunanje podobe na poslovno uspešnost menedžerjev*);
- *Izmerjen Dopplerjev efekt ustvarja resen dvom o vedno enaki svetlobni hitrosti na ponoru. Če se vrnem na opisovan primer* ***tenisačev*** *z ogledali*, *mirujoči* ***tenisač***,

[12] Upoštevali smo 40 najpogostejših kolokacij.

[13] Iskali smo po lemah *prek*, *preko* in *po*.

[14] Navajamo le nekaj zgledov izmed mnogih.

> *kjer koli na poti od izvora do ogledala prestreže foton, opazi vedno enako valovno dolžino.* (vir: *Kvarkadabra*, Vprašanja & odgovori, Vprašanja za Einsteina).

5.3.2. *duševen – mentalen*[15]

Pogostnostna primerjava sopomenk *duševen* in *mentalen* v različnih korpusih kaže, da je pridevnik *duševen* skoraj trikrat pogostejši od pridevnika *mentalen*.

	GF	KAS	SlWaC	Janes	Σ
mentalen	5751	13.211	11.218	2405	32.585
duševen	31.679	61.832	25.723	3037	122.271

Tabela 7: Prikaz pogostnosti pridevniškega para *mentalen – duševen* v štirih izbranih korpusih

Pregled rabe v referenčnem korpusu Gigafida je pokazal, da je frekvenca pojavljanja pridevnikov po besedilnih tipih zelo podobna (čeprav sta pogostnostno različna – v razmerju 85 : 15 %). Oba sta najbolj značilna za besedilni tip **tisk/knjižno/strokovno**, najbolj pa se njuna relativna frekvenca razlikuje pri **internetnih besedilih** (zanje je bolj tipičen pridevnik *mentalen*) in tipu **tisk/drugo** (pogostejši je pridevnik *duševen*). To je deloma v skladu z opisom pridevnika *mentalen* v SSKJ, saj je tudi pridevnik *duševen* pogost v strokovnih besedilih, hkrati pa se pridevnik *mentalen* pojavlja tudi na internetu, kjer prevladuje praktičnosporazumevalna zvrst.[16]

Pregled in primerjava kolokacij v vseh štirih korpusih sta pokazala, da oba pridevnika nastopata v naslednjih besednih zvezah:[17]

- vsi štirje korpusi: *mentalno*/*duševno zdravje*, *mentalna*/*duševna zaostalost*, *mentalno*/*duševno stanje*;
- samo Gigafida: *mentalna*/*duševna prizadetost*, *mentalna*/*duševna motnja*;
- samo Janes: *mentalna*/*duševna higiena*, *mentalni*/*duševni bolnik*.

Pridevnik *mentalen* se v vseh **štirih** korpusih povezuje v besedne zveze *mentalna higiena*, *mentalna sposobnost*, *mentalno stanje*, *mentalna zaostalost* in *mentalno zdravje*.

V **treh** korpusih najdemo besedni zvezi *mentalna kondicija* (Gigafida, KAS in SlWaC) in *mentalna lenoba* (Gigafida, SlWaC in Janes).

V **dveh** korpusih pa so pogoste naslednje besedne zveze: *mentalni zemljevid*, *mentalna izčrpanost* (Gigafida in KAS); *mentalna retardacija*, *mentalna reprezentacija* (KAS in SlWaC); *mentalna stimulacija* (Gigafida in SlWaC); *mentalni domet* (Gigafida in Janes); *mentalni napor* (SlWaC in Janes).

Zanimive besedne zveze, ki se pojavljajo samo v enem od korpusov, so npr. še:

- *mentalna sirota*, *mentalni horizont*, *mentalni slepec*, *mentalna prizadetost*, *mentalni revček* (Gigafida);
- *mentalni trening*, *mentalna predstava*, *mentalni leksikon* (KAS);
- *mentalna jasnost* (slWaC);
- *mentalni invalid*, *mentalna pohabljenost*, *mentalna okužba* (Janes).

Pridevnik *duševen* se v vseh **štirih** korpusih povezuje v besedne zveze *duševna bolečina*, *duševna bolezen*, *duševni bolnik*, *duševna celovitost*, *duševna motnja*, *duševno počutje*, *duševna prizadetost*, *duševno ravnovesje*, *duševni razvoj*, *duševno stanje*, *duševna stiska*, *duševna zaostalost* in *duševno zdravje*.

V **treh** korpusih se pojavita besedni zvezi *duševni mir* in *duševna muka* (mn.) (Gigafida, slWaC in Janes).

V **dveh** korpusih najdemo besedne zveze *duševno trpljenje* (Gigafida in slWaC); *duševna obremenitev* (Gigafida in Janes); *duševna zmožnost*, *duševna sprostitev* (KAS in slWaC).

Zanimive besedne zveze, ki se pojavijo le v **enem** od korpusov, so npr. še: *duševni pretres* (Gigafida); *duševna manjrazvitost*, *duševna integriteta* (KAS); *duševni aparat* (Janes).

Iz analize je mogoče razbrati, da se pridevnik *duševen* pogostejši v besednih zvezah, ki se pojavijo v vseh štirih korpusih. To je lahko znak večje razširjenosti in funkcijskozvrstne nevtralnosti tega pridevnika v nasprotju s pridevnikom *mentalen*. Vendar pa tudi pridevnik *mentalen* v sodobni jezikovni praksi ni omejen zgolj na leposlovna in znanstvena besedila, kot so izkazovale leksikalne raziskave v preteklosti, temveč opažamo pojavljanje v različnih zvrsteh, npr. v korpusu spletnih besedil slWaC in v korpusu jezika družbenih omrežij oz. nestandardnega jezika Janes (razmerje v odstotkih je med *mentalen* in *duševen* tudi glede na korpus GF precej bolj izenačeno, 56 : 44 %). Prav nestandardni jezik predstavlja za sodobno normativistiko enega od prvih detektorjev napovedujočih se pomenskih premikov.

Naj izpostavimo še očitne pomenske prenose oz. čustveno zaznamovanost pridevnika *mentalen* v besednih zvezah korpusa nestandardne slovenščine Janes (*mentalni invalid*, *mentalna pohabljenost*, *mentalna okužba*), ki kažejo na primerno ustaljenost pridevnika v nevtralnem pomenu, da lahko razvija pomenske odmike.

6. Sklep

Glavni izziv pri korpusno podprtih raziskavah je za pisce jezikovnih priročnikov, slovnic in slovarjev v izbiri orodja, ki lahko potrdi intuitivne domneve o spremembah jezikovnih zgradb, oblik, tvorb ... ali ki odkrije še neopisane jezikovne vzorce. Opis sodobne slovenščine temelji na slovarskih in slovničnih raziskavah leksike iz 70. in 80. let

[15] Iskali smo po lemah *mentalen* in *duševen*.

[16] Takšna primerjava po besedilnih tipih za druge korpuse ni bila niti mogoča niti smiselna, saj ima le Gigafida enakovredno zastopanost različnih besedilnih tipov.

[17] Upoštevali smo prvih 20 kolokacij, in sicer samo zveze pridevnika v levem prilastku in samostalnika v jedru.

ter normativnem opisu iz 90. let, poskusi izkoriščanja korpusa kot gradivske osnove pa so se izkazali kot uspešni pri pripravi leksikalne baze za slovenščino, ki je predstavljena v monografiji *Leksikografski opis slovenščine v digitalnem okolju* (Gantar, 2015). Tudi prenova druge izdaje SSKJ (2014) je potekala s pomočjo referenčnega korpusa Gigafida, ki časovno zajema obdobje do leta 2010.

Pri interpretaciji sodobne jezikovne prakse Gigafida ne zadostuje več kot edini vir, še zlasti za novejše besedje ne (npr. *brexit*, *finteh*, *slimacid* ...), novi korpusi akademske, nestandardne in spletne slovenščine pa so dobrodošla dopolnila, ki ponujajo nove rezultate, a so poizvedbe zelo zamudne, saj moramo postopke ponavljati za vsak korpus posebej.

6.1. Povezovanje korpusov

V prispevku smo zato želeli pokazati na prednosti, ki bi jih imel enoten iskalni vmesnik (še zlasti, če bi ta ponudil tudi primerjave deležev med obsegovno različnimi korpusi). Po vseh korpusih je sicer mogoče iskati z orodjem Sketch Engine, ni pa možnosti primerjave rezultatov iskanj (moramo jih izvoziti v Excell ali kako drugo obliko). Primerjava rezultatov iskanja po vsakem korpusu posebej je zamudna, zaradi količine podatkov je tudi več možnosti za napake. Primerjava rezultatov iskanj znotraj skupnega vmesnika bi možnost napake zmanjšala, hitrost iskanja in primerjave pa bi se bistveno povečala. Dobrodošla bi bila tudi možnost iskanja po velikem združenem korpusu (sestavljenem iz vseh korpusov, ki so zdaj na voljo), če bi omogočal ohranitev podatkov iz izvirnih »podkorpusov«.

6.2. Pomen rezultatov poizvedb

Pri prikazu korpusne preverbe stanja smo se zato odločili za tri poizvedbe, ki so nujne in jezikoslovno pereče, hkrati pa do končnih rezultatov ni mogoče priti brez ročnega dela. Pri treh primerih jezikovne neustaljenosti smo ugotovili naslednje.

Na **oblikoslovno-besedotvornem področju** je mogoče zaznati odstopanja od zapovedanih normativnih vzorcev enakomerno v vseh korpusih (*Franzov : Franzev*), razlike so le v korpusu Janes, v katerem nekaterih lem ni bilo mogoče preveriti. Izrazito sistemsko utemeljene normativne zapovedi, ki očitno niso bile preverjene z rabo (*Keatsev*, *Yeatsev*), kljub določilom v pravopisu tudi po 20 letih v rabi niso prisotne – ne glede na tip korpusa.

Na **področju (obliko)skladnje** smo raziskovali tožilniško in orodniško vezavo pri glagolu *upravljati* in rabo predloga *prek/preko* za izražanje sredstva. Tu se stanje v Gigafidi temeljito razlikuje od drugih korpusov predvsem pri vezavi, saj je razmerje med dvojnicama (*upravljati kaj* : *upravljati z/s*) v Gigafidi enakomerno, medtem ko nepriporočljiva orodniška vezava prednjači v vseh drugih treh korpusih. Pri rabi predloga *preko*/*prek* razlike med korpusi niso izrazite.

Tudi na področju **stilnega prevrednotenja** (presoja žargonskosti pri samostalniku *tenisač* in raba pridevniškega para *mentalen – duševen*) se največje razlike izkazujejo med korpusoma Gigafida in Janes, pri čemer v slednjem zasledimo inovativne pomenske prenose, ki jih drugi korpusi ne izkazujejo, kar kaže na pomensko produktivnost prevzetega pridevnika *mentalen* v nestandardnem jeziku oz. na ustvarjalnost govorcev v nestandardnih zvrsteh.

Preverbe kažejo predvsem na to, da so razlike med rabo (oz. stanjem v korpusih) in priročniki manj razvidne na ravni sistemskega normiranja oz. vrednotenja jezikovnih prvin kot pri stilni in zvrstni kategorizaciji leksike ter stilističnih napotkih, ki koreninijo v jezikoslovju polovice 20. stoletja, kjer je bila v slovenščini očitno narejena »tiha« revolucija, s katero se sociolingvistika še ni spoprijela.

7. Literatura

Besedje slovenskega knjižnega jezika 16. stoletja. 2014. www.fran.si. Dostop 3. 4. 2018.

Douglas Biber, Susan Conrad in Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Helena Dobrovoljc. 2014. Raba predloga »prek« v pomenu ʽizražanje sredstva'. *Jezikovna svetovalnica*, https://svetovalnica.zrc-sazu.si/topic/9/raba-predloga-prek-v-pomenu-izražanje-sredstva. Dostop 9. 4. 2018.

Helena Dobrovoljc. 2015. Kako je prav: »Francova« ali »Frančeva« sestra?. *Jezikovna svetovalnica*, https://svetovalnica.zrc-sazu.si/topic/ 789/kako-je-prav-francova-ali-frančeva-sestra. Dostop 9. 4. 2018.

Helena Dobrovoljc in Tina Lengar Verovnik. 2018. Pregibanje imen dveh avstrijskih politikov: »Kurz« in »Van der Bellen«. *Jezikovna svetovalnica*, https://svetovalnica.zrc-sazu.si/ topic/ 2604/pregibanje-imen-dveh-avstrijskih-politikov-kurz-in-van-der-bellen. Dostop 9. 4. 2018.

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar in Milan Ojsteršek. 2016. Slovenska znanstvena besedila: prototipni korpus in načrt analiz. V: Erjavec, Tomaž (ur.), Fišer, Darja (ur.). *Zbornik konference Jezikovne tehnologije in digitalna humanistika.* Znanstvena založba Filozofske fakultete. 58–64.

Tomaž Erjavec, Nikola Ljubešić in Nataša Logar Berginc. 2015. The slWaC corpus of the Slovene Web. *Informatica*: *an international journal of computing and informatics*, št. 1, 35–42.

Darja Fišer, Tomaž Erjavec in Nikola Ljubešić. 2016. JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0*, 4 (2): 67–99.

Polona Gantar. 2015. *Leksikografski opis slovenščine v digitalnem okolju.* Ljubljana: Filozofska fakulteta.

Rafał Górski. 2007. Representativeness of a written part of a Polish general-reference corpus. Primary notes. V: *Corpus linguistics*, *computer tools*, *and applications - state of the art*: PALC.

Nataša Logar in Nikola Ljubešić. 2013. Gigafida in slWaC: tematska primerjava. *Slovenščina* 2.0. 78–110.

Christine Möhrs, Meike Meliss in Dolores Batinić. 2017. LeGeDe – Towards a Corpus-based Lexical Resource of Spoken German. V: *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Ur. Iztok Kosem. Carole Tiberius. Miloš Jakubíček. Jelena

Kallas. Simon Krek. Leiden, the Netherlands, 19–21 September 2017. 281–298.

NoSketch Engine. URL: https://www.clarin.si/noske/. Dostop 23. 3. 2018 – 14. 4. 2018.

Serge Sharoff. 2010. Analysing Similarities and Differences between Corpora. V: T. Erjavec, J. Žganec Gros (ur.): *Zbornik Sedme konference Jezikovne tehnologije*. 5−11.

SSKJ = *Slovar slovenskega knjižnega jezika*. Prva knjiga A–H (1970); druga knjiga I–Na (1975); tretja knjiga Ne–Pren (1979); četrta knjiga Preo–Š (1985); peta knjiga T–Ž (1991) z dodatki od A–Š. Ljubljana: SAZU – Državna založba Slovenije.

SSKJ2 = *Slovar slovenskega knjižnega jezika*. 2014. Druga, dopolnjena in deloma prenovljena izdaja. Ljubljana: Založba ZRC, ZRC SAZU.

SP 1950 = *Slovenski pravopis*. 1950. Ljubljana: SAZU – Državna založba Slovenije.

SP 1962 = *Slovenski pravopis*. 1962. Ljubljana: SAZU – Državna založba Slovenije.

SP 2001 = *Slovenski pravopis*. 2001. Ljubljana: Založba ZRC.

Vera Smole. 1997. Sovplivanje samoglasnikov in soglasnikov v vzhodnodolenjskih govorih. *Jezikoslovni zapiski*. 167–174.

Jerica Snoj idr. 2016. *Sinonimni slovar slovenskega jezika*. Ljubljana: Založba ZRC.

Jože Toporišič. 2004. *Slovenska slovnica*. Četrta izdaja. Maribor: Založba Obzorja.

Urška Vranjek Ošlak. 2017. Nekoč nepravilno, danes dovoljeno: upravljati s skladom in upravljati sklad. *Jezikovna svetovalnica*. URL: https://svetovalnica.zrc-sazu.si/topic/2666/nekoč-nepravilno-danes-dovoljeno-upravljati-s-skladom-in-upravljati-sklad. Dostop 3. 4. 2018.

Frekvenčni seznami n-gramov v korpusih slovenskega jezika

Kaja Dobrovoljc

Laboratorij za umetno inteligenco, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
kaja.dobrovoljc@ijs.si

Povzetek

V prispevku predstavimo postopek luščenja besednih n-gramov, ki za dani korpus izdela in izpiše sezname nizov poljubnega tipa pojavnic poljubne dolžine. Izdelano programsko orodje poleg modula za izdelavo običajnega frekvenčnega seznama vseh n-gramov vključuje še modul za njegovo nadaljnje filtriranje ter modul za izdelavo skupnega t. i. prilagojenega frekvenčnega seznama, ki pri štetju n-gramov upošteva medsebojno vsebovanost nizov različnih dolžin. Predstavimo in primerjamo rezultate pilotnega luščenja za štiri referenčne korpuse slovenskega jezika (pisna, govorjena, spletna, zgodovinska slovenščina) ter jih ovrednotimo z vidika možnosti nadaljnjih jezikoslovnih in jezikovnotehnoloških raziskav.

N-gram Frequency Lists for Reference Corpora of Slovenian Language

This paper presents a procedure for extraction of word n-grams that produces a list of n-grams of any type and any length for a given corpus. In addition to the compilation of a common n-gram frequency list, the extraction tool also includes an additional module for subsequent filtering of the common frequency lists and a module for compilation of the so-called joint adjusted frequency list that takes into account the overlapping n-grams of different lengths. We describe and compare the results of a pilot application of this tool to four reference corpora of Slovenian (written, spoken, user-generated and historical Slovenian), and discuss their value for future applications in linguistics and language technologies.

1. Uvod

V jezikoslovnih raziskavah, ki svoja spoznanja gradijo na analizah obsežnih zbirk avtentičnih primerov jezikovne rabe (besedilnih korpusov), enega temeljnih metodoloških postopkov predstavlja analiza frekvenčnih seznamov besedišča, tj. nabora vsebovanih besed s pripisanim podatkom o pogostosti v opazovanem korpusu. Ti seznami so koristni za splošne leksikološke analize besedišča jezika (Gorjanc, 2005), za ugotavljanje leksikalnih specifik korpusov (Kosem in Verdonik, 2012; Zwitter Vitez in Fišer, 2015; Verdonik in Maučec, 2016), za izdelavo geslovnikov v leksikalnih podatkovnih zbirkah (Gantar, 2015; Dobrovoljc et al., 2015) ali za statistično modeliranje jezika, če naštejemo le nekaj najpogostejših jezikoslovnih in jezikovnotehnoloških aplikacij v slovenskem prostoru.

Poleg frekvenčnih seznamov posameznih besed pa se danes pojavlja tudi vse večja potreba po frekvenčnih seznamih daljših enot oz. besednih nizov, zlasti ob spoznanju raziskav formulaičnega jezika, ki dokazujejo, da je jezik prepreden z večbesednimi vzorci, ki vsaj na neki točki jezikovne rabe delujejo kot nerazstavljiva celota in se kot taki tudi shranjujejo v mentalni leksikon govorcev (Sinclair, 1991; Wray, 2005). Čeprav se je v slovenskem korpusnem jezikoslovju besednim nizom doslej namenjalo manj pozornosti kot drugim tipom večbesednih enot, kot so kolokacije, kombinacije bolj ali manj oddaljenih besed s statistično izstopajočo povezanostjo (Logar et al., 2014; Gantar et al., 2015; Ljubešić et al., 2015), so v zadnjem obdobju vse bolj aktualne tudi raziskave besednih nizov, denimo za potrebe analize večbesednih leksikalnih enot na ravni diskurza (Dobrovoljc, 2018a).

Izdelavo frekvenčnih seznamov besed oz. besednih nizov različnih dolžin (za katera bomo v nadaljevanju uporabljali splošnejši izraz n-gram) omogočajo številna različna specializirana korpusna orodja, kot so kfNgram,[1] N-Gram Phrase Extractor,[2] N-gram Extraction Tool[3] ali mwe toolkit,[4] kot tudi večina zmogljivejših orodij za splošno korpusno analizo, kot so SketchEngine,[5] WordSmith,[6] NooJ[7] ali AntConct,[8] vendar je uspešnost izdelave frekvenčnih seznamov največkrat odvisna od zmogljivosti računalniške infrastrukture, na katerih ti programi gostijo. Za obsežnejše, referenčne korpuse, do katerih raziskovalci pogosto tudi nimajo neposrednega dostopa, so ta orodja torej manj uporabna, zato je smiselno tovrstne spiske jezikoslovcem in drugim potencialnim uporabnikom ponuditi kot vnaprej pripravljene, samostojne jezikovne vire.

V nadaljevanju prispevka tako predstavimo proces izdelave tovrstnih frekvenčnih seznamov za izbrane referenčne korpuse slovenskega jezika, predstavljene v 2. razdelku, pri čemer poleg samega postopka izdelave frekvenčnih seznamov različnih tipov (3. razdelek) na podlagi objavljenih seznamov (4. razdelek) v jedrnem 5. razdelku predstavimo še njihovo pilotno kvantitativno analizo in medsebojno primerjavo, z namenom prikaza številnih možnosti nadaljnjih raziskav (6. razdelek).

2. Izbrani korpusi

V nadaljevanju opisani postopek luščenja (razdelek 3), ki ga je mogoče prenesti na katerikoli korpus v predvidenem vhodnem formatu, smo za potrebe izhodiščne evalvacije aplicirali na štiri referenčne korpuse slovenskega jezika različnih velikosti in jezikovnih zvrsti (Tabela 1): uravnoteženi korpus sodobne pisne slovenščine Kres (Logar Berginc et al., 2012), ki vsebuje uravnotežen nabor leposlovnih, stvarnih, periodičnih, spletnih in drugih pisnih oblik besedil iz obdobja 1990−2011; korpus sodobne

[1] http://www.kwicfinder.com/kfNgram/kfNgramHelp.html

[2] http://lextutor.ca/n_gram/

[3] http://homepages.inf.ed.ac.uk/lzhang10/ngram.html

[4] http://mwetoolkit.sourceforge.net/

[5] http://www.sketchengine.co.uk/

[6] http://www.lexically.net/wordsmith/index.html

[7] http://www.nooj4nlp.net/

[8] http://www.laurenceanthony.net/software/antconc/

govorjene slovenščine Gos (Zwitter Vitez in Verdonik, 2011), ki vsebuje transkripcije spontanega govora v različnih javnih in zasebnih, formalnih in neformalnih situacijah; korpus uporabniških spletnih vsebin Janes (Fišer et al., 2016), ki vsebuje besedila slovenskih tvitov, forumov, blogov, komentarjev in pogovornih strani Wikipedije; in korpus starejše slovenščine IMP (Erjavec, 2015), ki vsebuje leposlovna dela, rokopise in periodiko od konca 16. stoletja do leta 1918.

Korpus	Št. vseh pojavnic	Št. besednih pojavnic
Gos	1.110.649	1.033.024
IMP	17.723.874	14.405.281
Kres	120.447.573	97.135.649
Janes	252.904.238	191.292.328

Tabela 1: Seznam izbranih referenčnih korpusov slovenskega jezika s podatkom o številu besednih in vseh pojavnic.

3. Izdelava frekvenčnih seznamov

Postopek luščenja n-gramov smo zasnovali v obliki programske skripte, ki kot vhodno datoteko prejme korpus v tabelaričnem besedilnem formatu (Erjavec, 2013) in zanj izdela frekvenčni seznam nizov pojavnic v korpusu (besednih n-gramov), pri čemer poljubno določimo: **tip pojavnice** (originalni zapis, normaliziran zapis, lema, oblikoskladenjska oznaka ali različne kombinacije teh tipov), **dolžino niza** (velikost *n*) in pogoj **(ne)upoštevanja ločil**, glede na to, ali želimo kot relevantne gradnike n-gramov upoštevati tudi ločila ali ne.[9]

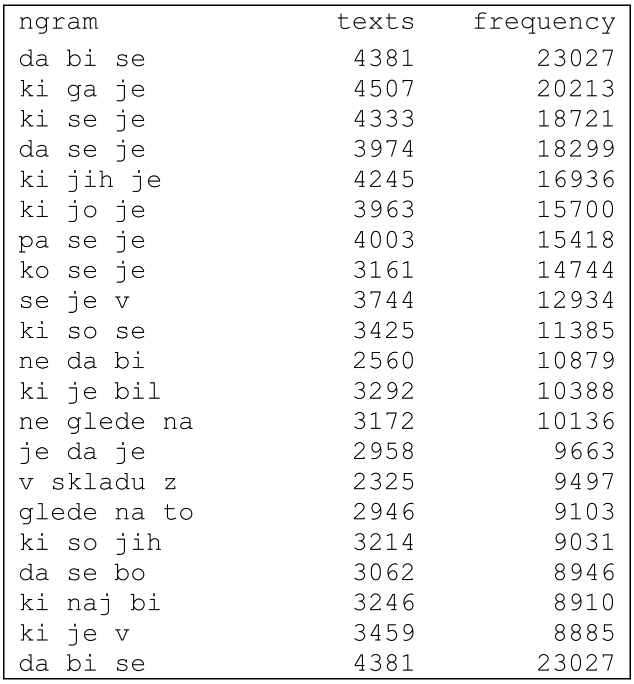
Glede na raznolike potrebe potencialnih uporabnikov je bil ta proces zasnovan kot niz treh zaporednih korakov (modulov), znotraj katerih nastajajo različni tipi samostojnih, zaključenih seznamov. Vsakega izmed modulov, tj. postopke luščenja, filtriranja in prilagajanja n-gramov, predstavimo v nadaljevanju.

3.1. Luščenje z običajnim štetjem

Na podlagi zgoraj navedenih parametrov program v prvem koraku za vsako poved vsakega besedila v danem korpusu izdela seznam vseh relevantnih nizov in jih po zaključku štetja v celotnem korpusu izpiše v tabelarični besedilni datoteki, skupaj s podatkom o številu različnih besedil, v katerih se n-gram pojavlja, in njegovi absolutni pogostosti v korpusu (Slika 1).

3.2. Filtriranje

Ker so tovrstnih frekvenčni seznami n-gramov zaradi velikega deleža pojavnic z enkratnimi oz. redkimi pojavitvami običajno zelo obsežni (glej denimo število različnic na primeru v Tabeli 2), v drugem koraku vpeljujemo modul za njihovo filtriranje. Poleg minimalnega frekvenčnega praga, tj. najmanjšega zahtevanega števila pojavitev danega n-grama v korpusu, uporabnik poljubno določi tudi minimalni besedilni prag, tj. najmanjše zahtevano število različnih besedil, v katerih se dani n-gram pojavi.



```
ngram              texts    frequency
da bi se            4381        23027
ki ga je            4507        20213
ki se je            4333        18721
da se je            3974        18299
ki jih je           4245        16936
ki jo je            3963        15700
pa se je            4003        15418
ko se je            3161        14744
se je v             3744        12934
ki so se            3425        11385
ne da bi            2560        10879
ki je bil           3292        10388
ne glede na         3172        10136
je da je            2958         9663
v skladu z          2325         9497
glede na to         2946         9103
ki so jih           3214         9031
da se bo            3062         8946
ki naj bi           3246         8910
ki je v             3459         8885
da bi se            4381        23027
```

Slika 1: Primer izpisa frekvenčnega seznama najpogostejših 20 3-gramov v korpusu Kres za nize normaliziranih pojavnic brez upoštevanja ločil.

Vpeljava tega pogoja je smiselna, kadar želimo iz končnega seznama izločiti n-grame, ki so vezani na avtorske, tehnične ali vsebinske specifike enega oz. majhnega števila besedil.[10]

Kot je razvidno iz primerjave števila različnih izluščenih n-gramov na seznamih brez filtriranja (Tabela 2) in s filtriranjem glede na izbrani minimalni frekvenčni in/ali besedilni prag (Tabela 3), so filtrirani seznami bistveno krajši in s tem primernejši za nadaljnje analize. Že razmeroma nizek frekvenčni prag relativne pogostosti 10 pojavitev na milijon in pojavljanja v vsaj 2 različnih besedilih nabor izluščenih normaliziranih n-gramov (Tabela 3) zoži na manj kot odstotek prvotnega seznama, denimo 0,76 % vseh nizov v korpusu Gos ali celo 0,005 % vseh nizov v korpusu Janes.

Besed	Gos	IMP	Kres	Janes
1	62.710	411.126	1.404.903	2.502.460
2	394.416	4.615.749	23.612.952	35.969.381
3	692.260	9.077.740	53.392.506	89.128.455
4	750.559	10.458.202	66.139.463	113.108.440
5	698.208	10.105.879	66.554.945	110.320.967
SUM	2.598.153	34.668.696	211.104.769	351.029.703

Tabela 2: Število vseh različnic 1–5-gramov v izbranih korpusih za normalizirane pojavnice brez upoštevanja ločil.

[9] Pogoj (ne)upoštevanja ločil nam tako omogoča skupno ali ločeno štetje nizov, ki se razlikujejo zgolj glede na vsebovana ločila (npr. *kljub , temu da - kljub temu , da - kljub temu da*).

[10] V izbranih korpusih so denimo najpogostejši normalizirani 3-grami s pojavitvijo v enem samem besedilu *na radiu center* (korpus Gos, 33 pojavitev), *s. frančišek zalaze* (IMP, 97), *dela z ekipo* (Kres, 635) in *Ubijeno od četnika* (Janes, 699).

Besed	Gos	IMP	Kres	Janes
1	6.628	8.564	10.560	9.266
2	9.387	6.321	5.215	6.412
3	3.343	1.154	950	1.350
4	287	50	49	251
5	47	6	11	171
SUM	19.692	16.095	16.785	17.450

Tabela 3: Število različnic 1–5-gramov v izbranih korpusih za normalizirane pojavnice brez upoštevanja ločil, ki se pojavijo v vsaj 2 različnih besedilih in z vsaj 10 pojavitvami na milijon pojavnic.

3.3. Prilagajanje štetja

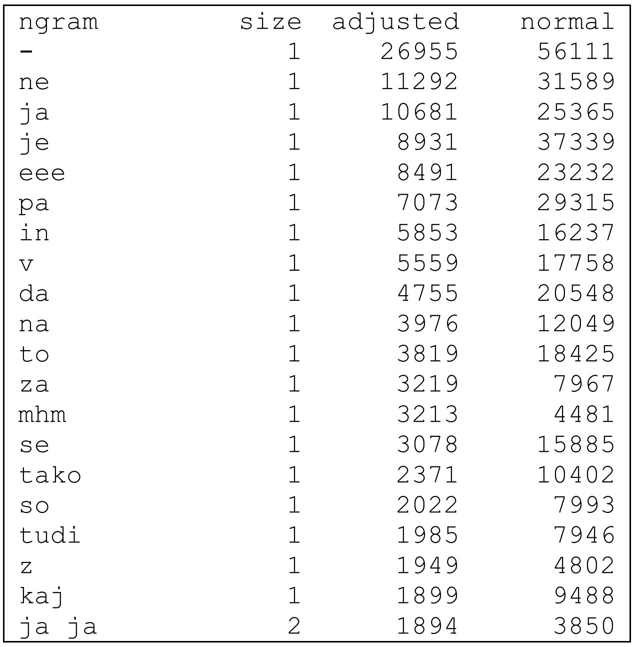
Frekvenčni seznami z običajnim štetjem pogostosti (in poljubnim načinom filtriranja), kakršne omogočajo tudi v uvodu našteta korpusna orodja, omogočajo različne jezikoslovne analize in jezikovnotehnološke aplikacije, vendarle pa ne dajejo zadovoljivega odgovora na vprašanje, kako pogost je določen niz v primerjavi z drugimi besedami ali besednimi nizi, saj pri njihovem štetju ne upoštevajo medsebojne vsebovanosti, torej dejstva, da je vsak niz dveh ali več besed (n-gram) sestavljen iz krajših nizov (n-1-gramov).

Za ponazoritev tega problema vzemimo pojavljanje dvobesednega niza *glede na* v korpusu govorjene slovenščine, ki se v kar 58 % vseh pojavitev (178 od skupno 309 pojavitev) v korpusu Gos pojavlja kot del daljšega besednega niza *glede na to*. To pomeni, da bi bilo na *skupnem* frekvenčnem seznamu n-gramov v korpusu Gos niz *glede na to* ustrezneje navajati pred nizom *glede na*, saj se ta izven te besedne zveze pojavlja manj pogosto kot znotraj nje. Po drugi strani bi morali na enak način tudi pri izračunu pogostosti besednega niza *glede na to* nato upoštevati, da se tudi sam pojavlja kot del pogostih daljših besednih nizov, npr. v 69 % (122 od 178 pojavitev) kot del niza *glede na to da*, ta pa se denimo včasih pojavi kot del nadrejenega besednega niza *ne glede na to da* (15 od 122 pojavitev).

Da bi uporabnikom omogočili tudi take medsebojne primerjave pogostosti nizov različnih dolžin, smo v tretji korak našega orodja vključili še modul za tovrstno statistično redukcijo podnizov. Med različnimi predlaganimi metodami za tako modificirano štetje (npr. Nagao in Mori, 1994; da Silva in Lopes 1999; Lü et al., 2005) smo izbrali algoritem za izdelavo t. i. prilagojenega frekvenčnega seznama (O'Donnell, 2010). Če njegovo delovanje na kratko povzamemo, ta v predhodno indeksiranem korpusu, v katerem ima vsaka pojavnica pripisan svoj unikatni številčni indeks, za vsak relevantni n-gram, ki smo ga izluščili v prvem koraku, tj. pri luščenju z običajnim štetjem, shrani podatek o njegovih konkretnih indeksih (mestih pojavljanja v korpusu) in nato preveri, ali se v povedi pojavi kot del daljšega relevantnega niza (n+1-grama).[11] Če to drži, se iz seznama vseh pojavitev danega n-grama ta pojavitev odstrani, s čimer se njegova končna pogostost zmanjša oz. ustrezno prilagodi. Ta postopek nato ponovimo še za vsako naslednjo dolžino, do največje določene dolžine nizov, ki jim frekvence ni mogoče prilagoditi.

S tem iterativnim postopkom dobimo nekoliko drugačen frekvenčni seznam n-gramov, kakršen je koristen predvsem za nadaljnje leksikološke raziskave. Ta vsebuje drugačno število različnic (odstranijo se npr. n-grami, ki se v korpusu pojavljajo zgolj kot gradniki daljših nizov, npr. [po/na] *eni strani*, [v] *zvezi z*, [se] *mi zdi*, *dame in* [gospodje]) in ustreznejše število pojavnic (vsaka besedna pojavnica lahko pripada zgolj nizu ene dolžine).[12] To posledično vodi v drugačno razvrščanje besednih nizov po pogostosti, saj so v nasprotju z običajnim štetjem na *skupnem* frekvenčnem seznamu daljši nizi (npr. *glede na to da*) lahko uvrščeni višje kot njihovi podnizi (npr. *glede na to* ali *na to da*), če se slednji večinoma pojavljajo zgolj znotraj daljših stalnih nizov.

Za razliko od načina izpisa frekvenčnih seznamov vseh (razdelek 3.1) oz. filtriranih (razdelek 3.2) n-gramov, ki so ločeni glede na dolžino niza (Slika 1), tretji modul vrne en sam, skupni frekvenčni seznam za vse n-grame izbranega intervala. Kot prikazuje Slika 2, za vsak n-gram poleg podatka o dolžini niza izpiše še podatek o prilagojeni in izhodiščni oz. običajni pogostosti v korpusu.



```
ngram          size  adjusted    normal
-                 1     26955     56111
ne                1     11292     31589
ja                1     10681     25365
je                1      8931     37339
eee               1      8491     23232
pa                1      7073     29315
in                1      5853     16237
v                 1      5559     17758
da                1      4755     20548
na                1      3976     12049
to                1      3819     18425
za                1      3219      7967
mhm               1      3213      4481
se                1      3078     15885
tako              1      2371     10402
so                1      2022      7993
tudi              1      1985      7946
z                 1      1949      4802
kaj               1      1899      9488
ja ja             2      1894      3850
```

Slika 2: Primer izpisa prilagojenega frekvenčnega seznama v korpusu Gos za nize normaliziranih pojavnic dolžine 1–5 besed brez upoštevanja ločil.

4. Objava seznamov

Z opisanim postopkom (razdelek 3) smo za izbrane referenčne korpuse (razdelek 2) v uvodni iteraciji luščenja izdelali običajne, filtrirane in prilagojene frekvenčne sezname za n-grame dolžine 1 do 5 pojavnic, za različne tipe pojavnic, z in brez upoštevanja ločil. Vsi seznami so pod licenco CC-BY-SA za prenos in nadaljnjo uporabo

[11] Kot relevanten nadrejeni niz (n+1) se upošteva vsak niz nad izbranim minimalnim frekvenčnim pragom, ki naj bi označeval stalnost oz. statistično relevantnost. Ta v sorodnih raziskavah običajno obsega od 5 (reference) do 10 (reference), 20 (reference) ali celo 40 (reference) pojavitev na milijon pojavnic.

[12] Ne moremo pa trditi, da vsaka pojavnica pripada zgolj enemu nizu, saj algoritem ne predvideva kakršnegakoli prilagajanja štetja prekrivnih nizov enake dolžine.

prosto dostopni na repozitoriju CLARIN.SI (Dobrovoljc, 2018b-d).

5. Primerjava seznamov

Čeprav izdelani seznami omogočajo številne nadaljnje analize in medsebojne primerjave, se z namenom uvodne ponazoritve njihove potencialne uporabne vrednosti v tem razdelku osredotočimo na splošno kvantitativno primerjavo prilagojenih frekvenčnih seznamov n-gramov v izbranih korpusih, in sicer z vidika njihovega nabora (4.1), pogostosti (4.2) in raznolikosti (4.3.).

Glede na raznolike zapisovalne posebnosti originalnih pojavnic v posameznih tipih korpusov primerjamo nize normaliziranih pojavnic, tj. ročno standardiziranega zapisa v korpusu Gos, strojno standardiziranega zapisa v korpusih IMP in Janes ter zapisa z malimi črkami v korpusu Kres, brez upoštevanja ločil. Upoštevali smo minimalni frekvenčni prag 10 pojavitev na milijon pojavnic in minimalni besedilni prag pojavljanja v vsaj 2 različnih besedilih, s čimer celotno množico identificiranih n-gramov, ki ustrezajo tem pogojem (Tabela 2), zamejujemo na razmeroma stalno oz. pogosto besedišče.

5.1. Raznolikost stalnega besedišča

Primerjava skupnega števila različnih 1–5-gramov v Tabeli 4 kaže, da se v korpusih nad izbranim minimalnim frekvenčnim pragom pojavlja od okoli 15 do 19 tisoč različnih enot stalnega besedišča. Medtem ko je število različnic v vseh treh korpusih pisnega jezika precej podobno (okoli 16 tisoč enot), korpus govorjene slovenščine Gos izkazuje večje število različnic (18.721). To kaže, da govorci v spontanem govoru uporabljajo večji nabor stalnega besedišča kot v pisnem jeziku, kjer se po drugi strani pojavlja večji nabor manj pogostega besedišča, kot kaže tudi primerjava števila različnic pred in po filtriranju glede na izbrani frekvenčni prag (Tabeli 2 in 3).

Št. besed	Gos	IMP	Kres	Janes
1	6.371	8.460	10.270	8.914
2	8.860	6.087	4.885	5.984
3	3.199	1.131	901	1.110
4	244	43	32	68
5	47	6	11	171
Skupaj	18.721	15.727	16.099	16.247

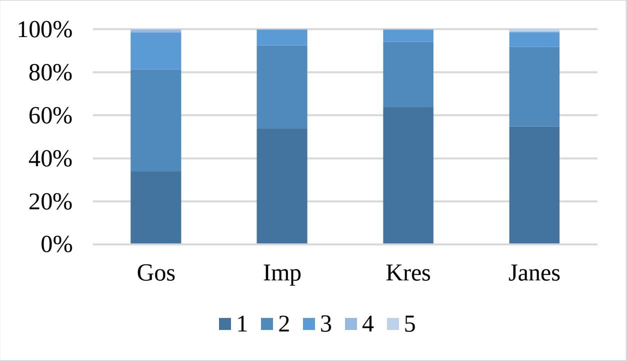
Tabela 4: Število različnic na prilagojenem frekvenčnem seznamu za n-grame normaliziranih pojavnic brez upoštevanja ločil, ki se pojavijo v vsaj 2 različnih besedilih in z vsaj 10 pojavitvami na milijon pojavnic.

Pri tem je treba poudariti, da rezultati te primerjave niso odvisni od precejšnjih razlik v velikosti korpusov (Tabela 1), saj zelo podobna razmerja med korpusi dobimo, če n-grame luščimo iz enako velikih vzorcev korpusov (Tabela 5), ki smo jih v našem konkretnem primeru izdelali z naključnim vzorčenjem stavkov v skupnem obsegu približno 1 milijon pojavnic.

Št. besed	Gos	IMP	Kres	Janes
1	5.746	7.482	9.248	7.487
2	8.064	5.415	4.231	4.952
3	2.736	1.009	778	885
4	208	39	30	61
5	40	4	13	110
Skupaj	16.794	13.949	14.300	13.495

Tabela 5: Število različnic na prilagojenem frekvenčnem seznamu naključnega vzorca vsakega korpusa v obsegu 1 milijon pojavnic za n-grame normaliziranih pojavnic brez upoštevanja ločil z vsaj 10 pojavitvami na milijon pojavnic.

Druga pomembna ugotovitev primerjave števila različnic v izbranih korpusih (Slika 3) pa izhaja iz dejstva, da se na vseh štirih frekvenčnih seznamih pojavlja razmeroma velik delež večbesednih enot (2- do 5-gramov), od 36 % vseh različnic v korpusu Kres do 66 % vseh različnic v korpusu Gos. To potrjuje določeno stopnjo formulaičnosti vseh oblik jezikovne rabe, pri čemer izstopajoči oz. večinski delež večbesednih enot na frekvenčnem seznamu korpusa Gos kaže, da je tudi v slovenščini govorjena raba izrazito bolj formulaična kot pisna (prim. npr. Biber (2009) za angleščino). V korpusih pisnega jezika po drugi strani prevladujejo enobesedne različnice, pri čemer pa oba specializirana korpusa (IMP in Janes) izkazujeta večjo stopnjo formulaičnosti (46,2 % oz. 45,1 % večbesednih enot) kot korpus sodobne standardne pisne slovenščine Kres.



Slika 3: Delež različnic posameznih dolžin na prilagojenih frekvenčnih seznamih izbranih korpusov za nize normaliziranih pojavnic dolžine 1–5 besed brez upoštevanja ločil, ki se v korpusu pojavijo v vsaj 2 različnih besedilih in vsaj 10-krat na milijon pojavnic.

V vseh štirih korpusih med večbesednimi enotami prevladujejo predvsem dvobesedni nizi, vendar nezanemarljiv delež predstavljajo tudi daljši nizi, od 5,9 % odstotkov vseh različnic v korpusu Kres do 18,6 % vseh različnic v korpusu Gos, kar potrjuje, da je na področju raziskav večbesedne leksike, ki se običajno osredotočajo na dvobesedne kolokacije, smiselno razvijati tudi metode za prepoznavo in analizo daljših večbesednih enot.

5.2. Pogostost stalnega besedišča

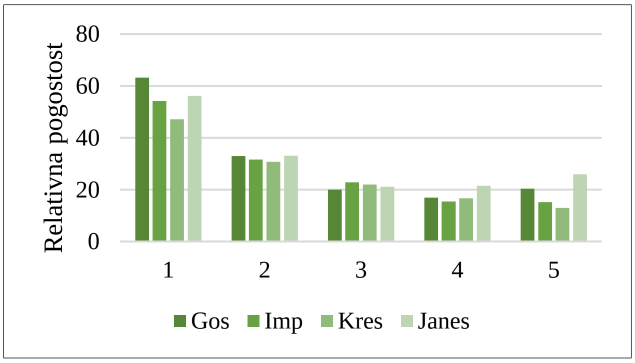
Primerjava skupne pogostosti 1–5-gramov v vsakem izmed korpusov v Tabeli 6 kaže, da se izluščeni stalni n-grami v vsakem korpusu pojavljajo s podobno povprečno pogostostjo (od 41 do 45 pojavitev na milijon pojavnic), tudi če primerjamo povprečno relativno pogostost nizov posameznih dolžin (Slika 4).

Št. besed	Gos	IMP	Kres	Janes
1	402.325	458.612	483.733	500.315
2	291.605	192.344	149.940	197.879
3	63.979	25.781	19.746	23.461
4	4.127	666	535	1.460
5	959	91	143	4.434
Skupaj	762.995	677.493	654.097	727.549
Povprečno	41	43	41	45

Tabela 6: Relativna pogostost različnic na prilagojenem frekvenčnem seznamu za n-grame normaliziranih pojavnic brez upoštevanja ločil, ki se pojavijo v vsaj 2 različnih besedilih in z vsaj 10 pojavitvami na milijon pojavnic.

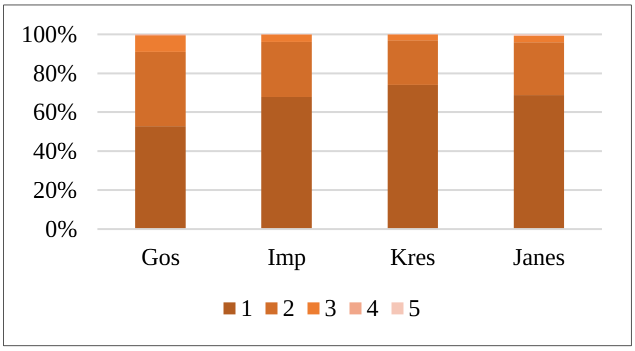
V vseh korpusih največjo povprečno pogostost rabe izkazujejo posamične besede, pri čemer nekoliko izstopajoča razlika v pogostosti povprečne besede v korpusu Gos (36 pojavitev na milijon) na eni strani in pogostost povprečne besede v korpusu Kres (47 pojavitev na milijon) potrjuje že izpostavljeno hipotezo, da se govorci v spontanem govoru ob pritiskih tvorjenja v realnem času poslužujejo manjšega nabora različnih besed, a te rabijo toliko pogosteje, medtem ko v pisni rabi zajemajo iz širšega nabora (manj pogostih) besed.

Za razliko od prepada v povprečni pogostosti eno- in večbesednih različnic je pogostost rabe daljših nizov bolj enakomerna, tako z vidika primerjave med nizi različnih dolžin kot z vidika primerjave med korpusi.[13]



Slika 4: Povprečna relativna pogostost različnic posameznih dolžin na prilagojenem frekvenčnem seznamu n-gramov normaliziranih pojavnic brez upoštevanja ločil, ki se pojavijo v vsaj 2 različnih besedilih in z vsaj 10 pojavitvami na milijon pojavnic.

Če pogostost n-gramov posameznih dolžin primerjamo še z vidika deleža glede na skupno pogostost vseh n-gramov prilagojenega frekvenčnega seznama (Slika 5), vidimo, da tudi ta analiza potrjuje visok delež formulaičnosti jezikovne rabe v slovenščini, saj v vseh korpusih vsaj četrtino vseh leksikalnih izbir (stalnega besedišča) predstavljajo stalni dvo- ali večbesedni nizi, pri čemer je v spontanem govoru raba besednih nizov celo skoraj enako pogosta kot raba posamičnih besed (47,3 % vseh pojavitev v korpusu Gos).



Slika 5: Delež pojavitev različnic posameznih dolžin na prilagojenih frekvenčnih seznamih izbranih korpusov za n-grame normaliziranih pojavnic dolžine 1–5 besed brez upoštevanja ločil, ki se v korpusu pojavijo v vsaj 2 različnih besedilih in vsaj 10-krat na milijon pojavnic.

5.3. Prekrivnost stalnega besedišča

Glede na ugotovljene kvantitativne podobnosti in razlike prilagojenih frekvenčnih seznamov izbranih korpusov nas je v tretjem koraku primerjave zanimala še njihova dejanska prekrivnost. Analizo povzemamo v obliki tabele (Tabela 7), ki za vsak korpus prikazuje tako delež prekrivnih n-gramov, ki se pojavljajo v enem ali več drugih korpusov, kot delež unikatnih n-gramov, ki se ne pojavljajo v nobenem drugem korpusu. Če za primer vzamemo korpus Janes, lahko iz tabele torej razberemo, da se 55,6 %, 42,0 % oz. 60,1 % n-gramov na frekvenčnem seznamu tega korpusa pojavlja tudi na frekvenčnem seznamu korpusa Gos, IMP oz. Kres, 25,3 % vseh n-gramov korpusa Janes pa je unikatnih, kar pomeni, da so bili kot relevantni identificirani zgolj v korpusu Janes. Rezultati te primerjavi razkrivajo več zanimivih ugotovitev.

	% prekrivnih n-gramov				% unikatnih
	v Gos	v IMP	v Kres	v Janes	
Gos		34,0	42,5	48,2	43,8
IMP	40,5		49,1	43,4	41,8
Kres	49,5	48,0		60,8	24,9
Janes	55,6	42,0	60,1		25,6

Tabela 7: Delež prekrivnih in unikatnih n-gramov na prilagojenem frekvenčnem seznamu n-gramov normaliziranih pojavnic brez upoštevanja ločil, ki se pojavijo v vsaj 2 različnih besedilih in z vsaj 10 pojavitvami na milijon pojavnic.

Prav vsi korpusi kažejo razmeroma velik delež unikatnih n-gramov − od 24,9 % unikatnega stalnega besedišča za korpus Kres do 43,8 % unikatnega stalnega besedišča za korpus Gos. Podrobnejša analiza seznama najpogostejših unikatnih nizov na eni strani razkriva, da so

[13] Nekoliko izstopajoča povprečna pogostost 4- in 5-gramov v korpusu uporabniških spletnih vsebin Janes je posledica dejstva, da se med njimi pretežno pojavljajo generični nizi tipa *People followed me and, a New Photo to Facebook, one person unfollowed me automatically* ipd., ki se ob deljenju spletnih povezav z drugimi uporabniki tvorijo samodejno.

med njimi pogosti nizi, vezani na označevalne specifike posamičnega korpusa, kot so na primer raba vezaja pri standardizaciji nebesednih ali anonimiziranih pojavnic v korpusu Gos (npr. - - in *gospod* -) ali posebnosti anonimizacije in standardizacije v korpusu Janes (npr. *[per] [per] [per]*, *[@per] [URL]*). Po drugi strani pa med unikatnimi n-grami posamičnih korpusov prevladujejo predvsem taki, ki razkrivajo specifike njihovega besedišča.

Kot prikazuje seznam desetih najpogostejših nizov različnih dolžin v vsakem izmed korpusov (Tabela 8) brez upoštevanja nizov z zgoraj navedenimi označevalnimi posebnostmi, v korpusu govorjene slovenščine Gos tako prevladujejo predvsem nizi z zapolnjenimi mašili, izrazi strinjanja, nedoločnosti in drugi pragmatični izrazi, pa tudi nestandardna govorjena leksika in izrazi, vezani na specifike samega nabora posnetih besedil; v korpusu IMP n-grami s časovno zaznamovanim besediščem, vključno z danes manj aktualnimi skladenjskimi vzorci; v korpusu Kres besedišče, vezano na zakonodajna in publicistična besedila; v korpusu Janes pa pojavnice, vezane na nebesedne vidike spletne komunikacije, pogosto rabo angleščine in spletno poizvedovanje.

Glede na delež unikatnih n-gramov največjo specializiranost besedišča torej kaže korpus govorjene slovenščine Gos, najbolj nevtralno oz. nezaznamovano pa je besedišče korpusa sodobne pisne slovenščine Kres, kar se odraža tudi pri analizi deleža prekrivnosti posameznih parov korpusov. Čeprav slednja odpira številne zanimive nadaljnje primerjave in analize, na tem mestu izpostavimo predvsem ugotovitev, da največjo podobnost besedišča izkazujeta sodobna standardna in spletna pisna slovenščina (60,8-% oz. 60,1-% prekrivnost med korpusoma Kres in Janes), najmanjšo pa sodobna govorjena in starejša pisna slovenščina (34,0-% oz. 40,5-% prekrivnost med korpusoma Gos in Imp).

	Gos
1	*eee, eem, tlele, nnn, tipo, čao, majčkeno, tukajle, šestdeset, devetdeset*
2	*in eee, eee eee, mhm mhm, eee v, ne eee, pa eee, eee in, eee ja, eee ne, pa pol*
3	*ja ja ja, ne ne ne, ja ne vem, na neki način, ne to je, mhm mhm mhm, eee to je, ne tako da, eee ne vem, eee tako da*
4	*ja ja ja ja, ne ne ne ne, to je to je, jaz mislim da je, ali pa kaj takega, zaradi tega ker je, in tako naprej ne, ja saj to je, da je da je, mhm mhm mhm mhm*
5	*ja ja ja ja ja, ne ne ne ne ne, šest osem nič osem nič, osem nič osem nič nič, šest šest osem nič osem, nič osem nič trinajst nič, osem nič trinajst nič ena, aha ja ja ja ja, s hiti na radiu city, zaslužite s hiti na radiu*
	IMP
1	*je., zavoljo, ondi, baron, lice, urno, zmerom, rekoč, dasi, čebele*
2	*ako se, je zopet, ako bi, ako je, dejal je, n. pr., in kakor, ne bil, ter je, moj bog*
3	*se je bil, kakor bi se, i. t. d., da bi ne, bi se bil, se je bila, na vse strani, mu je bil, mu je bila, se je bilo*
4	*kakor da bi se, od dne do dne, da se mu je, da bi se bil, in ko se je, se mu je zdelo, ki se mu je, kakor da bi bil, da bi se ne, se ji je zdelo*
5	*zdelo se mu je da, zdelo se mi je da, se mu je zdelo da, zdelo se ji je da, se mu je da je, in zdelo se mu je*
	Kres
1	*mag., članu, dodamo, določbe, priprava, varstva, odločbe, 1999, organa, odstavka*
2	*z dne, d. d., tega zakona, s področja, v postopku, v obdobju, osebnih podatkov, zaradi česar, foto reuters, za opravljanje*
3	*d. o. o., člena tega zakona, iz prejšnjega odstavka, pri tem pa, v republiki sloveniji, člena zakona o, državna revizijska komisija, po vsem svetu, v nasprotju s, v sodelovanju z*
4	*uradni list rs št., ki se nanašajo na, v skladu z zakonom, za okolje in prostor, cene izdelka franko tovarna, da se ne bi, ki se nanaša na, black process black plate, po drugi svetovni vojni, za šolstvo in šport*
5	*iz prvega odstavka tega člena, posneto v času terenskega dela, o spremembah in dopolnitvah zakona, spremembah in dopolnitvah zakona o, v uradnem listu republike slovenije, ne glede na to ali, med leti 1928 in 1947, objavi v uradnem listu republike, vrednost vseh uporabljenih materialov ne, vseh uporabljenih materialov ne presega*
	Janes
1	*:), ;), :d, :p, :-), #link, :)), slo., :(,* 😊
2	*v slo., v lj., p. s., for the, on the, to the, this is, to be, is a, is the*
3	*hvala za odgovor, tole je pa, še malo pa, ha ha ha, na to temo, me zanima če, zanima me če, je možno da, všeč mi je, in lep pozdrav*
4	*sledi oglasnik tip 1, km h zračni tlak, da ne bo pomote, people followed me and, 4 people followed me, a veš tisto ko, se mi ne da, ne zamudite ugodne ponudbe, sem mislil da je, ne da se mi*
5	*a new photo to facebook, i posted a new photo, posted a new photo to, unfollowed me automatically checked by, followed me automatically checked by, one person unfollowed me automatically, person unfollowed me automatically checked, photos on facebook in the, on facebook in the album, people unfollowed me automatically checked*

Tabela 8: Seznam 10 najpogostejših unikatnih n-gramov posameznih dolžin na prilagojenem frekvenčnem seznamu izbranih korpusov za n-grame normaliziranih pojavnic dolžine 1–5 besed brez upoštevanja ločil, ki se v korpusu pojavijo v vsaj 2 različnih besedilih in vsaj 10-krat na milijon pojavnic.

6. Zaključek in nadaljnje delo

V prispevku smo predstavili postopek luščenja n-gramov iz korpusov slovenskega jezika z namenom izdelave frekvenčnih seznamov korpusnih pojavnic različnih tipov in dolžin, pri čemer izdelano programsko orodje poleg modula za izdelavo običajnega frekvenčnega seznama vseh n-gramov vključuje še modul za njegovo nadaljnje filtriranje ter modul za izdelavo skupnega t. i. prilagojenega frekvenčnega seznama, ki pri štetju n-gramov upošteva medsebojno vsebovanost nizov različnih dolžin.

Ti seznami predstavljajo pomemben doprinos na področju jezikovnih virov za slovenščino, ki raziskovalce in druge potencialne uporabnike razbremenjujejo časovno potratne obdelave korpusnih baz ter jim omogočajo številne možnosti nadaljnjih analiz in aplikacij. Smiselnost nadaljnjih raziskav na podlagi tovrstnih frekvenčnih seznamov navsezadnje potrjujejo tudi prve ugotovitve predstavljene kvantitativne primerjave najpogostejših besednih n-gramov v štirih različnih referenčnih korpusih slovenskega jezika, ki razkrivajo pomembne podobnosti in razlike v naboru stalnega besedišča, njegovi pogostosti in raznovrstnosti, tudi z vidika nezanemarljivega deleža večbesednih enot.

Z jezikoslovnega vidika te ugotovitve spodbujajo predvsem nadaljnje raziskave formulaičnosti različnih jezikovnih zvrsti in njihovih podtipov (Biber, 2009; Simpson-Vlach in Ellis, 2010), podrobnejše analize lastnosti najpogostejših (stalnih) besednih nizov (Biber et al., 2004), s splošnejšega leksikološkega in kognitivnega vidika pa tudi vprašanja, povezana s shranjevanjem in priklicem besedišča v različnih sporazumevalnih okoliščinah (Schmitt, 2004).

Poleg uporabnosti v teoretičnem jezikoslovju, leksikografiji in jezikovni didaktiki pa so tovrstni frekvenčni seznami koristni tudi za razvoj jezikovnih tehnologij za slovenščino, zlasti tistih, ki temeljijo na podatkovnem modeliranju jezikovne rabe (Jurafsky in Martin, 2009), pri čemer rezultati naše analize kažejo, da je pri njihovem načrtovanju nujno upoštevati formulaičnost človeške komunikacije (pogostost večbesednih n-gramov) in njeno zvrstno raznolikost (velik delež unikatnih n-gramov v vsakem korpusu).

Z namenom dosega čim večjega nabora potencialnih uporabnikov in upoštevanja njihovih specifičnih raziskovalnih potreb nameravamo opisani postopek v prihodnosti implementirati v računalniško učinkovitejše prostodostopno spletno orodje,[14] ki bi ga bilo smiselno nadgrajevati tudi z dodatnimi funkcionalnostmi. Poleg možnosti obdelave korpusov v standardnem zapisu TEI XML se kot prioritetna denimo kaže potreba po dodajanju modulov za leksikalno filtriranje (dodajanje leksikona nezaželenih pojavnic) ter modulov za izpis dodatnih izkorpusnih metapodatkov (npr. podatkov o oblikoskladenjskih oznakah) in statističnih izračunov (npr. kolokabilnosti besed v besednih nizih).

7. Zahvala

Raziskavo je delno sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru programa usposabljanja mladih raziskovalcev (1000-15-2923) in nacionalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256).

8. Literatura

Douglas Biber. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3): 275–311.

Douglas Biber, Susan Conrad in Viviana Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25, str. 371–405.

Joaquim Ferreira da Silca in Gabriel Pereira Lopes. 1999. A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora. V: *Proceedings of the 6th Meeting on the Mathematics of Language*, str. 369–381.

Kaja Dobrovoljc. 2018a. Leksikalne prvine govorjenega jezika v uporabniških spletnih vsebinah: primer večbesednih diskurznih označevalcev. *Doktorska disertacija.* Ljubljana: Filozofska fakulteta Univerze v Ljubljani.

Kaja Dobrovoljc. 2018b. Janes corpus n-grams 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1192.

Kaja Dobrovoljc. 2018c. Kres corpus n-grams 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1193.

Kaja Dobrovoljc. 2018d. IMP corpus n-grams 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1194.

Kaja Dobrovoljc. 2018e. Gos corpus n-grams 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1195.

Kaja Dobrovoljc, Tomaž Erjavec in Simon Krek. 2015.. Leksikon besednih oblik Sloleks in smernice njegovega razvoja. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 80–105. Znanstvena založba Filozofske fakultete, Ljubljana.

Tomaž Erjavec. 2013. Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0*, 1(1): 24–49.

Darja Fišer, Tomaž Erjavec in Nikola Ljubešić. 2016. JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0*, 4(2): 67–100.

Polona Gantar. 2015. *Leksikografski opis slovenščine v digitalnem okolju.* Znanstvena založba Filozofske fakultete, Ljubljana.

Polona Gantar, Simon Krek, Iztok Kosem in Vojko Gorjanc. 2015. Collocation dictionary for Slovene: challenge for automatic extraction of data and crowdsourcing. V: *Proceedings of Europhras 2015*, str. 87–89.

Vojko Gorjanc. 2005. *Uvod v korpusno jezikoslovje*. Izolit, Domžale.

Dan Jurafsky in James H. Martin. 2009. *Speech and Language Processing*. Upper Saddle River, ZDA: Prentice-Hall.

Iztok Kosem in Darinka Verdonik. 2012. Key word analysis of discourses in Slovene speech: differences and similarities. *Linguistica*, 1(52): 309–322.

Nikola Ljubešić, Kaja Dobrovoljc in Darja Fišer. 2015. *MWELex – MWE Lexica of Croatian, Slovene and Serbian Extracted from Parsed Corpora. *Informatica*, 39(3): 293–300.

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

Nataša Logar, Polona Gantar in Iztok Kosem. 2014. Collocations and examples of use: a lexical-semantic approach to terminology. *Slovenščina 2.0*, 2(1): 41–61.

Xueqiang Lü, Le Zhang in Junfeng Hu. 2005. Statistical substring reduction in linear time. V: *Natural Language Processing – IJCNLP 2004*, str. 320–327.

[14] Tako orodje za širšo statistično analizo referenčnih korpusov že nastaja v okviru aktualnega nacionalnega projekta Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256).

Makoto Nagao in Shinsuke Mori. 1994. A new method of N-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. V: *COLING '94 Proceedings of the 15th conference on Computational linguistics - Volume 1*, str. 611–615.

Matthew Brook O'Donnell. 2010. The adjusted frequency list: A method to produce cluster-sensitive frequency lists. *ICAME Journal* 35: 135–170.

Rita Simpson-Vlach in Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics* 31(4), str. 487–512.

John Sinclair. 1991. *Corpus, Concordance, Collocation.* Oxford University Press.

Norbert Schmitt. 2004. *Formulaic sequences: acquisition, processing and use.* Amsterdam: John Benjamins Publishing.

Tomaž Erjavec. 2015. The IMP historical Slovene language resources. *Language resources and evaluation* 49(3): 753−775.

Darinka Verdonik in Mirjam Sepesy Maučec. 2017. A speech corpus as a source of lexical information. *International Journal of Lexicography* 30(2): 143−166.

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

Alison Wray. 2005. *Formulaic Language and the Lexicon*. Cambridge University Press.

Ana Zwitter Vitez in Darja Fišer. 2015. Elementi interakcije v govorjenih in spletnih besedilih. V: *Zbornik konference Slovenščina na spletu in v novih medijih*, str. 87–90.

Razvoj smernic za predajo in arhiviranje kvalitativnih podatkov v Arhivu družboslovnih podatkov

Maja Dolinar,* Janez Štebe,† Sonja Bezjak‡

* Arhiv družboslovnih podatkov, Fakulteta za družbene vede, Univerza v Ljubljani
Kardeljeva ploščad 5, 1000 Ljubljana
* maja.dolinar@fdv.uni-lj.si
† janez.stebe@fdv.uni-lj.si
‡ sonja.bezjak@fdv.uni-lj.si

Povzetek

Namen članka je predstaviti smernice za prevzem in arhiviranje kvalitativnih raziskav v Arhivu družboslovnih podatkov (ADP), ki smo jih pripravili kot dopolnilo obstoječim delovnim postopkom arhiviranja raziskav. V ta namen smo oblikovali smernice za prevzem in arhiviranje kvalitativnih raziskav in priporočila za urejanje kvalitativnih podatkov za raziskovalce in podatkovne arhiviste.

Development of the guidelines for ingest and archiving of qualitative data in the Social Science Data Archives (ADP)

The purpose of the article is to present guidelines for the acquisition and archiving of qualitative research in the Social Science Data Archives (ADP), which we have prepared as a complement to the existing working procedures of archiving studies. To this end, we have developed guidelines for the acquisition and archiving of qualitative research and recommendations for the editing of qualitative data for researchers and data archivists.

1. Uvod

Arhiv družboslovnih podatkov (ADP) se je v svojem več kot dvajsetletnem delovanju v slovenskem prostoru uveljavil kot osrednji repozitorij družboslovnih podatkov, katerega glavne naloge so izvajanje ciljnim uporabnikom namenjenih storitev prevzema, shranjevanja in dostopa do kakovostnih in za različne namene uporabnih raziskovalnih podatkov s področja družboslovja iz Slovenije in širše (glej ADP, 2017a). V skladu s tem poslanstvom dolgotrajne digitalne hrambe ADP in upoštevanjem strategije odprtega dostopa do raziskovalnih podatkov (glej Vlada, 2015 in Vlada, 2017) si prizadevamo razširiti svoje storitve na področja, ki so bila do sedaj zaradi različnih razlogov manj zastopana. Eno od področij, s katerega si prizadevamo uporabnikom ponuditi več podatkov, je kvalitativno raziskovanje, ki je v Sloveniji dobro uveljavljeno, s številnimi raziskovalci in ustanovami (glej Štebe, Hudales in Kragelj, 2011). Oranje ledine smo pričeli z razvojem in dopolnjevanjem delovnih procesov, ki smo jih v 90ih prejšnjega stoletja osnovali na izkušnjah arhivske obdelave kvantitativnih podatkov, v zadnjih letih pa se je pokazala potreba po razširitvi na področje kvalitativnih raziskav in prilagajanju procesov posebnostim kvalitativnih družboslovnih raziskav.

V zadnjem času je več pobud v skupnosti uporabnikov glede ustvarjanja in uporabe različnih vrst in formatov kvalitativnih podatkov. Skepsa o ponovni uporabnosti kvalitativnih podatkov iz začetnih obdobij (Fielding, 2004) se ni potrdila, saj v okoljih z vzpostavljeno podatkovno infrastrukturo opazimo, da raziskovalci kvalitativne podatke uporabljajo raznoliko, bodisi samostojno ali v kombinaciji z drugimi vrstami podatkov (Corti, 2007; Bishop in Kuula-Luumi, 2017). V Katalog ADP je že vključenih nekaj kvalitativnih raziskav. Arhivirali smo jih na enak način, kakor arhiviramo kvantitativne raziskave. To pomeni, da smo prevzeli primarni raziskovalni material kvalitativnih raziskav in ga dolgotrajno ohranili na podoben način kot neobdelane podatke v kvantitativnih raziskavah (npr. odprta vprašanja pri anketah). Kvalitativne raziskave, ki so že vključene v Katalog ADP, so večinoma omejene na besedilne podatke, tipično so to prepisi spraševanj.

Namen članka je predstaviti smernice za prevzem in arhiviranje kvalitativnih raziskav, ki smo jih pripravili kot dopolnilo obstoječim delovnim postopkom arhiviranja raziskav. V članku bomo prav tako predstavili priporočila za pripravo in urejanje podatkovnih datotek za kvalitativne podatke, ki smo jih pripravili za raziskovalce in podatkovne arhiviste, in ki vključujejo navodila in dobre prakse poimenovanja in urejanja posameznih datotek (slike, besedilne, zvočne in video datoteke) ter priporočila za anonimizacijo posameznih tipov kvalitativnih podatkov. Navodila vsebinsko dopolnjujejo priporočila ADP za urejanje podatkovne datoteke (2012), ki so pripravljena za urejanje datotek s kvalitativnimi podatki.

2. Sprejem kvalitativnih raziskav v Katalog ADP

2.1. Vrste kvalitativnih podatkov in merila za sprejem

V primerjavi s kvantitativnimi raziskavami se pri kvalitativnih pogosteje srečujemo z majhnimi vzorci,(od n=1 dalje). Tovrstni podatki so običajno zbrani z intervjuji (globinski ali nestrukturirani, posamezni ali skupinski), terenske dnevnike in zapiske opazovanj, strukturirane ali nestrukturirane dnevnike, osebne dokumente, fotografije ipd. Kvalitativno raziskovanje razumemo kot katerokoli raziskovanje ljudi, pri katerem podatki niso bili kvantificirani (npr. spremenjeni v številke in razvrščeni v tabelo, podatkovno bazo ali kvantitativni statistični program) (Corti, 2007). Izhajajoč iz te definicije lahko posamezne raziskave proizvedejo različne raziskovalne podatke, ki so primerni za arhiviranje (npr. podatki družbenih omrežij, komentarji na spletnih blogih, skice terenskih opazovanj, zvočni posnetki ipd.). Za ADP so v skladu s terminologijo vrst podatkov CESSDA ERIC (glej

DDI, 2018) za dolgotrajno arhiviranje zanimive štiri skupine kvalitativnih podatkov, in sicer:

- Besedilni podatki: lahko vključujejo črke, številke, posebne znake ali simbole. Tovrstni podatki so lahko transkripcije intervjujev, zgodbe ali eseji, ki so jih zapisali raziskovalni subjekti, časopisni članki, besedila s spletnih strani, besedila iz blogov, besedila iz družbenih omrežij itd.
- Mirujoča slika: je digitalna slika (fotografija) brez besedila (glasbeni tiski, zemljevidi, fotografije, gravure/odtisi, risbe, plakati, razglednice, slike, grafike, druga dvodimenzionalna umetniška dela).
- Avdio gradivo: vključuje zvočne posnetke, npr. posnetki intervjujev, fokusnih skupin, glasbeni posnetki itd.
- Avdiovizualno gradivo: vključuje premikajoče se podobe. Lahko vključuje filme, animacije, digitalne posnetke, vizualne podobe simulacij, posnete televizijske programe ipd. Posnetki lahko vključujejo zvok ali ne.

V skladu s politiko gradnje Kataloga ADP, v ADP dolgotrajno hranimo raziskovalne podatke, zanimive za družboslovne analize, s poudarkom na problemih, povezanih s slovensko družbo. Pri tem upoštevamo še dva osnovna kriterija: potencial nadaljnje uporabe podatkov in skladnost s strategijo odprtega dostopa do podatkov. Raziskave morajo ustrezati naslednjim merilom ADP za sprejem podatkov (glej ADP, 2017b):

- morajo biti podatki vsebinsko bogati v smislu ustreznosti konceptualizacije in tematskega dopolnjevanja zbirke ADP,
- uporabljene metode morajo biti izpopolnjene, podatki celoviti in ustrezni ter podkrepljeni z ustrezno dokumentacijo, ki omogoča nadaljnje analize,
- dajalec mora avtorsko razpolagati s podatki in biti pripravljen podatke izročiti arhivu za razširjanje.

V ADP kvalitativnim raziskovalnim podatkom določimo prioriteto in presodimo njihov pomen z vidika stroškov dolgotrajne hrambe glede na oceno njihove prihodnje uporabnosti (glej Corti 2007). Prioritete določimo glede na:

- zgodovinsko in kulturno vrednost podatkov, enkratnost podatkov, pomen za dopolnjevanje Kataloga ADP,
- znanstveno relevantnost: dimenzionalnost, izčrpnost zajema pojava in primerljivost, metodološka izvrstnost,
- relevantnost za pedagoške (primerni za izobraževanje)) in druge namene (primerni za državljansko znanost, popularizacijo, novinarsko rabo ipd.).

Posebnost presoje kakovosti in dolgotrajne uporabnosti kvalitativnih podatkov bomo podrobneje izpostavili v 3. poglavju.

2.2. Prevzem kvalitativnih podatkov in njihovo arhiviranje

V ADP najprej pregledamo vlogo za predajo raziskave (obrazec Evidentiranje raziskave na spletni strani, s katerim nas dajalec obvesti o pripravljenosti predaje raziskave) in presodimo, ali predani podatki ustrezajo merilom. Če je raziskava primerna za prevzem, pozovemo dajalca, da pripravi vsa potrebna gradiva za predajo. Podatkovni arhivist nudi strokovno pomoč, dajalcu pa so na voljo vodiči in dodatna priporočila za pripravo podatkov, pri čemer raziskovalcem priporočamo seznanitev s »Priporočili za urejanje podatkovnih datotek kvalitativnih raziskav« (2018a), ki sledi dobrim praksam sorodnih organizacij (UKDA, 2018; FSD, 2018; FORS, 2018) in mednarodnim smernicam.

V fazi prevzema raziskave je pomembno dobro sodelovanje med dajalcem in podatkovnim arhivistom, kajti le tako je mogoče opraviti kvaliteten opis in zagotoviti razumljivost in dolgoročno uporabnost podatkov za širše namene.

Pričakuje se, da so podatki in celotna dokumentacija v digitalni obliki. To lahko vključuje besedila, slike, avdio ali video posnetke. V kolikor podatki ali katero od gradiv niso v digitalni obliki, jih mora dajalec digitalizirati, v kolikor je mogoče (glej ADP, 2018a za priporočila digitalizacije različnih tipov gradiv). Pridružujemo si pravico, da zavrnemo arhiviranje projektov, ki zahtevajo digitalizacijo v velikem obsegu. Ob prevzemu razjasnimo tudi tveganja in možnosti glede zaščite oz. razkritja identitete udeležencev raziskav oz. morebitnih poslovnih skrivnosti.

2.2.1. Zaupnost podatkov, anonimizacija, informirano soglasje in dostop

V ADP pričakujemo, da bo dajalec anonimiziral podatke že pred predajo raziskave v arhiv. Le izjemoma namreč opravljamo anonimizacijo v ADP. V skladu s Politikami dolgotrajne hrambe ADP (glej ADP, 2017d) omogočamo zaupnost podatkov na tri načine:

1. Osnovni anonimizacijski postopki in odstranitev informacij o udeležencih raziskave (npr. odstanitev osebnih identifikatorjev v transkriptih, zamegljeni obrazi pri fotografijah ipd.), ki bi lahko identificirali posameznika:
2. Omejen dostop do raziskave na raziskovalce, ki pripadajo raziskovalni instituciji;
3. Pogodbe z uporabniki, ki jasno določajo posebne pogoje uporabe, vključujoč pravilno rabo podatkov in spoštovanje zaupnosti posameznikov.

Obstoječa zakonodaja in pravila etičnega raziskovanja zahtevajo od raziskovalcev, da že tekom raziskave pridobijo soglasja za sodelovanje pri raziskavi od udeležencev raziskave (v tem smislu spodbujamo raziskovalce k pripravi načrtov za ravnanje z raziskovalnimi podatki, v katerem med drugim predvidijo tudi bodoče dileme glede varovanja zasebnosti in morebitne potrebe po kasnejši anonimizaciji podatkov). Ob preučevanju možnosti prevzema raziskave v Katalog ADP, v ADP preverimo, ali je raziskovalec pridobil soglasja udeležencev za sodelovanje pri raziskavi. Za prevzem podatkov v ADP mora dajalec ustrezno anonimizirati podatke.

Dajalci podatkov morajo ob predaji raziskave v ADP predati vzorec soglasja oz. izjavo, na kakšen način je pridobili tovrstno soglasje (v primeru pridobitve ustnega soglasja), vso morebitno komunikacijo z udeleženci raziskave, ki se nanaša na zaupnost, ter protokole anonimizacije.

2.2.2. Anonimizacija podatkov

Za zagotovitev zaupnosti podatkov je anonimizacija kvalitativnih podatkov nujna v primerih, ko bi se z razkritjem povzročila škoda udeležencu raziskave, ali če ne obstaja soglasje udeležencev za sodelovanje pri raziskavi, ki določa pogoje deljenja podatkov. Anonimizacija je utemeljena, dokler le-ta ne zmanjša analitične vrednosti podatkov. Dajalci morajo pred predajo podatkov v ADP le-te kolikor je mogoče anonimizirati, spremembe pa ne smejo zmanjšati kvalitete in uporabnosti podatkov. Anonimizacija transktriptov bi tako morala vsebovati odstranitev vseh direktnih identifikatorjev, kot so imena in priimki, naslovi, telefonske številke, e-poštni naslovi. Anonimizacija objav na družbenih medijih bi morala vsebovati odstranitev imen uporabnikov (zameglitev v primeru slike zaslona). Imena je potrebno zamenjati s psevdonimi, pri tem pa vse spremembe na podatkih beležiti v ločeni tabeli (protokol anonimizacije), ki povezuje originalne lastnosti s psevdonimi.

Anonimizacija zvočnih in avdio-vizualnih gradiv je bolj zapletena, saj vpliva običajno na kakovost in primernost ponovne rabe tovrstnih podatkov. Anonimizacije zvočnih in avdio-vizualnih gradiv ne priporočamo, v teh primerih se raje odločamo za omejen dostop do teh datotek (npr. dostop pod posebnimi pogoji, varna soba ipd.) (glej točko 4 za možnosti dostopa).

Ob pregledu predanih podatkov podatkovni arhivist preveri, če so podatki ustrezno anonimizirani. Anonimizacija je delo in odgovornost dajalca, zato podatkovni arhivist ni odgovoren za anonimizacijo kvalitativnih datotek in tudi ne odgovarja v primeru identifikacij prepoznav (pravila ureja izjava o izročitvi). Lahko pa svetuje dajalcu in opozori na morebitne pomanjkljivosti.

2.2.3. Obveščeno soglasje za ponovno uporabo

Iz etičnega in pravnega vidika bi moral raziskovalec ob zbiranju podatkov udeležence nedvoumno vprašati, ali se strinjajo, da bodo pridobljeni podatki arhivirani v ADP in dostopni za ponovno uporabo. Obstajajo seveda primeri, ko ni možno tovrstnega soglasja pridobiti, kar je odvisno od raziskovalne teme in raziskovalnega načrta. Raziskovalcem priporočamo, da pridobijo soglasje od udeležencev »za uporabo pridobljenih informacij za raziskovalni namen« ali katerokoli variacijo tega zapisa.

Pridobitev obveščenih soglasij je pogoj za prevzem gradiv v ADP (lahko obstajajo izjeme), morebitna nepridobitev soglasij pa nujno vpliva na pogoje dostopa in uporabe tovrstnih podatkov.

2.3. Priprava podatkov in dokumentacije

Minimalno gledano mora biti dokumentacija o raziskavi in podatkih takšna, da še vedno omogoča novemu uporabniku razumevanje in možnost primere uporabe originalnih podatkov. zbiranja podatkov, potrebne za korektno in ustrezno interpretirati interpretacijo podatkov.

V ta namen raziskovalec oz. dajalec podatkov izpolni obrazec *Opis raziskav*e (glej ADP, 2018c), ki je enak tako za kvalitativne kot kvantitativne raziskave. Za opis raziskave v ADP uporabljamo mednarodni dokumentacijski standard DDI (Data Documentation Initiative) (DDI, 2018), ki je enak za vse vrste podatkov in omogoča iskanje po Katalogu ADP in vključitev v mednarodne kataloge.

V obrazcu *Opis raziskave* dajalec poleg osnovnih informacij o raziskavi (naslov, datum raziskave, avtorji, finančna podpora) med drugim določi ključne besede, vsebinska področja, opredeli čas zbiranja podatkov in geografsko pokritje, populacijo, vrsto podatkov, podrobno opiše vzorčenje in situacijo zbiranja podatkov ter morebitne postopke čiščenja in urejanja podatkov. Opisu doda tudi ostala gradiva, npr. dovoljenje naročnika za deljenje podatkov, podatkovno_e datoteko_e, primer informacijskega dokumenta, primer soglasja udeležencev v raziskavi, vprašalnik ali drug inštrument za zbiranje podatkov, protokol zbiranja podatkov, pokazne kartice, navodila anketirancu, navodila za anketarje, formularje o poteku raziskave, zloženke, pisma izbranim v vzorec, kodirno knjigo, šifrant, raziskovalno poročilo in seznam morebitnih publikacij, ki temeljijo na predanih podatkih.

Ključno za ponovno rabo podatkov kvalitativnih raziskav je podroben opis podatkovnih datotek. Opis datoteke mora minimalno vsebovati format datoteke, značilnosti udeležencev ali katerekoli druge ključne informacije (npr. jezik intervjuja). Posamezni transkript naj bi tako na primer vseboval vsaj:

- identifikator udeleženca,
- glavo dokumenta, ki opisuje situacijo zbiranja podatkov: datum, kraj, ime spraševalca in podrobnosti o intervjuvancu,
- enoten izgled podobnih dokumentov v raziskavi,
- jasne oznake, kaj je vprašanje in kaj odgovor,
- psevdonime, ki nadomeščajo osebne identifikatorje,
- številke strani.

Dodatno je priporočljivo, da dajalci podatkov za vsako kvalitativno podatkovno datoteko izpolnijo obrazec *Opis podatkov*, ki je prilagojen vrsti podatkovne datoteke (besedilni podatki, mirujoča slika, zvočno gradivo, avdiovizualno gradivo).

Naslov datoteke	
Opis	
URL povezava*	
Dolžina posnetka**	
Avtor/ji	
Dajalec (contributor)	
Organizacija	
Datum nastanka	
Kraj nastanka	
Obdobje zbiranja	
Vsebinsko področje	*Zaprt seznam CESSDA Vsebinska področja*
Format datoteke	
Vrsta podatka (CESSDA CV)	Besedilni podatki
Jezik datoteke	
Primarni vir	
Licenca	

*v primeru podatkovne datoteke, ki vsebuje podatke iz spletne strani, bloga

**v primeru podatkovne datoteke, ki vsebuje zvočne podatke

Tabela 1: Obrazec za opis datoteke (primer).

Zaradi lažje hrambe in uporabe e-gradiva je treba vsako datoteko ali skupino datotek opremiti tudi s tehničnimi metapodatki, ki so pomembni za njeno nadaljnjo uporabo (npr. format, s katerim orodjem je bila datoteka ustvarjena, prisotnost varnostnih vsebin ipd.).

Raziskovalci čedalje pogosteje uporabljajo različne programske rešitve za organizacijo in analizo kvalitativnih podatkov, kakršna sta *Nvivo* in *Atlas.ti*. Tovrstne aplikacije (programske rešitve za organizacijo in analizo kvalitativnih raziskav - CAQDAS) so pripomoček raziskovalcem pri organizaciji in upravljanju z velikimi količinami kvalitativnih podatkov – intervjuji, slikami, fotografijami, diagrami in celo avdio-vizualnimi in/ali zvočnimi posnetki.

Raziskovalec lahko v tovrstnem programu posameznim datotekam dodaja zapiske, komentarje in zabeležke ter kodirne besede, ki jih oblikuje tekom analize. Kodiranje je ključno orodje za beleženje analitičnih misli o podatkih in je način, kako raziskovalci razvijejo svoje razumevanje in interpretacijo podatkov. V ADP težimo k temu, da tovrstne informacije hranimo skupaj z raziskovalnimi podatki, saj to prispeva k boljšemu razumevanju in ponovni interpretaciji podatkov. Dajalce, ki uporabljajo tovrstne programske rešitve, zato spodbujamo, da ob predaji ostalih datotek, predajo tudi izvoz datoteke programov za kvalitativno analizo podatkov.

3. Obdelava in objava prevzetih raziskav v ADP

V primeru, ko dajalec izpolnjuje pravne in etične zahteve ter ima pristanek naročnika za deljenje podatkov, nadaljujemo s postopkom predaje oz. prevzema gradiv. Vsa prejeta gradiva v ADP najprej skrbno pregledamo, pri čemer se osredotočimo predvsem na popolnost dokumentacije po metapodatkovnem standardu DDI, vsebinsko bogastvo in zanimivost za drugo rabo, kakovost metodologije zbiranja podatkov in prepisov, medtem ko preverimo obstoj pristanka za sodelovanje in upoštevanje anonimizacije gradiv ter skladnost predanih formatov s priporočili že v fazi pred-prevzema (glej Priporočene in druge oblike zapisov posameznih vrst gradiv za predajo; ADP, 2017c). Presoja kakovosti in bodoče uporabnosti kvalitativnih podatkov obsega posebne vidike, ki jih bomo postopno z izkušnjami vključevali v fazo ocenjevanja in izbora podatkov za arhiv. Med drugim – namesto statistične napake vzorčenja, veljavnosti in zanesljivosti – izvedemo presojo ustreznosti izbora enot glede na pristop, refleksivnosti in prepričljivosti zbranih podatkov (glej Toolbox, 2018).

Nato sledi postopek arhiviranja raziskav, ki smo ga prilagodili posebnostim arhiviranja kvalitativnih raziskav. Pri tem se soočamo z izzivi, ki so povezani s prevzemom gradiv v lastniških formatih programov za analizo kvalitativnih podatkov. Prav tako so etične in pravne dileme glede varovanja avtorskih pravic in osebnih podatkov pri kvalitativnih podatkih lahko še bolj izrazite zaradi običajne večje občutljivosti vsebin, nezmožnosti popolne anonimiziacije, uporabe vsebin z nerazčiščenimi pravicami licenc itd., čemur je potrebno nameniti še posebno pozornost. Kvalitativno raziskovanje obsega širok spekter pristopov (opazovanje, osebno, skupinsko spraševanje, uporaba obstoječih virov itd.), kar prispeva k bolj kompleksni obravnavi podatkov. Dokument »Priporočila za prevzem kvalitativnih raziskav (ADP)« (2018) vključuje informacije o vrstah kvalitativnih podatkov in priporočila za pripravo podatkov za podatkovne arhiviste, definira tipe formatov podatkovnih datotek (prevzemni, arhivski in distribucijski formati) (glej Tabelo 4) ter določa potrebne metapodatke za posamezne vrste kvalitativnih podatkov. Pri standardizaciji in integraciji z drugimi informacijskimi storitvami smo uporabili »Smernice za zajem, dolgotrajno ohranjanje in dostop do kulturne dediščine v digitalni obliki« (Krstulovič, Hajtnik in Doma, 2013) ter dokumentirane dobre prakse sorodnih organizacij (npr. UKDA, 2018; FSD, 2018; FORS, 2018), v katerih so opredeljeni standardi in formati za tipične oblike multimedijskih dokumentov, kot so besedila, slike, filmi ipd.

ADP gradiva nadzira in jih ob spremembi formatov za dolgoročno hrambo prilagodi svojim potrebam, prav tako pa poskrbi tudi, da so vsa gradiva in metapodatki dolgotrajno strojno berljivi. Funkcija arhivske hrambe vključuje tudi številne varnostne mehanizme, kot so postopki preverjanja napak v paketu, ovrednotenje priprave gradiv za dolgotrajno hrambo, kot tudi politike ravnanja v primeru uničenja gradiv.

Po objavi v Katalogu ADP so metapodatki in ostala gradiva, povezana z raziskavo, brez registracije dostopna vsem obiskovalcem spletne strani ADP. Medtem ko se je za dostop do podatkov praviloma potrebno predhodno registrirati.

4. Dostop in uporaba podatkov

Arhivirani kvalitativni podatki so namenjeni ponovni uporabi, ali primerjavi z drugimi podatkovnimi viri. To je že dobro uveljavljena tradicija družboslovnih ved, čeprav opažamo, da raziskovalci nimajo vedno ustreznih metodoloških znanj za tovrstne analize (glej Corti, 2007; Bishop in Arja Kuula-Luumi, 2017). Ponovna uporaba kvalitativnih podatkov omogoča preučevanje surovih podatkov preteklih raziskav, pri čemer lahko raziskovalec dobi pomembne metodološke ali teoretične vpoglede v svoj raziskovalni problem, ki jih morebiti ne bi sam predvidel. Ker je zbiranje novih podatkov običajno drago (v finančnem in časovnem smislu), je uporaba že obstoječih virov ekonomsko smiselna.

Corti in Thompson (2004) opisujeta štiri pristope ponovne uporabe podatkov, ki jih definirata v smislu teoretičnih izzivov ponovne rabe in dejanskih izkušenj raziskovalcev. Ponovna uporaba kvalitativnih podatkov se bistveno ne razlikuje od ponovne uporabe anketnih podatkov (Corti 2007):

- ponovna analiza ali sekundarna analiza podatkov: ponovna interpretacije podatkov ali oblikovanje novih raziskovalnih vprašanj,
- raziskovalni načrt in metodološki napredek: preučevanje raziskovalnih metod pretekle raziskave,
- verifikacija rezultatov: ponovna preučitev zaključkov raziskave,
- poučevanje in učenje: uporaba podatkov za učenje metodologije in priprave raziskovalnih načrtov.

V skladu s strategijo odprtega dostopa in akcijskim načrtom (glej Vlada, 2015 in Vlada, 2017) v ADP težimo k temu, da bi bile raziskave prosto dostopne končnim uporabnikom Kataloga ADP. Pri tem sledimo ideji odprte znanosti, da so podatki odprti, kolikor je mogoče in zaščiteni, kolikor je nujno potrebno. Dajalec podatkov ob predaji podatkov v ADP podpiše *Izjavo o izročitvi* (ADP, 2018e), s katero določi pogoje dostopa do svoje raziskave. Ob predaji dajalci s podpisom pogodbe potrdijo, da imajo pravice razpolaganja s podatki, da zagotavljajo zaupnost podatkov ter opredelijo licence, pod katerimi se podatki in ostala dokumentacija lahko distribuira uporabnikom, vključno z morebitnimi izjemami glede dostopa. Podatki in dokumentacija so uporabnikom na voljo pod pogoji Creative Commons 4.0 licenc (glej Creative Commons, 2018) – med katerimi avtor/dajalec izbere ustrezno.

V ADP razlikujemo tri tipe uporabnikov: registrirane raziskovalce, ki lahko dostopajo tudi do manj zaščitenih mikropodatkov, študente, ki lahko dostopajo do večine podatkov iz kataloga ADP, ter komercialne uporabnike, ki lahko dostopajo do omejenega nabora mikropodatkov, ki niso distribuirani pod licenco CCBYNC. Uporabnik se preko opisa raziskave na spletni strani ADP seznani z morebitnimi posebnimi omejitvami in izjemami glede dostopa, ki jih je določil dajalec ob predaji.

V t.i. *prostem dostopu*, kjer ni potrebna registracija, lahko uporabnik dostopa do kataloga ADP, metapodatkov raziskav in podatkov nekaterih prostodostopnih raziskav. Kljub vsemu pa je raba podatkov omejena z zakonodajo, etičnimi pravili stroke in organizacije ter avtorskimi pravicami.

Večina podatkov znotraj Kataloga ADP (tudi kvalitativnih, če so ustrezno anonimizirane) je v ADP dostopna preko t.i. *standardnega dostopa*, ki v primeru kvantitativnih podatkov omogoča analize s pomočjo spletnega vmesnika Nesstar oz. prenos datotek na lokalni računalnik. V primeru dostopa do kvalitativnih podatkov je trenutno možen zgolj prenos datotek na lokalni računalnik, saj je razvoj spletnega orodja za analizo kvalitativnih podatkov stvar prihodnjega razvoja ADP. Za standarden dostop do podatkov je potrebno izpolniti »Registracijski obrazec za dostop do gradiv kataloga ADP« (ADP, 2018d). V obrazcu se uporabnik identificira, opredeli namen uporabe gradiva ter izrazi strinjanje s »Splošnimi določili in pogoji uporabe podatkov«.

Nekatere podatkovne datoteke s kvalitativnimi podatki so izjemoma dostopne pod posebnimi pogoji in je zanje potrebno pridobiti dovoljenje izvirnih avtorjev. Na primer:

- Podatki morda niso v celoti anonimizirani, zato je potrebno posebno varstvo. Gre za t.i. Scientific Use File (SUF). Sem spadajo neanonimizirane podatkovne datoteke kvalitativnih raziskav.
- Avtorji določijo, da bo podatkovna datoteka dostopna s časovnim zamikom, 6 mesecev po objavi raziskave v katalogu ADP, ko se sprosti t. i. embargo.
- Podatkovna datoteka je dostopna le naročniku raziskave in izvirnim avtorjem (dovoljujemo le izjemoma).

Dostop pod posebnimi pogoji zahteva poleg običajne registracije še izpolnjeno »Vlogo za dostop do gradiva na zahtevo« (glej ADP, 2018d) s točno navedbo iskane raziskave ter dodatna pojasnila glede namena rabe. Komisija za zaščito zaupnosti ADP obravnava vlogo in uporabnika najkasneje v 7 delovnih dneh po oddaji vloge obvesti o odločitvi glede možnostih dostopa. Komisija za zaščito zaupnosti lahko uporabniku dodeli dostop do zahtevanega gradiva preko varne povezave ali preko varne sobe, kjer uporabnik pred dostopom podpiše posebno pogodbo, ki ureja pravice in obveznosti uporabe zahtevanih raziskovalnih podatkov. Gre za t.i. Secure Use File (ScUF).

Vrsta podatkov	**Prevzem**		**Formati za arhivsko hrambo**	**Formati za dostop**
	Priporočeni formati	**Sprejemljivi formati**		
Besedilni podatki	Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	Hypertext Mark-up Language (.html)	• PDF/A • W3C XML • ODF Pri spletnih vsebinah: • HTML • WARC	• PDF/A • HTML
Mirujoča slika	TIFF 6.0 uncompressed (.tif)	JPEG (.jpeg, .jpg, .jp2), če je datoteka izvorno v tem formatu GIF (.gif) TIFF druge verzije (.tif, .tiff) RAW slikovni format (.raw) Photoshop datoteke (.psd) BMP (.bmp) PNG (.png) Adobe Portable Document Format (PDF/A, PDF) (.pdf)	TIFF JPEG (stiskanje z izgubami) PNG JPEG2000 SVG 2D v1.1 WRC (vektorske slike) DWG 3D (de facto standard; 3D grafični objekti, vektorski podatki, CAD)	TIFF, JPEG, JPEG2000, PNG, GeoTIFF, JPEG (primerno za prenos), EPS (primerno za prenos), GIF (primerno za prenos).

Zvočno gradivo	Free Lossless Audio Codec (FLAC) (.flac)	MPEG-1 Audio Layer 3 (.mp3), če je datoteka izvorno v tem formatu Audio Interchange File Format (.aif) Waveform Audio Format (.wav)	BWF FLAC MPEG-2 Audio Layer III (stiskanje z izgubami) MPEG-2 Audio AAC (stiskanje z izgubami) MPEG-4 Audio AAC (stiskanje z izgubami)	Mp3 MPEG-2 Audio AAC MPEG-4 Audio AAC Free Lossless Audio Codec (FLAC) Waveform Audio Format (.wav)
Avdio-vizualno gradivo	AVI MPEG-4 (mp4) OGG video (.ogv, .ogg) Motion JPEG2000 (.mj2)	AVCHD video (.avchd) WMF	ANSI/SMPTE 268M (DPX) Motion JPEG2000 MPEG-4 AVC ((stiskanje z izgubami)	Za prenos k uporabniku (download), original na zahtevo ali ogledne kopije, npr.: MPEG-1 AVI WMV Quicktime Ogledne kopije za pretakanje (streaming), npr: ASF WMV Quicktime

Tabela 4: Prevzemni, arhivski in distribucijski formati različnih vrst kvalitativnih podatkov.

5. Zaključek

Vključitev kvalitativnih raziskav z vidika arhiviranja raziskav pomeni troje. Prvič, da se kvalitativni podatki digitalizirajo – skeniranje morebitnih dokumentov, slik in drugih gradiv, ki obstajajo v fizični obliki, in njihovo preoblikovanje v digitalno obliko, ki omogoča dolgotrajno hranjenje in uporabnost (PDF, XML). Drugič, da podatkovno zbirko opremimo z dodatnimi informacijami in gradivi, s čimer pripomoremo k njeni ponovni uporabnosti (opis konteksta zbiranja podatkov, namena raziskave, dodatne publikacije in gradiva). Tretjič, da zagotovimo, da je mogoče podatke najti in do njih dostopati preko spleta, skladno z načeli FAIR (F - *findable*: podatke je mogoče najti, A - *accessible*: podatki so dostopni, I - *interoperable*: podatke je mogoče povezati z drugimi podatki, R- *reusable*: podatke je mogoče ponovno uporabiti.), tako da vključimo raziskave v spletne kataloge in spletne aplikacije za analizo podatkov.

Razširitev vloge ADP je nujna za ohranjanje in izmenjavo dragocenih kvalitativnih podatkov, ki jih proizvajajo slovenski raziskovalni centri (Fakulteta za družbene vede, Filozofska fakulteta, Fakulteta za socialno delo ipd.) (glej Štebe, Hudales in Kragelj, 2011). ADP je edina raziskovalna infrastruktura za arhiviranje in dostop do podatkov družboslovnih raziskav v Sloveniji in je nedavno pridobila certifikat zaupanja vrednega arhiva CoreTrustSeal (CoreTrustSeal, 2018). Kot taka ima trdno podlago za razširitev storitev tudi na področju arhiviranja kvalitativnih družboslovnih raziskav.

Naš obstoječi delovni postopek arhiviranja raziskav smo razširili z značilnostmi arhiviranja kvalitativnih raziskav, česar smo se lotili z oblikovanjem smernic za prevzem in arhiviranje kvalitativnih raziskav in z oblikovanje priporočil za urejanje kvalitativnih podatkov za raziskovalce in podatkovne arhiviste. Postopke, pravila in merila bomo tudi v prihodnje preverjali, dopolnjevali ter nadgrajevali na podlagi izkušenj. Za kvalitativne podatke, ki se jih ne da anonimizirati brez izgube, razvijamo in preizkušamo različne načine nadzorovanega omejenega dostopa. Podatki se lahko razlikujejo tudi z vidika, ali gre za zbirko referenc v povezavi s člankom ali tematiko, ali pa samostojno zbirko. Glede na naštete posebnosti je smiselno v bodoče prilagajati postopke obravnave (glej Elman in Kapszewski, 2013; QDR, 2018), saj zaradi vse večje ozaveščenosti raziskovalcev pričakujemo vedno večje število predanih raziskovalnih podatkov, skladno s tehnološkim in vsebinskim razvojem ADP pa povečujemo tudi možnost arhiviranja različnih vrst podatkov.

6. Literatura

ADP. 2017a. Naše poslanstvo. https://www.adp.fdv.uni-lj.si/spoznaj/adp/poslanstvo/.

ADP. 2017b. Merila za sprejem raziskav. https://www.adp.fdv.uni-lj.si/deli/merila/.

ADP. 2017c. Priporočeni formati. https://www.adp.fdv.uni-lj.si/deli/postopek/priprava/formati/.

ADP. 2017d. Politike digitalne hrambe. https://www.adp.fdv.uni-lj.si/spoznaj/politika/.

ADP. 2018a. Priporočila za urejanje podatkovnih datotek kvalitativnih raziskav. https://www.adp.fdv.uni-lj.si/deli/postopek/priprava/#kvalitativni.

ADP. 2018b. Priporočila za prevzem kvalitativnih raziskav (za ADP) (interno gradivo). ADP, Ljubljana.

ADP. 2018c. Opis raziskave. https://www.adp.fdv.uni-lj.si/deli/postopek/opis_raziskave/.

ADP. 2018d. Registriraj se. https://www.adp.fdv.uni-lj.si/registracija/.

ADP. 2018e. Izjava o izročitvi. https://www.adp.fdv.uni-lj.si/deli/postopek/izjava/.

Bishop, Libby in Arja Kuula-Luumi. 2017. Revisiting Qualitative Data Reuse: A Decade On. *SAGE* Open, 7(1):1–15.

CoreTrustSeal. 2018. Implementation of the CoreTrustSeal for ADP. https://www.coretrustseal.org/wp-content/uploads/2018/01/ADP-Social-Science-Data-Archives.pdf.

Corti, Louise. 2017. Re-using archived qualitative data – where, how, why? *Archiving Science*, 7(37):37–54.

Creative Commons. 2018. Creative Commons 4.0. https://creativecommons.org/licenses/by/4.0/

DDI. 2018. DDI Controlled Vocabulary for General Data Format. http://ddi.icpsr.umich.edu/Specification/DDI-CV/GeneralDataFormat_2.0.html.

Elman, Colin in Diana Kapiszewski. 2013. A Guide to Sharing Qualitative Data. Qualitative Data Repository (QDR), Center for Qualitative and Multi Method Inquiry (CQMI), Syracuse University, Syracuse. https://qdr.syr.edu/sites/default/files/QDR-A_Guide_to_Sharing_Qualitative_Data_v1.3.pdf.

Fielding, Nigel. 2004. Getting the most from the archived qualitative data: epistemological, practical and professional obstacles. *International Journal of Social Research Methodology*, 7(1):97–104.

FORS. 2018. Policy on Archiving Qualitative Data. http://forscenter.ch/en/data-and-research-information-services/2221-2/qualitative-data/.

FSD. 2018. Processing Qualitative Data Files. http://www.fsd.uta.fi/aineistonhallinta/en/processing-qualitative-data-files.html.

Krstulović, Zoran, Tatjana Hajtnik in Mitja Doma. 2013. Smernice za zajem, dolgotrajno ohranjanje in dostop do kulturne dediščine v digitalni obliki. Ljubljana: Ministrstvo za kulturo.

Lužar, Sanja, Maja Ojsteršek in Irena Vipavc Brvar. 2012. Priporočila za urejanje podatkovne datoteke. https://www.adp.fdv.uni-lj.si/media/img/datoteke/PriporocilaZaPodatkovnoDatoteko2.pdf.

QDR. 2018. Qualitative Data Repository: Managing Data. https://qdr.syr.edu/guidance/managing.

Štebe, Janez, Jože Hudales in Boris Kragelj. 2011. Archiving and Re-using Qualitative and Qualitative Longitudinal Data in Slovenia. *IASSIST Quarterly*, 34/35:1/4:50–41.

Toolbox. 2018. Quality Criteria for Qualitative Research in Health Sciences. http://wp.unil.ch/qualityofqualitativeresearch/toolbox-3/.

UKDA. 2018. Format your data. https://www.ukdataservice.ac.uk/manage-data/format.

Vlada Republike Slovenije. 2015. Nacionalna strategija odprtega dostopa do znanstvenih objav in raziskovalnih podatkov v Sloveniji 2015–2020. http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/Znanost/doc/Zakonodaja/Strategije/Nacionalna_strategija_odprtega_dostopa.pdf.

Vlada Republike Slovenije. 2017. Akcijski načrt izvedbe Nacionalne strategije odprtega dostopa do znanstvenih objav in raziskovalnih podatkov v Sloveniji 2015–2020. http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/Znanost/doc/Odprti_dostop/Akcijski_nacrt_-_POTRJENA_VERZIJA.pdf.

Prehod iz statističnega strojnega prevajanja na prevajanje z nevronskimi omrežji za jezikovni par slovenščina-angleščina

Gregor Donaj, Mirjam Sepesy Maučec

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

Povzetek

Strojno prevajanje z nevronskimi omrežji je najnovejši pristop k strojnemu prevajanju. V primerjavi s klasičnim statističnim prevajanjem, ki temelji na modelu šumnega kanala, sestavljenega iz množice neodvisnih komponent, pri nevronskem prevajanju učimo en sam model oziroma nevronsko omrežje, ki ga optimiramo v smeri čim kvalitetnejših prevodov. V članku predstavljamo naše prve izsledke nevronskega strojnega prevajanja za jezikovni par slovenščina-angleščina in rezultate primerjamo s klasičnim statističnim prevajanjem. Analiziramo prevajanje v obe smeri z različnimi hiperparametri učenja oz. modelov. Primerjava rezultatov kaže, da lahko z nevronskimi omrežji tvorimo prevode, ki so boljši za 6,2 točke BLEU (smer angleščina-slovenščina) oz. za 2,9 točke BLEU (smer slovenščina-angleščina) v primerjavi s statističnimi prevajanjem.

From statistical machine translation to translation with neural networks for the Slovene-English language pair

Neural machine translation is a newly proposed approach to machine translation. In comparison to the traditional statistical machine translation, which is based on the noisy channel model with many independent components, neural machine translation system is a single neural network trained to optimize the translation performance. In this paper, we present our first experiments with neural machine translation on Slovene-English language pair and compare the obtained results with classical statistical machine translation. Translation in both directions is analyzed with different model and learning hyperparameters. We found that neural machine translation outperforms statistical machine translation by 6.2 BLEU points in the translation from English to Slovene and by 2.9 BLEU points in the translation from Slovene to English.

1. Uvod

Statistično strojno prevajanje (SMT) je še do nedavnega veljalo za najuspešnejši pristop k strojnemu prevajanju. Vsaj 20 let smo lahko sledili kontinuiranemu izboljševanju kvalitete prevodov, ki so jih generirali frazni statistični prevajalniki različnih tipov: klasični frazni, faktorski, hierarhični ali temelječi na sintaktičnih strukturah (Koehn, 2010). V zadnjih nekaj letih pa lahko vidimo, da se je v raziskovalni srenji povečal interes za nevronsko strojno prevajanje (NMT), ki je dotlej veljajo za računsko preveč zahteven pristop. Že prvi rezultati so pokazali, da so NMT prevodi po kvaliteti primerljivi s SMT prevodi, za določene jezikovne pare pa občutno boljši (Junczys-Dowmunt et al., 2016). Zanimivo je, da je izboljšanje najbolj očitno pri najtežjih jezikovnih parih, kot je na primer prevajanje, ki vključuje kitajščino, arabščino in nemščino (Junczys-Dowmunt et al., 2016; Bentivogli et al., 2018). Hiter napredek nevronskega prevajanja izvira iz uporabe ponavljajočih nevronskih omrežij (tudi rekurentnih ali povratnih; ang. recurrent neural network – RNN) in arhitekture kodirnik-dekodirnik (ang. encoder-decoder), ki uporablja več nevronskih omrežij tipa RNN. Takšen pristop so med prvimi predlagali Cho et al. (2014).

Leto kasneje so Bahdanau et al. (Bahdanau et al., 2014) predlagali še rešitev za problem prevajanja dolgih povedi, ki so jo poimenovali mehanizem poudarka (ang. attention mechanism). Da je nevronsko strojno prevajanje “vroča” tema, kaže tudi osrednja konferenca na področju strojnega prevajanja WMT[1], kjer se je leta 2015 med tekmovalnimi MT sistemi prvič pojavil NMT, leto zatem pa je na NMT temeljila velika večina zmagovalnih sistemov. V letu 2017 so med tekmovalne naloge na novo uvrstili tudi učenje NMT.

Zakaj so nevronska omrežja tako učinkovita? Razlog je lahko v njihovi sposobnosti razločevanja in izločanja informacij iz zapletenih vzorcev, ki jih druge tehnike prevajanja ne zaznajo. Kot osnovno enoto uporabljajo poved, kar pomeni, da upoštevajo širši kontekst kot SMT, pri katerem je osnovna enota podatkovno definirana fraza. Prednost je tudi v strukturi NMT sistema, ki je en sam velik model, v katerem se prevodi oblikujejo na kontekstno odvisen način, za razliko od SMT, ki je sestavljen iz povezane množice neodvisnih komponent (model prevajanja, jezikovni model, model preurejanja ipd.).

Naš cilj v tem članku je preizkusiti NMT prevajalnik na jezikovnem paru slovenščina-angleščina in ga primerjati z rezultati SMT prevajalnika. Članek je organiziran v naslednja poglavja: v poglavju 2 predstavimo splošne značilnosti NMT arhitekture in njene osnovne mehanizme. V poglavju 3 predstavimo zasnovo našega NMT sistema. Najprej podamo osnovne podatke učnega korpusa, sledi opis konfiguracije NMT sistema in tudi SMT sistema, ki smo ga uporabili za primerjavo. Rezultati in analiza eksperimentov so v poglavju 4. Članek zaključimo v poglavju 5, kjer povzamemo ključne ugotovitve in podamo smernice za naprej.

[1] `http://www.statmt.org/wmt17/`

2. Prevajanje z nevronskimi omrežji

Nevronsko omrežje je model za obdelavo informacij, ki posnema delovanje živčnega sistema bioloških organizmov. Uči se iz primerov, zato je primeren model tudi za strojni prevajalnik, ki ga zasnujemo iz vzporednega korpusa poravnanih prevodov. Nevronsko omrežje sestavlja množica nevronov, ki jih v procesu učenja prilagodimo izbrani nalogi, v našem primeru nalogi prevajanja.

Obstajajo različne arhitekture nevronskih omrežij. Pri prevajanju se uporabljajo ponavljajoča nevronska omrežja RNN, pri katerih lahko signali potujejo v obe smeri: naprej in nazaj, kar dosežemo z vpeljavo povratnih zank. Ponavljajoča nevronska omrežja so zelo zmogljiva, a tudi zelo zapletena.

Nevronska omrežja so sestavljena iz treh plasti oz. skupin enot: vhodna plast, ena ali več skritih plasti in izhodna plast. Vhodna plast je povezana s skrito plastjo, le-ta pa z izhodno plastjo. Aktivnosti v vhodni plasti predstavljajo vhodno informacijo, ki jo vnesemo v omrežje. V skriti plasti določimo aktivnost vhodne plasti in uteži med vhodno in skrito plastjo. Kako bo odreagirala izhodna plast je odvisno od aktivnosti v skriti plasti in uteži med skrito in izhodno plastjo.

Delovanje nevronskega omrežja je, razen od uteži, odvisno tudi od aktivacijske (tudi pragovne ali prenosne) funkcije, ki povezuje vhod z izhodom. Navadno se uporablja hiperbolični tangens, ki iz neomejenega definicijskega območja slika na interval [-1, 1]. V bolj izpopolnjenjih RNN, imenovanih RNN z vrati (ang. gated RNN), se uporabljajo GRU (ang. gated recurrent unit) enote, ki se prilagajajo različnim odvisnostim (Cho et al., 2014).

Nevronska omrežja so v splošnem omejena s fiksno dolžino vhodnega zaporedja, pri čemer je tudi izhodno zaporedje enake dolžine. Povedi, ki jih prevajamo, pa so različnih dolžin in tudi prevod se običajno v dolžini ne ujema z izvorno povedjo. Lahko ima več ali manj besed. Ta problem rešuje arhitektura kodirnik-dekodirnik (ang. encoder-decoder), kjer so dovoljena vhodna in izhodna zaporedja različnih dolžin. Kodirnik bere vhodno poved in jo kodira v vektorje fiksne dimenzije. Dekodirnik iz teh vektorjev tvori prevod. Kodirnik in dekodirnik za izbrani jezikovni par učimo sočasno, tako da maksimiziramo verjetnost pravilnega prevoda za izbrano vhodno poved.

Kodirnik je dvosmerno ponavljajoče nevronsko omrežje (ang. bidirectional RNN), sestavljeno iz naprej in nazaj usmerjene RNN. Naprej usmerjena RNN bere vhodno poved v pravilnem vrstnem redu, tj. od leve proti desni, in izračuna zaporedje naprej usmerjenih skritih stanj (ang. forward hidden states), medtem ko nazaj usmerjena RNN bere besede v obratnem vrstnem redu, tj. od desne proti levi, in generira zaporedje nazaj usmerjenih skritih stanj (ang. backward hidden states). Na ta način vsako besedo označimo s spetimi naprej in nazaj usmerjenimi skritimi stanji. To pomeni, da vsako besedo opremimo z levim in desnim kontekstom v povedi.

Dekodirnik preiskuje izvorno poved in jo po principu veriženja pogojnih verjetnosti dekodira v prevod. Tudi dekodirnik je RNN, pri katerem si iterativno sledijo tri faze; look-update-generate. V fazi look je izbrano novo skrito stanje. Izračunano je iz treh podatkov: njegovega kontekstnega vektorja, predhodnega skritega stanja in predhodno generiranje besede v prevodu. Sledi faza update, v kateri se generira novi kontekstni vektor. Kontekstni vektor skritega stanja je odvisen od vseh označb, ki jih je generiral kodirnik za celotno izvorno poved. Izračunan je kot utežena vsota teh označb. Prehod v novo skrito stanje ima za posledico tudi generiranje nove besede v prevodu. To fazo imenujemo generate.

Problem omenjenega principa delovanja dekodirnika so dolge povedi in pridruženi vektorji fiksnih dolžin. Posebej problematično je prevajanje povedi, ki so daljše od povedi v učnem korpusu. Rešitev predstavlja mehanizem poudarka (ang. attention mechanism), ki nevronskemu omrežju omogoča učenje poudarka v izvorni povedi, tj. odsekov, ki vsebujejo pomembne informacije za generiranje posamezne besede v prevodu. Mehanizem poudarka je uporabljen med dvema GRU prehodoma dekodirnika (Miceli Barone et al., 2017; Sennrich et al., 2017).

Nevronska omrežja so se šele v zadnjih letih začela bolj pogosto uporabljati na področju strojnega prevajanja. Tako je tudi pred kratkim bilo izpostavljenih nekaj ključnih izzivov pri uporabi nevronskih omrežij (Koehn in Knowles, 2017). Med drugim je izpostavljena problematika dolžine stavkov, ki jih hočemo prevajat. Avtorja sta pokazala, da se kvaliteta prevodov z nevronskimi omrežji poslabša pri stavkih z več kot 60 pojavnicami. Drugi izpostavljen izziv je bilo prevajanje dokumentov izven domene učne množice.

V naši raziskavi smo tako dodali tudi primerjavo med sistemoma SMT in NMT glede na dolžino povedi in rezultate na množici izven domene učnega korpusa.

3. NMT sistem

3.1. Učni korpus

V eksperimentih smo uporabili Europarl korpus[2] (Koehn, 2005). Korpus je sestavljen iz besedil zbornika Evropskega parlamenta. Korpus pokriva 20 jezikovnih parov, pri katerih je en jezik v paru vedno angleščina, kot drugi jezik pa nastopajo jeziki držav članic Evropske unije. Za naše eksperimente smo uporabili vzporedni korpus za jezikovni par slovenščina-angleščina, ki obsega gradivo iz obdobja med letoma 2007 in 2011. Korpus vsebuje 623.490 stavkov, pri čemer je na slovenski strani 12,5 milijonov besed, na angleški pa 15 milijonov. Korpus smo razdelili na učni, razvojni in testni del. Razvojni in testni del obsegata vsak 2000 stavkov, ki smo jih izločili iz konca korpusa. Preostali stavki so v učnem korpus. Učni korpus vsebuje na slovenski strani 144.671 različnih besed, na angleški pa 66.604. Pred učenjem prevajalnikov smo korpus tokenizirali in normalizirali.

3.2. Konfiguracija NMT sistema

Tip modela je ponavljajoče nevronsko omrežje. Uporabljeni modeli temeljijo na arhitekturi omrežja kodirnik-dekodirnik (Bahdanau et al., 2014), kjer se vhodni podatki v nevronsko omrežje najprej preslikajo na enote v skritih plasteh omrežja (kodirajo) in nato preslikajo na izhodne podatke (dekodiranje).

[2] `http://www.statmt.org/europarl/`

	Angleško→slovensko	Angleško→slovensko (dtdn)	Slovensko→angleško	Slovensko-angleško (dtdn)
BLEU	35,5	29,8	44,1	39,1
METEOR	31,1	28,1	41,8	38,1
TER	46,0	52,1	38,9	43,3

Tabela 1: Rezultati prevajanja testne množice s sistemom SMT.

V postopku učenja smo določili hiperparametre modela oz. nevronskega omrežja. Prvi hiperparameter je dimenzija vgrajenih vektorjev v modelih. Ti vektorji predstavljajo besede, njihovo dimenzijo pa smo nastavili na 512. Drugi hiperparameter modela je dimenzija skritega stanja v RNN, ki smo ga nastavili na 1024.

Dodatno k temu smo še omejili dolžino stavkov v učnem korpusu na 50 (tukaj štejemo tako besede in vse ostale pojavnice, npr. ločila), kar pomeni, da se pri učenju daljši stavki odstranijo. V postopku učenja dodatno določimo še velikost mini serije (ang. mini-batch). To je število stavkov iz učne množice, ki se v vsaki iteraciji učenja uporabijo za učenje omrežja – njegovo posodobitev. Kot velikost mini-serije smo izbrali 64.

3.3. Trajanje učenja

Značilnost nevronskih omrežij je, da pri velikem številu iteracij učenja prihaja do prekomernega prilagajanja (ang. overfitting) modela na učno množico. Pri tem pojavu začne model vse bolje izražati primere v učni množici, s tem pa izgubi na splošnosti in slabše deluje na novih podatkih.

Z namenom iskanja optimalnega trajanja učenja za nevronsko omrežje uporabimo razvojni del korpusa. V prvem poskusu učenja smo kot trajanje učenja določilu 500 epoh (prehodov celotnega učnega korpusa). Po porabljenem času na primerljivi strojni opremi (približno 1 teden), je to učenje primerljivo s sistemi drugih raziskovalcev (Junczys-Dowmunt et al., 2016).

Med postopkom učenja smo modele shranjevali na vsakih 10.000 iteracij (posodobitev omrežja glede na eno mini serijo). Med testiranjem smo kasneje iskali optimalno število iteracij oz. epoh učenja.

3.4. Programska oprema

Za učenje modelov in prevajanje smo uporabljali orodje Marian (prej AmuNMT). Orodje AmuNMT (Junczys-Dowmunt et al., 2016) je bilo sprva razvito za hitro prevajanje z uporabo modelov, naučenih z orodjem Nematus (Sennrich et al., 2017). Kasneje je bila razvita ponovna implementacija orodja Nematus v jeziku C++, ki je bila nato združena z AmuNMT in imenovana Marian. Orodje je odprtokodno in prostodostopno[3].

Za tokenizacijo in detokenizacijo ter normalizacijo in denormalizacijo smo uporabljali skripte, ki so sestavni del orodja Moses (Koehn et al., 2003). Preveden tekst smo ocenjevali z orodjem Multeval (Clark et al., 2011) in pri tem vrednotili prevode z metrikami BLEU, METEOR in TER.

3.5. Strojna oprema

Učenje modelov in prevajanje se izvajata le na grafičnem procesorju in grafičnem delovnem pomnilniku. Oba sta del grafične kartice, v našem primeru Nvidia GeForce GTX 1080 Ti. Izkušnje kažejo, da je najpomembnejša lastnost grafične kartice pri učenju nevronskih omrežij pasovna širina za prenos podatkov do grafičnega spomina. V primeru naše grafične kartice je ta 484 GB/s.

Ostala strojno oprema ni bistvena za rezultate ali hitrost učenja oz. prevajanja.

3.6. SMT sistem za primerjavo

NMT sistem smo primerjali s fraznim statističnim prevajalnikom, ki smo ga zgradili na istem korpusu. Slovar prevajalnika je na slovenski strani vseboval 144.671 besed, na angleški pa 66.604. Pri gradnji prevajalnika smo uporabili orodje Moses (Koehn et al., 2003) in standardne nastavitve: besede smo poravnali v zaporedju iteracij IBM modela 1, HMM modela in modelov 3 in 4; uporabili smo "grow-diag-final-and" simetrizacijo; jezikovni model je bil besedni 3-gramski z modificiranim Kneser-Ney glajenjem frekvenc. Za učenje jezikovnih modelov smo uporabili celoten učni korpus. Iz statistike smo izločili besede, ki se pojavijo le enkrat. Perpleksnost jezikovnega modela slovenskega jezika je bila 109, angleškega pa 62.

Ideja raziskave v tem članku je primerjava NMT in SMT prevajalnikov, ki temeljijo izključno na poravnanem korpusu, brez uporabe dodatnih jezikovnih ali kakršnihkoli drugih informacij. Več podatkov o SMT prevajalniku je v (Sepesy Maučec in Donaj, 2016), kjer smo uspešnost osnovnega SMT prevajalnika v nadaljevanju še izboljšali z uporabo jezikovno-specifičnih oznak, ki jih pa v tej raziskavi ne vključujemo.

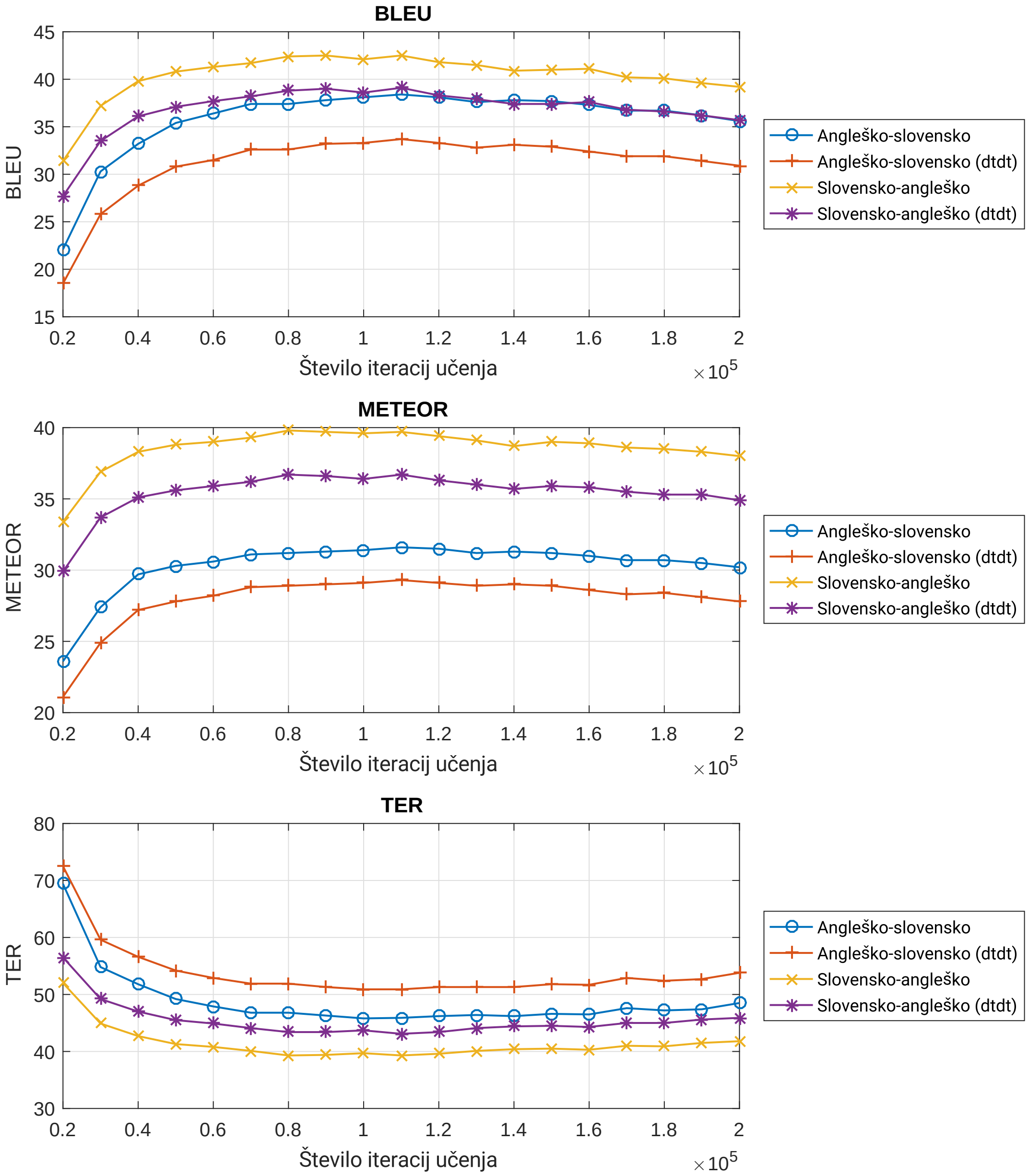
4. Rezultati

4.1. Hitrost učenja

Hitrost učenja modelov je odvisna od hiperparametrov modela in učenja. Pri naših modelih je bila hitrost učenja modela v smeri slovensko-angleško 368 povedi na sekundo, za učenje modelov v smeri angleško-slovensko pa 390 povedi na sekundo. Za učni korpus Europarl je to pomenilo 2,3 epohe na uro (slovensko-angleško) oz. 2,5 epoh na uro (angleško-slovensko). Rezultate na razvojni množici bomo predstavili na prvih 22,5 epohah, za katere je potrebnih približno 9 ur učenja za vsakega izmed obeh modelov.

Hitrost učenja velja za prvotne nastavitve hiperparametrov učenja in modela. Pri spremenjenih hiperparametrih (npr. povečana ali pomanjšana kompleksnost modelov) se čas učenja spremeni. Sprememba dimenzije vektorjev, ki predstavljajo besede ali pa dimenzije skritega stanja, skoraj

[3] `https://marian-nmt.github.io/`



Slika 1: Rezultati metrik BLEU, METEOR in TER na razvojni množici za obe smeri prevajanja s sistemom NMT.

premosorazmerno vpliva na čas učenja. Povečanje števila povedi v mini-seriji pri učenju pa pohitri učenje, vendar je vpliv manj izrazit.

4.2. Rezultati SMT

Najprej smo izvedli učenje modelov in prevajanje s sistemom SMT. Rezultati BLEU, METEOR in TER so prikazani v tabeli 2.. Prikazani so rezultati vseh treh metrik za prevajanje v obe smeri, ki jih dobimo z ocenjevanjem pred detokenizacijo in denormalizacijo, kot tudi po detokenizaciji in denormalizaciji (dtdn). Rezultati nam služijo za primerjavo obeh sistemov.

4.3. Rezultati na razvojni množici

Na sliki 1 so prikazani rezultati metrik BLEU, METEOR in TER, ki jih dobimo na razvojni množici pri različnih trajanjih učenja. Optimalne vrednosti so maksimalni rezultati BLEU in METEOR oz. minimalni rezultati TER. Prikazani so rezultati za trajanja od 20.000 do 200.000 iteracij. Rezultati so prikazani za obe smeri prevajanja tako pred detokenizacijo in denormalizacijo kot tudi po detokenizaciji in denormalizaciji (dtdn). Iz rezultatov lahko razberemo, da imamo pri vseh 4 potekih maksimum metrike BLEU pri 110.000 iteracijah učenja, kar ustreza približno 12 epoham učenja.

	Angleško→slovensko	Angleško→slovensko (dtdn)	Slovensko→angleško	Slovensko→angleško (dtdn)
BLEU	40,8	36,0	46,4	42,7
METEOR	33,2	30,9	41,7	38,7
TER	43,0	48,1	37,1	40,9
Δ BLEU	5,3	6,2	2,3	2,9
Δ METEOR	2,1	2,8	-0,1	0,6
Δ TER	-3,0	-4,0	-1,8	-2,4

Tabela 2: Rezultati prevajanja testne množice s sistemom NMT in primerjava z rezultati, dobljenimi s sistemom SMT (Δ).

	Angleško→slovensko	Angleško→slovensko (dtdn)	Slovensko→angleško	Slovensko→angleško (dtdn)
Batch 16	40,4	35,6	45,2	41,7
Batch 32	40,0	35,0	45,6	41,9
Batch 64	40,8	36,0	46,4	42,7
Batch 128	39,4	34,5	45,6	41,8
Batch 256	39,7	34,8	44,4	40,7
EMB 256	40,2	35,3	44,8	41,0
EMB 512	40,8	36,0	46,4	42,7
EMB 1024	40,5	35,8	45,8	42,1
RNN 512	40,8	35,9	45,7	42,0
RNN 1024	40,8	36,0	46,4	42,7
RNN 2048	39,5	34,6	44,8	41,1

Tabela 3: Rezultati metrike BLEU za prevajanje testne množice s sistemom NMT in različnimi hiperparametri učenja oz. hiperparametri modela.

Čeprav lahko za optimalno trajanje učenja pri drugih metrikah opazimo manjša odstopanja, so idealni rezultati še vedno pri ali blizu 110.000 iteracijam učenja.

Na vseh grafih lahko vidimo slabšanje rezultatov pri večjem številu iteracij, kar je posledica prekomernega prilagajanja učni množici. Tako smo za idealni model določili model po 110.000 iteracijah, s katerim smo nato izvajali eksperimente na testni množici. Pripomniti velja, da bo to število iteracij veljalo le v primeru mini-serije z velikostjo 64. Ob večjih oz. manjših velikostih mini-serij se idealno število iteracij sorazmerno pomanjša oz. poveča. Število epoh učenja pa ostaja nespremenjeno.

4.4. Rezultati na testni množici

Rezultati metrik BLEU, METEOR in TER za testno množico so prikazani v tabeli 4.. Prav tako je prikazano izboljšanje rezultatov (pozitivna sprememba BLEU in METEOR oz. negativna sprememba TER) pri vseh metrikah v primerjavi s sistemom SMT.

Pri prevajanju iz slovenščine v angleščino vidimo minimalno poslabšanje rezultata metrike METEOR pred detokenizacijo in denormalizacijo. Vsi ostali rezultati kažejo izboljšanje rezultatov pri prehodu na sistem NMT.

Če kot najbolj uveljavljeno metriko smatramo BLEU, vidimo pomembna izboljšanja v obe smeri prevajanja, in sicer za 6,2 točki v smeri angleščina-slovenščina in 2,9 točke v smeri slovenščina-angleščina. Oba rezultata sta dobljena po detokenizaciji in denormalizaciji.

4.5. Rezultati pri različnih hiperparametrih

V tabeli 4. so prikazani še rezultati, ki jih dobimo z različnimi hiperparametri učenja in modelov. Naš osnovni model je bil učen z dolžinami mini serij 64 (Batch 64), dodali pa še smo modele, naučene z dolžinami serij 16, 32, 128 in 256. V osnovnem modelu smo uporabljali dimenzijo vektorjev za besede 512 (EMB 512), dodali pa še smo modele z dimenzijami 256 in 1024. Zadnji hiperparameter, ki smo ga spreminjali, je dimenzija skrite plasti, ki je bil v osnovnem modelu 1024 (RNN 1024), dodali pa še smo dimenzije 512 in 2048.

Primerjava vseh rezultatov kaže, da je naš osnovni model NMT, ki smo ga zgradili na priporočenih vrednostih hiperparametrov, v vseh primerih tudi v naših eksperimentih najuspešnejši. Pri ostalih modelih opazimo poslabšanja rezultatov do 2 točki BLEU.

4.6. Primerjava z rezultati na testni množici izven domene

Za primerjavo kvalitete prevodov teksta izven domene smo pripravili novo testno množico, ki smo jo dobili iz korpusa IJS-ELAN (Erjavec, 2002). Korpus je prosto dostopen[4] in vsebuje besedila iz različnih virov. Kot besedilo izven domene učnega korpusa smo izbrali leposlovje in sicer roman “1984” G. Orwella. Testno množico smo sestavili iz prvih 1000 segmentov poravnanega korpusa v slovenskem (elan-orwl-sl.xml) in angleškem (elan-orwl-

[4] http://nl.ijs.si/elan/c/

	Angleško→slovensko	Slovenko→angleško
SMT	9,0	10,5
NMT	8,5	10,0
Δ	0,5	0,5

Tabela 4: Rezultati metrike BLEU pri prevajanja v obe smeri za testno množico izven domene učnega korpusa.

en.xml) jeziku. Tako obe množici vsebujeta 16.000 oz. 18.000 besed.

Rezultati prevajanja so prikazani v tabeli 4.6.. Iz rezultatov vidimo, da oba sistema dajeta bistveno slabše rezultate na besedilu, ki ne spada v domeno učnega korpusa. Je pa razlika med obema sistemoma, saj daje prevajalnik SMT v obeh smereh prevajanja rezultate, ki so boljši za 0,5 BLEU točke.

Ta ugotovitev je skladna z ugotovitvami za jezikovni par angleščina-nemščina (Koehn in Knowles, 2017), vendar pa zaradi majhne razlike med našimi rezultati in dejstva, da smo preverili le eno testno množico iz druge domene, lahko zaključimo le, da oba prevajalnika dajeta primerljive rezultate na besedilih izven domene.

4.7. Rezultati glede na dolžine stavkov

Vrednotenje kakovosti prevodov smo ponovili tako, da smo testno množico razdelili na več podmnožic glede na dolžino stavka. Delitev se je izvedla za obe smeri prevajanja ločeno, pri tem pa smo vedno gledali število pojavnic v izvirnem jeziku. Delili smo na množice, ki vsebujejo:

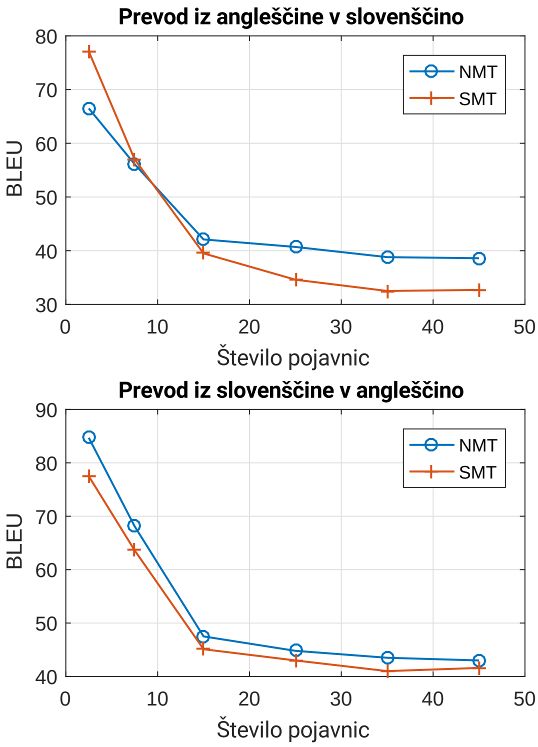
- od 1 do 5 pojavnic,
- od 6 do 10 pojavnic,
- od 11 do 20 pojavnic,
- od 21 do 30 pojavnic,
- od 31 do 40 pojavnic in
- 41 ali več pojavnic.

Rezultati kvalitete prevodov so prikazani na sliki 2. Iz rezultatov vidimo, da oba sistema dajeta boljše rezultate pri krajših stavkih in da celo prevajalnik SMT pri prevajanju iz angleščine v slovenščino daje boljše rezultate v prvih dveh množicah (stavkih do 10 pojavnic). V ostalih primerih pa vidimo, da daje prevajalnik NMT boljše rezultate.

5. Zaključek

Prve raziskave uporabe nevronskih omrežij za strojno prevajanje so pokazale, da lahko z NMT dosežemo boljše prevode, kot pa s klasičnimi statističnimi sistemi.

Prišli smo tudi do zaključka, da je izboljšanje kvalitete prevodov bolj izrazito pri prevajanju iz angleščine v slovenščino kot pri prevajanju v obratni smeri. Ta izsledek je posebej pomemben, saj prevajanje iz morfološko enostavnejših v morfološko kompleksnejše jezike velja kot zahtevnejša smer prevajanja. Zato so izboljšave v tej smeri bolj pomembne, še posebej za slovenski prostor.



Slika 2: Rezultati metrike BLEU glede na število pojavnic v izvirnem jeziku pri prevajanju testne množice v obe smeri s sistemoma SMT in NMT.

Če primerjamo naše ugotovitve glede časa učenja modelov z drugimi raziskavami, vidimo, da dobimo optimalne rezultate pri primerljivem številu epoh. To pomeni, da bo optimalno trajanje učenja odvisno od velikosti učnega korpusa.

V nadaljevanju bomo raziskave usmerili v vključevanje morfoloških informacij v modele prevajanja. Dodatno želimo preučiti uporabo klasičnih jezikovnih modelov za ponovno ocenjevanje hipotez prevajalnika, saj takšnega modela ni v osnovni zasnovi nevronskega omrežja. Preučevali bomo tudi adaptacijo modelov prevajanja pri prehodu v novo domeno besedil.

6. Zahvala

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

7. Literatura

Dzmitry Bahdanau, Kyunghyun Cho in Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo in Marcello Federico. 2018. Neural versus phrase-based MT quality: An in-depth analysis on english-german and english-french. *Computer Speech & Language*, 49:52–70.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau in Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. V: *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*.

Jonathan H. Clark, Chris Dyer, Alon Lavie in Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. V: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, str. 176–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomaž Erjavec. 2002. Compiling and using the ijs-elan parallel corpus. *Informatica*, 26:299–307.

Marcin Junczys-Dowmunt, Tomasz Dwojak in Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. V: *International Workshop on Spoken Language Translation*, IWSLT '16.

Philipp Koehn, Franz Josef Och in Daniel Marcu. 2003. Statistical phrase-based translation. V: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, str. 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn in Rebecca Knowles. 2017. Six challenges for neural machine translation. V: *The First Workshop on Neural Machine Translation*, str. 28–39, Vancouver, Canada, August. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. V: *Conference Proceedings: the tenth Machine Translation Summit*, str. 79–86, Phuket, Thailand. AAMT, AAMT.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, prva izd.

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow in Alexandra Birch. 2017. Deep architectures for neural machine translation. V: *Proceedings of the Second Conference on Machine Translation*, str. 99–107. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry in Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. V: *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, str. 65–68, Valencia, Spain, April. Association for Computational Linguistics.

Mirjam Sepesy Maučec in Gregor Donaj. 2016. Morphosyntactic tags in statistical machine translation of highly inflectional language. V: *Proceedings of the artificial intelligence and natural language conference (AINL FRUCT)*, str. 99–102.

Analiza tvitov slovenskih korporativnih uporabnikov

Darja Fišer*, Monika Kalin Golob**

Oddelek za prevajalstvo Filozofske fakutete Univerze v Ljubljani, Odsek za tehnologije znanja Instituta »Jožef Stefan«
Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si
** Katedra za novinarstvo Fakultete za družbene vede Univerze v Ljubljani
Kardeljeva ploščad 5, 1000 Ljubljana
monika.kalin-golob@fdv.uni-lj.si

Povzetek

V prispevku predstavljamo korpusno analizo korporativnega komuniciranja na družbenem omrežju Twitter, ki smo jo s kombinacijo metapodatkov in besedilnih podatkov izvedli na korpusu Janes-Tviti. Opravili smo analizo računov in dinamiko objav, nato smo analizirali rabo novomedijskih elementov in uporabljenega jezika v korporativnih objavah ter preverili ključne besede korporativnih objav. Analiza korpusa Janes-Tviti je pokazala, da v primerjavi z zasebnimi računi v korporativnih tvitih prevladujejo standardne jezikovne prvine formalnega sporočanja ter da so neformalne in nestandardne izbire redkejše, a premišljene glede na naslovnika in namen sporočanja. Poleg rezultatov opravljene analize je prispevek dragocen zato, ker na sistematičen in metodološko zrel način pokaže potencial korpusnih pristopov v komunikologiji, medijskih š tudijah in drugih sorodnih družboslovnih disciplinah, ki proučujejo jezikovno rabo.

The analysis of tweets of Slovene corporate users

The paper presents a corpus analysis of corporate communication on Twitter which was performed with a combination of metadata and textual data on the Janes-Tweet corpus. We compare the amount, posting dynamics and use of social-media specific communication elements by corporate and private users. Next, we analyse the language of corporate users. Our analysis shows that, in comparison to private accounts, corporate tweets predominantly use formal communication and standard language characteristics with seldom usage of informal and non-standard choices. In the event of those, however, they are chosen deliberately to address a specific target audience and meet the desired communicative goals. In addition to the results of the analysis, a major contribution of the paper is also a systematic and methodologically advanced showcase of the potential of corpus-based approaches in communication studies, media studies and other related disciplines in social sciences which study language use.

1. Uvod

V nedavni raziskavi o besedilnih vrstah na področju odnosov z javnostmi v Sloveniji (Kalin Golob et al., 2018) smo analizirali, katere besedilne vrste so danes najpogosteje uporabljene v slovenski praksi odnosov z javnostmi. K sodelovanju smo povabili 20 strokovnjakov za odnose z javnostmi, po pet iz organizacij iz profitnega, javnega in nevladnega sektorja. Tem predstavnikom smo dodali še predstavnike iz petih slovenskih agencij za odnose z javnostmi, ki med svoje storitve uvrščajo tudi kreativno zasnovo, pripravo in izvedbo različnih pisnih izdelkov za svoje naročnike (npr. sporočila za javnost, letno poročilo, letak, infografika, e-glasilo, revija, besedilo za tvite, facebook, itd.).

Evropski komunikacijski monitor (2017)[1] v analizi pomembnosti komunikacijskih kanalov za praktike odnosov z javnostmi postavlja družbene medije z 62,9 % na šesto mesto. Zato smo analizirali tudi njihove tvite, ki jih v monografijo zaradi premajhnega vzorca in odločitve, da se posvetimo najprej klasičnim žanrom, sicer nismo vključili, a se jim zato posvečamo v pričujočem prispevku.

V tvitih, ki so nam jih posredovali praktiki, prevladujeta dve vrsti jezikovnih izbir:

a) celotno sporočilo je pisano nevtralno, v standardnem (knjižnem) jeziku in krajših povedih. Sledi povezava z ”Več na: ...” Od popolne nezaznamovanosti odstopa le pogostejša raba klicajev in samih velikih črk ob zaključnih vzkličnih povedih (npr. ISKRENA HVALA VSEM DAROVALCEM!; Do obvladovanja stresa vas loči le 6 korakov!; VEDNO obstajajo razlogi za življenje!) in

b) sporočilo je zapisano v govorjenem tonu, značilnem za digitalna sporočila, pojavlja se več neposrednih ogovorov naslovnikov, predvsem z vprašanimi povedmi (Poglejte, kako smo se zabavali na snemanju ...; Ali veste, kod vodi spust ...?; Je tvoj predplačniški račun prazen?; Dobro poznate svoj avtomobil?); ob klicajih se pojavlja še vprašaj (Prvi korak k ukinitvi neutemeljenih stroškov telekomunikacijskih operaterjev?!); frazemov in pogovornih besed (dati na stran, jekleni konjiček, pofočkati se, fajn); medmetov (uuu to je pa fajn!!!!).

Prav tako kot v sporočilih za javnost tudi v tvitih prevladuje pozitivna predstavitev novosti, izdelkov, oseb, institucije ali podjetja. Pozitivni sentiment se ujema z definicijo Guya Cooka (2012), da je jezik odnosov z javnostmi na makroravni določen z namenom “predstaviti posameznika ali organizacijo v prijetni luči”. Iz tega po Cooku izhajajo nejasne ubeseditve in nenatančnost, ubeseditev zgolj pozitivnih elementov. Na mikroravni se po Cookovo to kaže v naslednjih jezikovnih strategijah: nenatančni prislovi (*mnogo ljudi*), naklonski izrazi (*lahko prispeva k*); pomanjkanje natančnosti pri virih (*anketa iz leta 2005*), nenatančni izrazi vrednotenja (*ugodno za naše porabnike in okolje*); izrazi jačanja (*močno poudarjamo osebno verodostojnost*), primerjave in presežniki brez reference (*manj kalorij*), zaokroževanje števnikov (*kakšnih 10 milijonov ljudi*)[2] itn. (Cook, 2007; Cook, Reed in Twiner, 2009 v Cook, 2012). Analiza besedil odnosov z javnostmi

[1] www.communicationmonitor.eu.

[2] Zgledi so iz Kalin Glob et al. (2018).

(Kalin Golob et al., 2018) je na jezikovni ravni kot pozitivne predstavitvene prvine izluščila v rabi pridevnikov in prislovov (*izjemen dosežek*), ki pomensko pozitivno izražajo lastnosti dogodka ali stanja, presežnikov in vljudnostnih fraz (*največji; veseli bomo odziva*), lahko razumljivih metafor, frazemov in drugih tropov s pozitivno konotacijo (*ideja se je rodila*).

Če pregledamo tuja navodila za pisanje v digitalnih medijih, strokovnjaki (Treadwell in Treadwell, 2005; Brown, 2009, Solis in Breakenridge, 2009) in praktična navodila[3] uporabnikom svetujejo, da pišejo čim bolj kratko, jasno in enostavno, se izogibajo žargonskim izrazom, se po vsebini, tonu in stilu prilagajajo naslovnikom. Ker je bistvo digitalnega sporočanja dvosmernost, interaktivnost in hitra povezljivost v druga omrežja, je pomembno pisati na način, ki nagovarja posameznika, ga povezuje in vključuje v dialog. Stil pisanja za družbene medije naj bi bil precej oseben, neformalen in vključujoč.

Bolj kot pravila pisanja avtorji poudarjajo načela, ki jih morajo uporabniki omrežij poznati in v praksi uporabljati. Pisanje mora biti avtentično, relevantno, transparentno, resnicoljubno, odgovorno, vključujoče (Solis in Breakenridge, 2009; Brown, 2009). Organizacije, institucije, podjetja ta kanal uporabljajo pri obveščanju in promociji organizacije, izdelkov, storitev, blagovnih znamk. Twitter omogoča neposredno konverzacijo s potrošniki in pridobivanje neposrednih odzivov, zato je priljubljeno komunikacijsko orodje. Posebej zanimivo pa je, kako se je ta način sporočanja prijel med politiki, predsedniki držav (npr. Obama, Trump, Narendra Modi), vladami (angleška, kanadska), zvezdniki (Katy Perry, Justin Biber, Taylor Swift), blagovnimi znamkami (Playstation, Channel, Samsung Mobile).[4]

Kvalitativen pregled gradiva, ki so nam jih posredovali praktiki, se kaže dober premislek o izbiri stila glede na vsebino (formalnost in standardni jezik obvestil, pogovornost sporočil o družabnih dogodkih, pozivih k udeležbi ali predstavitve novosti), naslovnike (splošna javnost = težnja po nevtralnosti stila; ciljne publike = variacije med nevtralnim in govornim stilom) in sporočevalca (javne institucije = večja nevtralnost in standardnost; podjetja = opaznost, govornost, nestandardne prvine).

Kvantitativna primerjava komuniciranja slovenskih korporativnih in zasebnih uporabnikov na družbenem omrežju Twitter (Ljubešić in Fišer, 2016) je pokazala precejšnje razlike v dinamiki, načinu in vsebini tvitanja teh skupin uporabnikov, ki v veliki meri odražajo njune različne komunikacijske funkcije: korporativni uporabniki največ sporočil objavijo v dopoldanskih urah in med delovniki, zasebni pa v večernih urah in med vikendi. Tviti korporativnih uporabnikov so izrazito pozitivno nastrojeni, medtem ko je sentiment v tvitih zasebnih uporabnikov pretežno nevtralen oz. negativen. Objave korporativnih uporabnikov ostali uporabniki družbenega omrežja veliko pogosteje posredujejo naprej, medtem ko objave zasebnih uporabnikov večkrat všečkajo.

Omenjena raziskava je bila omejena zgolj na kvantitativno analizo metapodatkov, ne pa tudi na analizo dejanske jezikovne produkcije, kar smo izvedli v pričujočem prispevku. Cilj te raziskave je ugotoviti, kakšne so posebnosti korporativnega komuniciranja na družbenih omrežjih v primerjavi s komuniciranjem zasebnih uporabnikov. S korporativnim komuniciranjem imamo v mislih eno od osnovnih elementov odnosov z javnostmi (ang. public relations, nem. Öffentlichkeitsarbeit), za katerega se v literaturi uporabljata tudi strateško komuniciranje in komunikacijski menedžment (prim. Kalin Golob et al., 2018). Pod korporativne uporabnike družbenega omrežja Twitter pa štejemo račune podjetij, institucij, medijev in interesnih združenj, ki ne tvitajo kot posamezniki v zasebne namene. Analiza je bila opravljena na gradivu korpusa Janes-Tviti (Erjavec et al., 2018), v njej pa ves čas kombiniramo metapodatke, dostopne v korpusu, z besedilnimi podatki, kar nam omogoča natančnejše umeščanje, parametrizacijo, primerjavo in posplošitve jezikovne rabe glede na specifične sporazumevalne okoliščine.

2. Korpusna analiza korporativnega komuniciranja na družbenem omrežju Twitter

Korpusno analizo smo opravili na korpusu Janes-Tviti (Erjavec et al. 2018), ki vsebuje 11,3 milijone tvitov oz. 160 milijonov pojavnic, ki jih je objavilo nekaj več kot 10.200 uporabnikov. Ti so glede na namen komuniciranja ročno razvrščeni v dve skupini: zasebni in korporativni uporabniki. V prvem delu smo s pomočjo orodja za korpusno analizo SketchEngine (Killgarriff et al., 2014) analizirali produkcijo in dinamiko objav teh dveh skupin uporabnikov. Nato smo analizirali rabo novomedijskih elementov, kot so heštegi, emodžiji in emotikoni. Sledi analiza jezika v korporativnih objavah ter primerjava ključnih besed.

2.1. Analiza računov

uporabniki	št. uporabnikov (%)	št. pojavnic (%)	št. tvitov (%)
korporativni	2612 (25,57 %)	30.003.182 (18,70%)	2.112.910 (18,64%)
zasebni	7627 (74,44 %)	130.401.083 (81,30%)	9.223.736 (81,36%)
skupaj	10.248 (10,00 %)	160.404.265 (100,00%)	11.336.646 (100,00%)

Tabela 1: Delež korporativnih in zasebnih uporabnikov in njihova produkcija v korpusu Janes-Tviti.

Delež uporabnikov. Kot kaže Tabela 1, je v korpusu razmerje med korporativnimi in zasebnimi 1 : 3. Še večji razkorak med njimi je pri njihovi produkciji, saj so korporativni uporabniki objavili zgolj petino vseh objav v korpusu. Iz tega sklepamo, da je Twitter med slovenskimi uporabniki namenjen pretežno zasebni rabi.

Spol uporabnikov. Glede na oznake spola uporabnikov, ki so bile v korpusu Janes-Tviti pripisane avtomatsko, nato pa ročno pregledane, za veliko večino tvitov korporativnih uporabnikov glede na uporabniško ime, podatke v uporabniškem profilu in glagolske oblike v tvitih spola ni bilo mogoče identificirati (82 %), pri

[3] Npr.: A Scientific Guide to Writing Great Tweets https://blog.bufferapp.com/writing-great-tweets-scientific-guide; 14 Point Checklist: How To Write The Perfect Twitter Update https://getstencil.com/blog/perfect-twitter-update/; How to write a Perfect Tweet http://www.adweek.com/digital/the-perfect-tweet/.

[4] Twitter Statistics Directory – https://www.socialbakers.com/statistics/twitter/. 12. april 2012.

zasebnih uporabnikih je takšnih uporabnikov zelo zelo malo (1,5 %). Podatki, predstavljeni v tabeli 2, so pričakovani, saj v korporativnem komuniciranju uporabniki tvitajo v imenu podjetja oz. organizacije, čemur prilagajajo stil pisanja, kot je npr. raba množinskih glagolskih oblik.

	korporativni		zasebni	
spol	št. tvitov	%	št. tvitov	%
nedoločljiv	1.730.258	81,89%	134.048	1,45%
moški	271.729	12,86%	6.136.470	66,53%
ženski	110.923	5,25%	2.953.218	32,02%
skupaj	2.112.910	100,00%	9.223.736	100,00%

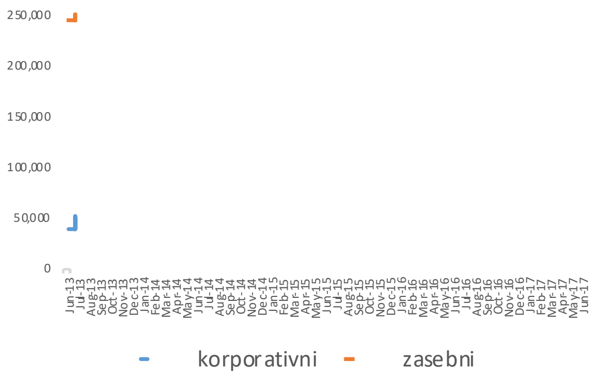
Tabela 2: Distribucija tvitov korporativnih in zasebnih uporabnikov glede na spol v korpusu Janes-Tviti.

2.2. Analiza objav

Količina objav. Korporativnih uporabnikov, ki so na družbenem omrežju zelo dejavni in objavijo več kot 10.000 tvitov, je le 29 (1,11 %). Srednje dejavnih uporabnikov, ki objavijo med 1.000 in 10.000 tvitov, je 422 (16,16 %). Večina (1640 oz. 62,79 %) korporativnih uporabnikov sodi v kategorijo malo aktivnih računov, s katerih je objavljenih med 100 in 1.000 tvitov. Med najmanj aktivne uporabnike, ki objavijo manj kot 100 tvitov, pa sodi 521 (19,95 %) računov. V primerjavi z zasebnimi uporabniki je največ razlik v 2. in 4. skupini. Med zasebnimi uporabniki je namreč 9 % več takšnih uporabnikov, ki objavijo med 1.000 in 10.000 tvitov, ter podoben delež manj računov, ki imajo objavljenih med 1.000 in 100 tvitov. V letih, ki so zajeti v korpus Janes-Tviti, je delež vsebin korporativnih uporabnikov stabilen, medtem ko pri zasebnih uporabnikih nekoliko upada, kar prikazuje Slika 1. Občasni veliki padci v količini objav, ki so simultani pri obeh skupinah uporabnikov, niso povezani s sezonskim nihanjem ali kakšnim drugim vsebinskim pojavom, temveč s tehničnimi težavami pri pridobivanju gradiva.

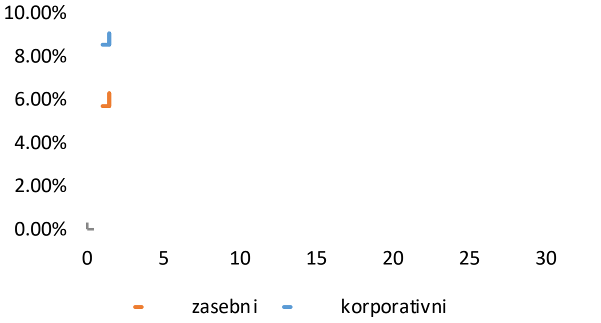
	korporativni		zasebni	
št. vseh računov	2612	%	7627	%
> 10.000 tvitov	29	1,11%	129	1,69%
med 10.000 in 1.000 tviti	422	16,16%	1867	24,48%
med 1.000 in 100 tviti	1640	62,79%	4055	53,17%
< 100 tvitov	521	19,95%	1576	20,66%

Tabela 3: Dejavnost korporativnih in zasebnih uporabnikov v korpusu Janes-Tviti.

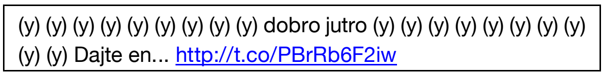


Slika 1: Dinamika objavljanja zasebnih in korporativnih uporabnikov v korpusu Janes-Tviti glede na št. objavljenih tvitov med junijem 2013 in junijem 2017.

Dolžina objav. Kot je razvidno iz Slike 2, je dolžina tvitov korporativnih uporabnikov bolj homogena od zasebnih uporabnikov, ki objavljajo tako več krajših kot več daljših tvitov od korporativnih uporabnikov, največji delež tvitov pri korporativnih uporabnikih ima med 7 in 11 besed (zasebni med 4 in 7 besede), kar je povezano z manjšo dejansko dvosmernostjo tvitov v službi odnosov z javnostmi. To se kaže tudi v tem, da je tvitov, ki ne vsebujejo nobene besede (samo emodži, hešteg, hiperpovezavo ali multimedijske vsebine), le 0,1 % (pri zasebnih uporabnikih je takšnih tvitov šestkrat več, saj se s temi znaki pač odgovarja dvosmerno). Najdaljši slovenski korporativni tvit v korpusu je prikazan na sliki 3.



Slika 2: Dolžina tvitov zasebnih in korporativnih uporabnikov v korpusu Janes-Tviti.


(y) (y) (y) (y) (y) (y) (y) (y) (y) dobro jutro (y) (y) (y) (y) (y) (y) (y) (y) (y) (y) Dajte en... http://t.co/PBrRb6F2iw

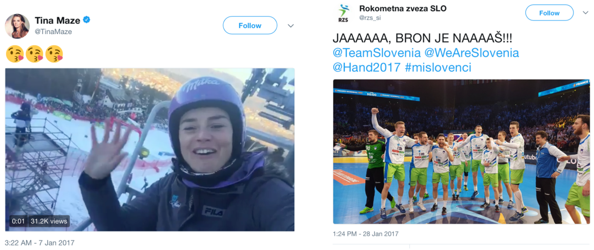
Slika 3: Najdaljši korporativni slovenski tvit v korpusu Janes-Tviti.

2.3. Analiza interaktivnih elementov

Všečkanje objav. Da je dvosmernost manjša in so korporativni tviti le eden od kanalov enakega (enosmernega) sporočanja znotraj različnih žanrov, kaže tudi podatek, da štiri petine korporativnih tvitov ne prejme nobenega všečka, po enega jih ima 12 %, 2 ali več všečka pa le 9 % tvitov. Pri zasebnih uporabnikih opazimo precejšnje razlike, saj vsaj 1 všeček prejme tretjina vseh tvitov, nezanemarljiv delež (0,7 %) jih prejme celo več kot 10 všečkov.

	št. všečkov			
	korporativni uporabniki		zasebni uporabniki	
	št. tvitov	%	št. tvitov	%
0	1.663.755	78,74%	610.9048	66,23%
1	265.385	12,56%	1.890.549	20,50%
2-10	175.788	8,32%	1.160.057	12,58%
>10	7.982	0,38%	64.082	0,69%
skupaj	2.112.910	100,00%	9.223.736	100,00%
	št. retvitov			
	korporativni uporabniki		zasebni uporabniki	
0	1.754.988	83,06%	8.414.713	91,23%
1	219.698	10,40%	490.346	5,32%
2-10	134.184	6,35%	300.319	3,26%
>10	4.040	0,19%	18.358	0,19%
skupaj	2.112.910	100,00%	9.223.736	100,00%

Tabela 4: Delež všečkanih in posredovanih tvitov korporativnih in zasebnih uporabnikov v korpusu Janes-Tviti.



Tina Maze @TinaMaze

😘😘😘

0:01 31.2K views

3:22 AM - 7 Jan 2017

Rokometna zveza SLO @rzs_si

JAAAAAA, BRON JE NAAAAŠ!!! @TeamSlovenia @WeAreSlovenia @Hand2017 #mislovenci

1:24 PM - 28 Jan 2017

Sliki 4 in 5: Najbolj všečkan (levo) in najbolj retvitan korporativni slovenski tvit v korpusu Janes-Tviti.

Raba heštegov			
	abs. frekv.	na milijon	na tvit
korporativni	922.504	30.746,9	0,44
zasebni	2.241.693	17.190,8	0,24
Raba emodžijev			
	abs. frekv.	na milijon	na tvit
korporativni	1.285.696	42.852,0	0,61
zasebni	12.061.885	92.498,3	1,31
Raba hiperpovezav			
	abs. frekv.	na milijon	na tvit
korporativni	1.989.643	66.314,4	0,94
zasebni	2.583.651	19.813,1	0,28
Raba omemb			
	abs. frekv.	na milijon	na tvit
korporativni	659.211	21.971,4	0,31
zasebni	9.216.857	57.460,2	1,00

Tabela 5: Raba heštegov, emodžijev in hiperpovezav pri korporativnih in zasebnih uporabnikih.

Posredovanje objav. Pri posredovanju tvitov je situacija nekoliko obrnjena, saj je v tem primeru vsaj enkrat posredovan bistveno večji delež korporativnih (17 %) kot zasebnih tvitov (8 %), pri zelo pogosto posredovanih tvitih pa se tipi računov izenačijo.

Raba heštegov. Relativno gledano, korporativni računi skoraj dvakrat pogosteje uporabljajo heštege kot zasebni, v povprečju skoraj vsak drugi korporativni tvit vsebuje hešteg (pri zasebnih računih šele vsak četrti). Kot je razvidno iz Tabele 5 med 10 najpogostejšimi heštegi korporativnih uporabnikov močno prevladujejo športne teme, kar je zelo podobno zasebnim uporabnikom. Prav tako je na seznamu najpogostejših 10 uporabljenih heštegov pri obeh tipih uporabnikov zelo veliko prekrivanja, saj se polovica heštegov pojavi na obeh seznamih (šport, novice, Ljubljana). Pri obeh so na vrhu seznama skoraj izključno heštegi, vezani na športno tematiko. Med 10 korporativnimi uporabniki, ki relativno gledano uporabijo največ heštegov, so manj formalne revije in podjetja. Za natančnejšo analizo korporativnega komuniciranja bi bilo zato zanimivo korporativne uporabnike nadrobneje razvrstiti med medije (dnevnike in revije), podjetja, državne institucije ter nevladne organizacije, kar načrtujemo za nadaljnje raziskave.

Raba emotikonov in emodžijev[5]. Pri emotikonih in emodžijih je situacija obratna kot pri heštegih, saj so relativno gledano emodžiji več kot dvakrat pogostejši pri zasebnih uporabnikih. Pri zasebnih uporabnikih namreč posamezni tvit v povprečju vsebuje 1,3 emodžije oz. emotikone, medtem ko je ta element pri korporativnih uporabljen le v vsakem drugem tvitu, kar nakazuje večjo formalnost korporativnega sporočanja. Med 10 korporativnimi računi, ki relativno gledano uporabljajo največ emodžijev in emotikonov, so večinoma prodajalci modnih artiklov.

Kot kaže Tabela 6, so vsi najpogostejši emodžiji oz. emotikoni pozitivni, kar ponovno kaže pozitivno usmerjenost sporočanja v odnosih z javnostmi. Zanimivo je, da v tvitih korporativnih računov med najpogostejšimi 10 tovrstnimi elementi najdemo kar 8 emotikonov in le 2 emodžija, kar lahko nakazuje na večjo konzervativnost komuniciranja korporativnih uporabnikov, saj so emodžiji veliko mlajši fenomen kot emotikoni, po drugi strani pa lahko kažejo na to, da korporativni uporabniki več tvitajo z računalnikov, saj so emodžiji značilni predvsem za komuniciranje prek pametnih telefonov.

korporativni uporabniki		zasebni uporabniki	
hešteg	frekvenca	hešteg	frekvenca
#plts	18.703	**#plts**	26.370
#slonews	18.247	**#slonews**	18.270
#PLTS	9.620	**#junaki**	18.167
#Ljubljana	5.724	#slochi	13.195
#izvršba	5.167	**#PLTS**	10.943
#NKDomzale	4.437	#Slovenia	10.780
#olimpija	4.176	**#Ljubljana**	10.141
#rokomet	4.143	#radiobattleSI	9.184
#junaki	3.941	#ligaprvakov	9.091
#skupajdovrha	3.864	#sp14si	8.351

Tabela 6: Deset najpogostejših heštegov v tvitih korporativnih in zasebnih uporabnikov.

emo.	frekv.	uporabnik	frekv.	rel. frekv.[6]
:)	114.602	_RecycleMan	530	12.711,5
;)	55.763	JennParisBags	188	11.522,1
:D	17.715	EtiVelikonja	160	10.409,8
&lt;3	13.688	ApartmaNet	184	10.104,9
:-)	9.672	TRENDtrgovina	436	10.049,3
;-)	4.926	Pawla40	228	9.720,0
:))	4.680	iPlace_si	125	8.860,0
❤️	3.679	bozicluka	92	8.290,2
:P	3.558	matejgaber22	99	7.222,6
😉	3.436	Modniovitki	424	7.010,9

Tabela 7: Deset najpogostejših emotikonov in emodžijev v tvitih korporativnih uporabnikov in seznam 10 korporativnih računov z najvišjo relativno frekvenco emotikonov in emodžijev.

Raba hiperpovezav. Pri navajanju povezav na druge spletne strani opazimo zelo velike razlike med zasebnimi in korporativnimi uporabniki. Relativno gledano, korporativni uporabniki uporabljajo več kot trikrat več hiperpovezav kot zasebni. V povprečju korporativni uporabniki v skoraj vsakem tvitu dodajo kakšno hiperpovezavo, zasebni pa le v vsakem četrtem; to se sklada z ugotovitvami preliminarne analize, da so tviti pogosto le skrčena sporočila za javnost, ki daljše sporočilo in dodatne informacije prinašajo na povezavah.

[5] Emotikoni (npr. ;)) so kombinacije standardnih tipografskih znamenj, s katerimi izražamo čustva. Emodžiji pa so piktogrami (npr. 🎂), ki poleg čustev obsegajo še širok nabor drugih tematik, njihova raba in interpretacija pa se od posameznika do posameznika razlikujeta.

[6] Relativna frekvenca je povprečna frekvenca fenomena v milijonu pojavnic.

Omembe drugih uporabnikov. Zelo velike razlike med zasebnimi in korporativnimi uporabniki opazimo tudi pri omenjanju drugih uporabniških računov v tvitih. Relativno gledano, zasebni računi druge uporabnike v svojih tvitih omenjajo več kot dvakrat pogosteje. V povprečju zasebni uporabniki v vsakem tvitu navedejo drugega uporabnika, korporativni pa to storijo le v vsakem tretjem. Tviti v odnosih z javnostmi so usmerjeni k samopredstavitvi, zato je sklicevanja na druge po pričakovanjih manj. Med desetimi najpogosteje omenjanimi računi so pri korporativnih uporabnikih to večinoma mediji, politične institucije/stranke/posamezni politiki in športne organizacije, pri zasebnih računih pa t. i. vplivneži na drugih omrežji, dva novinarja, en politik. Na obeh seznamih se znajdeta samo dva ista računa, in sicer YouTube ter Janez Janša.

korporativni uporabniki		zasebni uporabniki	
omemba	pogostost	omemba	pogostost
@YouTube	8.325	@petrasovdat	91.328
@Nova24TV	6.903	**@YouTube**	71.859
@Val202	3.992	@MarkoSket	57.333
@rtvslo	3.866	**@JJansaSDS**	53.482
@kzs_si	3.736	@lucijausaj	51.391
@union_olimpija	3.616	@leaathenatabako	44.453
@JJansaSDS	3.464	@petra_jansa	44.102
@radioPrvi	3.128	@savicdomen	43.394
@vladaRS	2.764	@darkob	42.363
@nkmaribor	2.758	@zzTurk	40.534

Tabela 8: Deset najpogosteje omenjenih računov v tvitih korporativnih in zasebnih uporabnikov.

2.4. Analiza jezika

Jezik sporočil. Korporativni uporabniki objavljajo sporočila skoraj izključno v slovenščini (93 %), s čimer se precej razlikujejo od zasebnih uporabnikov, pri katerih najdemo dvakrat večji delež tujejezičnih tvitov. Med tujimi jeziki v tvitih korporativnih uporabnikov močno prednjači angleščina (5 %), precej manj pa objavljajo v ostalih jezikih (1,6 %). To se ujema z našimi preliminarnimi ugotovitvami, saj gre v večini primerov za nagovarjanje slovenskega naslovnika v uradnem komuniciranju o zadevah, ki jih želi institucija prenesti v poslovne ali inoformativne namene. Izjema so računi slovenskih veleposlaništev, ki veliko tvitov objavljajo v lokalnem jeziku (npr. v francoščini), ter zunanjega ministrstva, predsednika države in vlade, ki s tviti v angleščini občasno obveščajo tudi mednarodno javnost o pomembnejših dogodkih (npr. o arbitraži).

	korporativni		zasebni	
jezik	št. tvitov	%	št. tvitov	%
slv	1.973.677	93,41%	8.074.681	87,54%
eng	104.955	4,97%	983.141	10,66%
hbs	16.058	0,76%	57.017	0,62%
ostalo	18.220	0,86%	108.897	1,18%
skupaj	2.112.910	100,00%	9.223.736	100,00%

Tabela 9: Raba jezikov v tvitih korporativnih in zasebnih uporabnikov.

Sentiment sporočil. Vsakemu tvitu v korpusu je pripisana oznaka sentimenta (glej Erjavec et al., 2018). Polovica vseh tvitov, ki jih korporativni uporabniki objavijo, ima pozitivni sentiment, tretjina je nevtralnih, negativnih pa 17 %. To je zelo drugače kot pri zasebnih uporabnikih, ki objavijo polovico nevtralnih tvitov, 27 % negativnih in le četrtino pozitivnih.

	korporativni		zasebni	
sentiment	št. tvitov	%	št. tvitov	%
pozitiven	1.024.238	48,48%	2.320.841	25,16%
nevtralen	729.811	34,54%	4.411.516	47,83%
negativen	358.861	16,98%	2.491.379	27,01%
skupaj	2.112.910	100,00%	9.223.736	100,00%

Tabela 10: Sentiment tvitov korporativnih in zasebnih uporabnikov.

Standardnost jezika. Korporativni uporabniki v svojih tvitih večinoma uporabljajo standardno slovenščino (80 %), zelo nestandardnih vsebin je zelo malo (3 %), s čimer se močno razlikujejo od zasebnih uporabnikov, ki v standardni slovenščini objavljajo le slabo polovico tvitov, delež tvitov, napisanih v zelo nestandardni slovenščini, pa je pri zasebnih uporabnikih več kot štirikrat večji kot pri korporativnih. Izjema so računi nekaterih javnih osebnosti (npr. stand-up komiki, radijski voditelji, glasbeniki), ki jim je neformalno komuniciranje pomemben del korporativnega imidža in zato pogosto zavestno tvitajo v nestandardni slovenščini. Tovrstni računi so označevalcem tudi povzročali največ težav pri ročnem razvrščanju med zasebne in korporativne račune.

	korporativni		zasebni	
standardnost	št. tvitov	%	št. tvitov	%
L1	1.688.244	79,90%	4.515.310	48,95%
L2	353.397	16,73%	3.489.743	37,83%
L3	71.269	3,37%	1.218.683	13,21%
	2.112.910	100,00%	9.223.736	100,00%

Tabela 11: Stopnja standardnosti tvitov korporativnih in zasebnih uporabnikov.

besedna vrsta	korporativni (na milijon)	zasebni (na milijon)	razmerje [7]
lastni sam.	66.738,40	33.507,80	1,99
števniki	30.564,90	16.109,70	1,90
vezniki	54.381,10	33.302,10	1,63
predlogi	86.947,20	54.549,60	1,59
pridevniki	76.889,90	48.254,80	1,59
občni sam.	186.446,60	127.056,00	1,47
okrajšave	3.826,00	3.458,90	1,11
ločila	143.234,60	158.188,20	0,91
polnopom. gl.	62.631,90	75.795,70	0,83
pom. gl.	36.974,70	52.968,00	0,70
prislovi	38.192,10	55.483,10	0,69
zaimki	39.118,20	62.678,80	0,62
členki	19.816,60	35.540,70	0,56
medmeti	1.740,90	6.194,50	0,28

Tabela 12: Primerjava jezika korporativnih in zasebnih uporabnikov glede na besedne vrste.

Pravopisno gledano, se pojavljajo velike razlike pri krajšavah: v korporativnih tvitih prevladujejo standardne okrajšave akademskih in drugih nazivov (dr., mag., d. o. o.) in dogovorjene okrajšave (št., oz., min.), pri zasebnih pa nestandardne (tw), v njih tudi pogosto manjka krajšavna pika (slo, lj, min). Razlike se pojavljaj tudi pri rabi ločil. Korporativni računi uporabljajo večji nabor klasičnih ločil in jih stavijo bolj standaradno, torej pravopisno normirano.

[7] Razmerje med frekvenco v tvitih korporativnih in zasebnih uporabnikov.

Za zasebne uporabnike je značilno pogosto ponavljanje istega ločila za doseganje želenega čustvenega naboja sporočila, izrazito pogostejša je tudi raba tipičnih novomedijskih simbolov (#, @, *).

Besedne vrste. Skozi opazovanje rabe besednih vrst v jeziku korporativnih tvitov dobimo uvid v komunikacijske namene korporativnih računov. Relativno gledano, je pri korporativnih računih v primerjavi z zasebnimi uporabljenih skoraj dvakrat več lastnih samostalnikov in števnikov, izrazito pogostejši so tudi vezniki, predlogi, pridevniki in občni samostalniki. Kot je razvidno iz tabele 10, je pri zasebnih računih izrazito pogostejša raba medmetov (3,5-krat več), členkov (skoraj 2-krat več) ter zaimkov in prislovov, kar po eni strani potrjuje večjo formalnost korporativnih ter neposrednost in osebnost zasebnih uporabnikov, po drugi pa odraža precej različni komunikacijski funkciji družbenega omrežja Twitter pri teh skupinah uporabnikov, ki je izrazito obvestilna za korporativne račune in konverzacijska za zasebne. Obvestilna, deloma pa tudi vplivanjska funkcija korporativnih računov, se zrcalita tudi iz podrobnejših analiz posameznih besednih vrst, ki jih predstavljamo v nadaljevanju.

Samostalniška beseda. V korporativnih tvitih je sicer raba občnih imen enainpolkrat pogostejša, a je med prvih dvajset najpogostejših občnih samostalnikih presenetljivo veliko ujemanja, kar 70 %: dan, leto, tekma, ura, mesto, teden, čas, hvala, svet, delo, človek, konec, otrok, država. Samostalniki, ki se pojavijo med 20 najpogostejšimi v korporativnih tvitih, v zasebnih pa ne, pa so: video, foto, zmaga, novica, cena in sezona. Osebna lastna imena so v korporativnih tvitih dvakrat pogostejša, prekrivanje pri dvajsetih najpogostejših je 40 %: Slovenija, Ljubljana, Maribor, EU, Slovenc, Evropa, ZDA, Cerar, Janša. Lastni samostalniki, ki se pojavijo med 20 najpogostejšimi v korporativnih tvitih, v zasebnih pa me, pa so: Olimpija, Koper, Peter, Gorica, Janez, Domžale, Luka, Tina, Marko. Formalnost izražanja je večja v korporativnih, saj navajajo imena in priimke (zasebni le priimek), prav tako je v korporativnih tvitih večja raznolikost krajev in imen podjetij. Samostalniški zaimki kažejo pričakovane razlike: korporativni tviti vsebujejo množinske (nam, nas, vam), zasebni pa edninske oblike (jaz, me, ti, te). Množinskost gre pripisati uradnemu sporočanju v imenu institucije oz. podjetja in vikanju naslovnika.

Glagol. Raba polnopomenskih glagolov je pogostejša v zasebnih tvitih, najpogostejših dvajset se v 60 % ujema (imeti, iti, morati, vedeti, videti, priti, dobiti, začeti, čakati, dati, praviti, delati, dobiti) vendar so razlike v motivaciji za sporočanje: korporativni računi poročajo o dogodkih in izjavah, zasebni o lastnih aktivnostih in mnenju. Polnopomenski glagoli, ki se pojavijo med 20 najpogostejšimi v korporativnih tvitih, v zasebnih pa ne, so: želeti, preveriti, najti, iskati, prebrati, gledati, moči, hoteti, narediti.

Pridevniška beseda. Raba pridevnikov je enainpolkrat pogostejša v korporativnih tvitih, prekrivanje med 20 najpogostejših je 50 %: nov, dober, slovenski, velik, lep, zadnji, mlad, star, pravi, super. Pridevniki, ki se v korporativnih tvitih pojavijo med 20 najpogostejšimi, v zasebnih pa ne, so: vabljen, današnji, evropski, javen, spleten, svetoven, odličen, državen, visok, domač. V korporativnih tvitih torej prevladujejo izrazito pozitivni (nov, dober, slovenski, velik, lep), ki so tudi bolj formalni od zasebnih (vabljen, odličen, visok vs. hud, mali, sam). Pridevniški zaimki se prav tako kot samostalniški v korporativnih tvitih uporabljajo v prvoosebni množinski obliki (naše, naši), ko gre za istovetenje s podjetjem oz. institucijo in vključevanje v omejeni sporočanjski krog, ki povezuje tvorca sporočila v imenu institucije in naslovnika (Korošec 1998).

Členki. Razliko med formalnostjo in neformalnostjo kažejo tudi členki, ki se sicer v 80 % prekrivajo, vendar med neprekrivnimi v korporativnem sporočanju izrazito izstopajo bolj formalni (morda, predvsem, sicer, skoraj), v zasebnem pa nestandardni in neformalni (tud < tudi; ze < že, itak, pač).

Medmeti. Kot že omenjeno, so pri tej besedni vrsti razlike največje. Zabeležili smo 55 % prekrivanje 20 najpogostejših medmetov v korporativnih in zasebnih tvitih: bravo, hm, haha, uf, o, ej, ah, ha, aha, aja, oh. Neprekrivni med njimi so: živjo, zdravo, hej, hehe, gooool, opa, ups, na, ojoj. Ob manjši količini medmetov so tisti v korporativnih tvitih tudi bolj formalni in pozdravljalni (zdravo, ups), medtem ko so v zasebnem sporočanju pogosti tudi tujejezični (btw, lol) in kletvice (fak, wtf).

2.5. Analiza ključnih besed

V tem razdelku analiziramo ključne besede v korporativnih tvitih, pri čemer ključne besede razumemo kot tiste besede, ki so v tem korpusu nenavadno pogoste v primerjavi z referenčnim korpusom. Kot referenčni korpus pri nas služi kar celoten korpus Janes-Tviti.

negativen	ključnost	pozitiven	ključnost	nevtralen	ključnost
oviran	22,2	čestitka	3,5	novice.si	10,1
trčenje	19,1	vabljen	3,5	zemljišče	8,7
trčiti	18,0	bravo	3,4	pivniški	8,3
priključek	15,4	album	3,4	ebel	8,3
evakuirati	15,3	beautiful	3,4	katarinin	8,1
ranjen	15,1	hvala	3,4	petv	8,0
poškodovan	15,0	posted	3,4	šloganje	7,9
razcep	14,9	photos	3,4	solaten	7,8
novicejutro.si	14,9	odličen	3,3	ugnati	7,8
osumljen	14,6	polepšati	3,3	pripravljalen	7,7
nesreča	14,5	odlično	3,3	koel	7,6
aretirati	14,3	prijeten	3,3	novinec	7,6
avtocesta	14,1	super	3,3	napovednik	7,4
neurje	14,1	čudovit	3,3	zoofa	7,3
strmoglaviti	13,9	čestitati	3,3	prerokovanje	7,3
osumljenec	13,1	srečno	3,3	poiesis	7,2
magnituda	13,1	facebook	3,3	apod	7,1
prometen	12,8	welcome	3,3	wt	7,1
ubit	12,8	summer	3,3	sklepen	6,9

Tabela 13: Seznam 20 najbolj ključnih lem v korporativnih tvitih glede na sentiment.

Sentiment. Kot je razvidno iz Tabele 13, po ključnosti izrazito izstopa besedišče v negativnih korporativnih tvitih, med katerimi vseh 20 najbolj ključnih lem prihaja iz medijskih tvitov in se navezuje na črno kroniko, poročanje o nesrečah (npr. trčenje, evakuirati, ranjen, nesreča). 20 najpogostejših ključnih besed s pozitivnim sentimentom ustreza definicijam pozitivnega poročanja v odnosih z javnostmi (npr. čestitka, vabljen, bravo, čudovit, polepšati), na visokem mestu pogostosti so tudi pridevniki in prislovi, ki izražajo visoko stopnjo pozitivnega (beautiful, odličen/odlično, prijeten, super, čudovit). Tudi dvajset najpogostejših ključnih besed z nevtralnim sentimentom prihaja iz poročevalskih medijskih tvitov (novice.si, zemljišče, napovednik, sklepen), ubeseduje dogodke (pivniški, ebel, šloganje, prerokovanje) in imena (katarinin, ebel, zoofa, apod). Ta seznam torej kaže na to, da bi bilo pri analizi korporativnega komuniciranja na družbenem omrežju Twitter smiselno premisliti o ločevanju medijskih tvitov ter tvitov podjetij in institucij.

standardni tviti	ključnost	nestandardni tviti	ključnost
Izkl	6,4	Posetite	562,3
Novice.SI	6,4	potrazi	557,6
dražba	6,0	sjajan	553,5
[hiperpovezava]	5,9	Jeste	455,0
SiOL	5,8	tim	308,5
Petv	5,8	[hiperpovezava]	307,2
APOD	5,8	[hiperpovezava]	186,6
Moia	5,7	li	166,4
spletnem	5,7	koketo	145,9
Zurnal24	5,7	trombeto	143,3
ugodne	5,7	[hiperpovezava]	130,0
astronomska	5,7	belooranžnega	129,5
SMUČANJE	5,6	deejaytime	111,2
KOŠARKA	5,6	Živjo	111,0
oviran	5,6	Skupne	109,6
[hiperpovezava]	5,6	pritisne	92,8
ALPSKO	5,6	oglasiš	66,2
HOKEJ	5,6	[hiperpovezava]	65,9
zamudite	5,6	cheers	60,3
Preverite	5,5	hajskul	56,5
Nogometaši	5,5	[hiperpovezava]	49,6
TENIS	5,5	gnargnar	49,6
ciganskih	5,4	sporočimo	47,0
NOGOMET	5,4	najbrš	46,8
ROKOMET	5,4	pridte	45,3
[hiperpovezava]	5,4	javimo	41,9
Astrolife.si	5,4	Poslali	41,5
Izbrane	5,4	dm	41,2
Slovenske	5,4	javiš	41,2
SMUČARSKI	5,4	unc	41,0

Tabela 14: Primerjava ključnih besednih oblik v korporativnih tvitih, napisanih v standardni in nest. jeziku.

ženske	ključnost	moški	ključnost
foodwalks	7,7	Moia	41,7
Posodobljen	7,0	dražba	39,9
Patsy	6,1	APOD	37,2
KOEL	5,9	astronomska	36,4
[hiperpovezava]	5,9	premičnin	35,4
info@patsy.si	5,5	UGANKA	33,9
[hiperpovezava]	5,5	[hiperpovezava]	30,7
foodwalk	5,5	Izhodišče	30,3
Lylo	5,3	FOTOGRAFIJE	30,0
ORTO	5,1	GLASBA	29,6
UriKuri	4,6	Dopolni	29,5
yummy	4,6	UE	29,1
Ordered	4,4	javna	27,5
Shellac	4,4	sedežna	27,2
Cosmo	4,2	GCC	26,5
LPG	3,8	PRIPOROČAMO	26,4
Starševski	3,7	Espargaro	26,4
e-trgovine	3,5	[hiperpovezava]	26,3
[hiperpovezava]	3,5	zemljišča	26,0
Elle	3,3	[hiperpovezava]	25,3
info@tjasaseme.si	3,3	Pomurskem	24,8
boxa	3,2	ENERGIJE	24,5
derivatov	3,2	Žurnal24	24,4
IBU	3,1	LITERATURA	24,3
Onaplus	3,1	gozda	24,2
Aquafresh	3,0	[hiperpovezava]	23,5
naftnih	3,0	PRS	23,1
Watercolour	3,0	Ekipa24	22,8
[hiperpovezava]	3,0	[hiperpovezava]	22,3
foodwalks	7,7	Moia	41,7

Tabela 15: Primerjava ključnih besednih oblik v korporativnih tvitih glede na spol avtorja.

Standardnost. Primerjava 30 najbolj ključnih besednih oblik (glej Tabelo 14) v korporativnih tvitih, ki so bili napisani v standardni in v nestandardni slovenščini, kaže, da uporabniki standardno slovenščino uporabljajo za objavo obvestil in oglasov (npr. dražba, ugodne, zamudite, preverite). V korporativnih tvitih, napisanih v nestandardni slovenščini, je namen sporočanja zelo podoben, vendar veliko tujejezičnih prvin in nestandardnega zapisa slovenskih besed, kaže, da se v tovrstnih sporočilih uporabniki želijo približati svoji ciljni publiki in jim tako svojo ponudbo narediti privlačnejšo (npr. deejaytime, hajskul, najbrš, pridte, dm, javiš).

Spol. Primerjava korporativnih ženskih in moških računov sicer ne daje podatkov o morebitnih tipičnih razlikah v rabi jezika med moškimi in ženskami, je pa povedna z vidika tematskih in stilnih razlik pri jezikovnih izbirah za nagovarjanje ženske ali moške ciljne publike: pri ženskih računih gre za imena revij, spletne naslove in lastna imena, ki se nanašajo na modo, nakupovanje, hrano in starševstvo, pri moških pa na nepremičnine, šport in glasbo.

3. Zaključek

Jezikovnostilna in žanrska analiza pridobljenega materiala (Kalin Golob et al., 2018) je pokazala, da praktiki za odnose z javnostmi uporabljajo tvite v jezikovni obliki, kot so jo uporabili v drugih žanrih (npr. sporočilu za javnost), torej enako besedilo objavijo po različnih kanalih. Tako so tviti večinoma videni le kot možnost doseganja različnih javnosti, zato jezikovne izbire drugačnemu kanalu prilagajajo le redko.

Iz predstavljanja organizacije oz. institucije, ki želi informirati v pozitivni luči, izhaja značilnost, ki jo omenjajo raziskovalci najpogosteje v zvezi s sporočili za javnost, to je raba promocijskih, torej predstavitvenih elementov (Cameron in Marcus, 2002; Maat, 2007), s katerimi predstavijo informacije v pozitivni luči za podjetje ali institucijo. Predstavnik za odnose z javnostmi poskuša v sporočilu za javnost združevati zahtevo novinarske stroke, da promocijski elementi ne smejo nastopati v informativnih novinarskih besedilih, hkrati pa mora z objavo doseči pozitivno predstavitev dogodka ali stanja. Kot ugotavlja Maat (2007, 60), po eni strani veljajo priporočila, naj se pisci sporočil za javnost izogibajo pretirani rabi pridevnikov, posredujejo samo gola dejstva in ne uporabljajo vrednotenjskih sredstev, npr. presežnikov. Hkrati pa se zavedajo, da pozitivne trditve zvišujejo branost, zato se jim novinarji eksplicitno ne izogibajo.

Pregled ključnih besed v korpusu Janes-Tviti glede na sentiment prikazuje zanimivo sliko tvitov, označenih kot korporativni: negativne zajemajo medijsko poročanje o nesrečah, pozitivne pa v celoti ustrezajo definiciji promocijskih elementov. Prav zato, ker se kažejo večje razlike v negativnem sentimentu novičarskih tvitov in pozitivnem drugih podjetij, institucij ipd., bi bilo treba za nadaljnje raziskave jezika in stila odnosov z javnostmi medijske portale obravnavati ločeno, saj gre pri njih za značilno novinarsko poročanje v maniri: slaba novica je dobra novica, medtem ko podjetja, institucije in posamezniki za samopromocijo uporabljajo tipične značilnosti sporočil za javnost.

Analiza korpusa Janes-Tviti je v veliki meri potrdila rezultate delne analize gradiva tvitov, pridobljene za analizo žanrov v odnosih z javnostmi (Kalin Golob et al., 2018), da korporativni tviti prevladujoče vsebujejo standardne jezikovne prvine formalnega sporočanja ter da so neformalne in nestandardne izbire redkejše, vendarle pa premišljene glede na naslovnika, torej ciljno občinstvo. Prave dvogovornosti ni, saj gre za enosmerno sporočanje podjetja, institucije, ki tvite večinoma uporablja kot skrajšana sporočila za javnost, za katera je bil izbran le nov kanal, ki ga omogoča nova tehnologija, medtem ko je prevladujoča funkcija obvestilna in pozitivno predstavitvena. Vse to se kaže v manjšem številu emotikonov glede na zasebne tvite in vseh drugih prvin, ki kažejo na neposrednost stika v zasebnih tvitih, v katerih prevladuje konverzacijska funkcija: nestandardni in številni členki, številni medmeti, nestandardna raba ločil in njihov omejen izbor, število všečkov in nestandardnega jezika.

Rezultati analize, predstavljene v pričujočem prispevku, pa ne potrjujejo preteklih ugotovitev izogibanju rabe pridevnikov in vrednotenjskih sredstev, saj smo z analizo ključnosti glede na besedno vrsto v podkorpusu korporativnih tvitov zaznali ravno izrazito visok delež vrednotenjskih pridevnikov, med katerimi so pogosti superlativi.

Poleg rezultatov opravljene analize je prispevek dragocen zato, ker na sistematičen in metodološko zrel način pokaže potencial korpusnih pristopov v komunikologiji, medijskih študijah in drugih sorodnih družboslovnih disciplinah, ki proučujejo jezikovno rabo, kar v slovenskem okolju zaenkrat še ni uveljavljena praksa.

V nadaljevanju želimo raziskave korporativnega komuniciranja nadgraditi in poglobiti z ločevanjem različnih tipov korporativnih uporabnikov, kot so medijske hiše, korporacije in javne ustanove. Prav tako bomo podrobneje proučili recepcijo korporativnih tvitov, ki vsebujejo nestandardne jezikovne prvine in interaktivne novomedijske elemente, ki so sicer značilnejši za zasebne uporabnike.

Zahvala

Raziskava, opisana v prispevku, je bila opravljena v okviru nacionalnega temeljnega projekta “ Viri, metode in orodja za razumevanje, prepoznavanje in razvrščanje različnih oblik družbeno nesprejemljivega diskurza v informacijski družbi ” (J7-8280, 2017-2019), ki ga je financirala ARRS.

4. Literatura

Rob Brown. 2009. *Public Relations and the Social Web. How to use social media and web 2.0 in communicazions*. London, Philadelphia, Kogan Page.

Deborah Cameron in Thomas A. Marcus. 2002. The words between the space: Buildings and Language. London, Routledge.

Darja Fišer, Tomaž Erjavec in Nikola Ljubešić. JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. Slovenščina 2.0, 4 (2): 67–99.

Guy Cook. 2007. 'This we have done'. The differet vagueness of poetry and Publiv relations. V Cutting, Joan (ur.): *Vague Language Explored*. London, Palgrave, 21–39.

Guy Cook. 2012. British applied linguistics: imacts of and imacts on. *Applied Linguistics Rewiew, 3–1*: 25–45.

Nikola Ljubešić in Darja Fišer. Slovene Twitter Analytics. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana, Slovenia: 39–43.

Monika Kalin Golob, Nada Serajnik Sraka in Dejan Verčič. 2018 (v tisku). *Pisanje za odnose z javnostmi: temeljni žanri*. Zbirka Stičišča. Ljubljana: Fakulteta za družbene vede.

Kilgarriff, Adam, 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)*: 7-36.

Pander Maat. 2007. How promotional language in press releases is dealt with by journalists: Genre mixing or genre conflict? *Journal of Business Communication 44.1*: 59–95.

Brian Solis in D. K. Breakenridge. 2009. Putting the Public *Back in Public Relations: How Social Media Is Reinventing the Aging Buisness of PR.* Pearson Education LTD.

Donald Treadwell in Jill B. Treadwell. 2005. *Public Relations Writing: Principles in Practice.* Sage Publications Inc.,Thousand Oaks, California. 2nd Ed.

Citiranje jezikoslovnih podatkov v slovenskih znanstvenih objavah: stanje in priporočila

Darja Fišer[*†], Jakob Lenardič[*], Tomaž Erjavec[†]
[*] Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si
jakob.lenardic@ff.uni-lj.si

[†] Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

Odprta znanost temelji na prosto in odprto dostopnih znanstvenih publikacijah in podatkih. Slednji omogočajo preverjanje rezultatov predhodnih raziskav in njihovo nadgrajevanje, v kontekstu jezikovnih tehnologij in ročno označenih jezikovnih virov pa tudi šolanje novih orodij za procesiranje besedil. Vendar pa je, tako kot za znanstvene objave, tudi za podatke pomembno, da so korektno citirani, saj šele to omogoča ponovljivost raziskav, citati pa so tudi najpomembnejši pokazatelj zanimivosti in koristnosti delovanja znanstvenikov in pomembno vplivajo na njihovo možnost pridobivanja projektov in zaposlitev. V prispevku obravnavamo stanje citiranja jezikoslovnih podatkov, predvsem korpusov, v slovenskih znanstvenih publikacijah. Izvedli smo pregled večjega števila slovenskih revij in zbornikov in kvantitativno ter kvalitativno analizirali rezultate. Izsledke povzamemo in po ti. »austinskih načelih«, pokažemo, kaj je bilo že narejenega v sklopu raziskovalne infrastrukture CLARIN.SI ter predlagamo smernice za citiranje znanstvenih podatkov in načine za njihovo implementacijo.

Linguistic data citation in Slovene scientific publications: analysis and recommendations

Open science is based on freely and openly available scientific publications and data. The latter enable the verification and improvement of previous research. In the context of language technologies and manually annotated language resources, they also enable training of new text processing tools. However, just like scientific publications, research data need to be properly cited, as only this makes reproducibility of experiments possible and is the most important indicator of how interesting and useful researchers' work is in the community and plays a major role in their success with research grant proposals and career trajectory. In this paper, we survey the landscape of linguistic data (corpora) citation in Slovene scientific publications. The investigation was performed on key Slovene linguistic journals and proceedings with the results analysed both quantitatively and qualitatively. Our findings are organized according to the Austin Principles of data citation, where we present the developments in this field within the CLARIN.SI research infrastructure and propose recommendations for linguistic data citation as well as suggest solutions for their implementation.

1. Uvod

Odprti dostop do znanstvenih publikacij in podatkov pospešuje inovacije, spodbuja sodelovanje, zmanjšuje podvajanje dela in omogoča dograjevanje predhodnih rezultatov raziskav ter vključevanje državljanov in družbe (European Commission, 2012). Odprti dostop do rezultatov raziskav predvidevajo *Resolucija o nacionalnem programu za jezikovno politiko 2014–2018*,[1] *Nacionalna strategija odprtega dostopa do znanstvenih objav in raziskovalnih podatkov v Sloveniji 2015-2020*[2] ter *Akcijski načrt izvedbe nacionalne strategije odprtega dostopa do znanstvenih objav in raziskovalnih podatkov v Sloveniji 2015—2020*.[3]

V Sloveniji imamo na področju jezikovnih virov že dolgo tradicijo odprtih podatkov. Že od nastanka so bili odprto dostopni npr. jezikovni viri projektov MULTEXT-East[4], JOS[5] in SSJ,[6] leta 2013 pa je bila ustanovljena raziskovalna infrastruktura za jezikovne vire in orodja CLARIN.SI, v sklopu katere je bil vzpostavljen certificiran repozitorij, ki arhivira prek sto odprto dostopnih jezikovnih virov.

Med poglavitnimi cilji *Akcijskega načrta* je izvedba pilotnega programa *Odprti dostop do raziskovalnih podatkov v letih 2017—2020*, katerega namen je izboljšati dostop do raziskovalnih podatkov, mdr. z uvedbo novega sistema za vrednotenje raziskovalnih podatkov, v skladu s katerim bodo raziskovalni podatki, shranjeni v pooblaščenem podatkovnem središču, ki so prestali presojo pomena za znanost priznani kot znanstvena objava. Dobra praksa doslednega citiranja raziskovalnih podatkov je pomembna, ker zagotavlja in spodbuja transparentnost znanstvenega dela in posledično deluje kot ključni vzvod tovrstnega sistema vrednotenja. Raziskovalni podatki so v najboljšem primeru shranjeni v certificiranih repozitorijih (npr. repozitorij infrastrukture CLARIN.SI),[7] kar je skladno z *Akcijskim načrtom,* saj repozitoriji zagotavljajo tako trajni in transparentni dostop kot tudi jasno dokumentacijo za določen vir.

V pričujočem prispevku nas ne zanimajo jezikovni viri sami ali njihova dostopnost, niti ne njihova uporaba v raziskovalni skupnosti, temveč kako se le-ta citira v znanstvenih člankih slovenskih publikacij. Kot smo zapisali pred desetimi leti:

[1] http://www.pisrs.si/Pis.web/pregledPredpisa?id=RESO91#

[2] http://www.mizs.gov.si/delovna_podrocja/direktorat_za_znanost/sektor_za_znanost/strategije_s_podrocja_znanosti/nacionalna_strategija_odprtega_dostopa_do_znanstvenih_objav_in_raziskovalnih_podatkov_v_sloveniji_2015_2020/

[3] http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/Znanost/doc/Odprti_dostop/Akcijski_nacrt_-_POTRJENA_VERZIJA.pdf

[4] http://nl.ijs.si/ME/

[5] http://nl.ijs.si/jos/

[6] http://www.slovenscina.eu/

[7] http://www.clarin.si/

»*Citiranje je še posebej pomembno, ker je merljiv kazalec raziskovalne uspešnosti, zato bi se tudi moralo dosledno izvajati. Žal pa to ni v navadi pri citiranju publikacij o jezikovnih virih: vse prepogosto se nek vir omeni samo po imenu ali pa se v najboljšem primeru doda njegov spletni naslov, namesto da bi se v virih citiralo publikacijo, kjer je vir prvotno opisan*« (Erjavec, 2009).

Od takrat se je stanje spremenilo, tako da je sedaj mogoče citirati ne samo publikacije o izdelavi nekega vira, pač pa tudi vir sam, saj npr. repozitorij CLARIN.SI za vsak vnesen vir na samem vrhu njegove spletne strani točno navaja, kako naj se ga citira. Dostop do podatkov v certificiranih repozitorijih, kot je CLARIN.SI, je v skladu s t.i. Austinskimi načeli za ustrezno citiranje v jezikoslovju, ki so povzeti v dokumentu *The FORCE11 Joint Declaration of Data Citation Principles*.[8] Poleg tega, da so natančna navodila za citiranje jezikovnih virov skladna z drugo točko Austinskih načel (*Credit and Attribution*, »Priznanje zaslug in avtorstva«), so transparentni metapodatki in stalni spletni identifikatorji, ki jih repozitoriji nudijo za vsak vir, ključnega pomena za zagotavljanje odprtega dostopa in s tem interoperabilnost, trajnost in preverljivost podatkov.

Neposredni povod za pričujoč prispevek je bilo zasedanje Interesne skupine za jezikoslovne podatke, ki je potekalo v sklopu plenarnega sestanka »Research Data Alliance« v Berlinu 22. 3. 2018. Na sestanku je bilo odprtih več vprašanj o citiranju raziskovalnih podatkov v jezikoslovju, kar nam je zbudilo zanimanje, kakšno je stanje na tem področju v Sloveniji. Prispevek ima sledečo strukturo: v 2. razdelku podamo pregled mednarodnih načel in praks pri citiranju znanstvenih podatkov v jezikoslovju, 3. razdelek analizira stanje v izbranih slovenskih publikacijah, 4. razdelek predlaga smernice za boljšo prakso na tem področju, zadnji razdelek pa zaključi in poda smernice za nadaljnje delo.

2. Mednarodna načela citiranja podatkov v jezikoslovju

Odprta znanost, odprti podatki in citiranje le-teh je v svetu trenutno v središču pozornosti, saj so obstoječe prakse tudi mednarodno zastarele, manj v naravoslovju in posebej računalništvu, mnogo bolj pa v humanistiki in jezikoslovju; tako npr. relativno nova »Splošna pravila za oblikovanje jezikoslovnih prispevkov« (Haspelmath, 2014) citiranja podatkov sploh ne omenjajo.

Obširen pregled pomena odprte znanosti, odprtih podatkov in potrebe po korektnem citiranju v jezikoslovju je podan v Berez-Kroeker et al. (2018), ki je rezultat iniciative, v kateri je sodelovalo 41 jezikoslovcev in drugih znanstvenikov. Prispevek najprej osmisli odprte raziskovalne podatke in ponovljivost raziskav, tako na splošno kot v jezikoslovju, nato pa poda pregled trenutnega stanja v jezikoslovju, kar se tiče transparentnosti uporabljenih virov in raziskovalnih metodologij. Avtorji ugotavljajo, da je po eni strani nemogoče uveljaviti ponovljivost raziskav brez primernega citiranja virov, po drugi pa, da je stanje v jezikoslovju še vedno zelo nezadovoljivo. Nato sledijo ugotovitve avtorjev glede potrebe po mehanizmih, ki bi ovrednotila tudi »delo na podatkih« pri zaposlovanju in napredovanju znanstvenikov, in nujnosti po korenitem premiku v omogočanju ponovljivosti raziskav v jezikoslovju, kar naj bi dosegli skozi izobraževanje, promocijo in razvoj ustreznih politik. Strinjajo se, da bi zbiralci podatkov za svoje delo morali dobiti primerno priznanje avtorstva, posebej takrat, ko so izdelani podatki dostopni, ponovno uporabni in jih je mogoče citirati. Prispevek zaključijo priporočila za konkretne dejavnosti, ki bi jih morali izvesti jezikoslovci, oddelki, sveti in založniki. Te dejavnosti so v veliki meri osredotočene na zagotovitev odprtih podatkov oz. izobraževanje, kako se upravlja s podatki, da sploh lahko postanejo odprti, kot tudi, kako primerno ovrednotiti to delo. Zadnje priporočilo pa je neposredno posvečeno boljšemu citiranju raziskovalnih podatkov, kjer avtorji svetujejo urednikom ter založnikom znanstvenih revij in knjig uvedbo konkretnih politik tako za izmenjavo podatkov kot za njihovo citiranje, pri slednjem tako, da razvijejo formate za citiranje jezikoslovnih podatkov.

3. Analiza citiranja objav v slovenskih znanstvenih publikacijah

3.1. Izbor gradiva in zasnova analize

Za pričujoči prispevek smo pregledali ključne slovenske revije in zbornike za področje jezikoslovja in ugotavljali, v kolikšni meri in na kakšen način avtorji prispevkov omenjajo oz. navajajo jezikovne vire. Naj poudarimo, da nas v tej raziskavi ni zanimalo, kateri jezikovni viri so v objavljenih raziskavah uporabljeni in citirani, temveč, kako jih avtorji navajajo.

Pri revijah smo analizirali navodila za avtorje in izdane številke za zadnjih pet let (2013-2017), pri zbornikih pa navodila za avtorje oz. predloge prispevkov ter celoten zbornik zadnje edicije konference. Med zborniki smo v študijo zajeli *JTDH 2016* in *Obzorja 2016*, med revijami pa: *Linguistica*, *Jezik in slovstvo*, *Jezikoslovni zapiski*, *Slavistična revija*, *Slovene Linguistic Studies* in *Slovenščina 2.0*

Skupaj je bilo pregledanih 751 znanstvenih prispevkov, od katerih jih vire omenja 133 oz. dobrih 17 %. Navedbe virov v pregledanih prispevkih ločujemo na naslednje kategorije:

- **Povezava[9] na vir v besedilu prispevka (največkrat v opombi)**. Zgled takega citiranja je v Žele (2014), kjer je povezava na korpus *Gigafida* podana v opombi. Prispevki v tovrstni kategoriji ne navajajo ključne publikacije o viru, t.j. Logar et al. (2012).
- **Povezava na vir v bibliografiji**. Zgled takega citiranja je v Ribič (2016), kjer je povezava na korpus *Gigafida* podana v končnem seznamu virov. Prispevki v tovrstni kategoriji ne navajajo ključne publikacije o viru.
- **Povezava na vir v besedilu prispevka (največkrat v opombi) kot tudi v bibliografiji**. Zgled takega citiranja je v Žele (2015), kjer je povezava na korpus *Gigafida* podana večkrat v opombah ter v končnem seznamu virov. Prispevki v tovrstni kategoriji ne navajajo ključne publikacije o viru.
- **Publikacija o viru**. Zgled takega citiranja je v Verdonik in Sepesy Maučec (2013), kjer je za korpus *OPUS OpenSubtitles* navedena ključna publikacija o viru, t.j. Tiedemann (2009).

[8] https://www.force11.org/datacitationprinciples

[9] V to kategorijo vključujemo tudi navedbe stalnih spletnih identifikatorjev, kot so handle in DOI.

- **Povezava na vir v besedilu prispevka in publikacija o viru**. Zgled takega citiranja je v Bálint Čeh in Kosem (2017). Avtorja podajata povezavo na korpus *Gigafida* v opombi in navajata ključno publikacijo, t.j Logar et al. (2012).
- **Kombinacija različnih načinov navajanja virov**. Zgled takega navajanja je v Ljubešić et al. (2013), kjer je za označevalnik JOS navedena ključna publikacija (Erjavec et al., 2010), za korpus ssj500k pa zgolj povezava med besedilom.[10]
- **Brez navedbe vira**. Zgled takega citiranja je v Vidovič Muha (2015). Avtorica se sklicuje na uporabo označevalnika JOS, vendar ne podaja niti povezave na vir niti ne navaja njegove ključne publikacije (t.j. Erjavec et al., 2010).

3.2. Pregled navodil avtorjem

V tem razdelku podajamo kratek pregled navodil za avtorje, saj je od teh navodil močno odvisno, kako bodo avtorji navajali vire. Za revije *Jezikoslovni zapiski*, in *Slovene Linguistic Studies* ter za zbornik *Obdobja* navodil avtorjem na njihovih spletnih straneh nismo našli.

Najbolj podrobna navodila za navajanje virov podaja revija *Slovenščina 2.0,*[11] ki ločuje navajanje korpusov, spletnih strani in spletnih virov:

Korpus:

- Gigafida. Dostopno prek: http://www.gigafida.net (datum dostopa).
- Cambridge English Corpus.e Dostopno prek: http://www.cambridge.org/gb/elt/catalogue/subject/item2701617/Cambridge-International-Corpus/?site_locale=en_GB (datum dostopa).

Spletna stran:

- OpenWebSpider. Dostopno prek: http://www.openwebspider.org/ (datum dostopa).
- Creative Commons. Dostopno prek: http://creativecommons.org/ (datum dostopa).

Spletni vir:

- Pew Research Center (2010): Americans Spending More Time Following the News ? Ideological News Sources: Who Watches and Why. Dostopno prek: http://www.people-press.org/ (datum dostopa).
- TEI Consortium, ur. (2011): TEI P5: Guidelines for Electronic Text Encoding and Interchange: Version 1.9.1. Dostopno prek: http://www.tei-c.org/Guidelines/P5/ (datum dostopa).
- Scott, M. (2008): WordSmith Tools: Version 5. Dostopno prek: http://www.lexically.net/downloads/version5/HTML/index.html (datum dostopa).

Jezik in slovstvo avtorje poziva,[12] da vire in literaturo navajajo ločeno, kar se nam zdi dobra praksa, saj s tem avtorjem med drugim sporočajo, da je uporaba in navajanje virov pomemben sestavni del znanstvenega prispevka. Dodatno velja dodatno omeniti, da poziv k ločenemu navajanju jezikovnih virov omogoča bralcem lažji dostop in preveritev citiranih podatkov, ki podpirajo neko znanstveno trditev, kar je skladno z npr. austinskimi načeli (glej razdelek 4) . Podrobneje ta revija načina za navajanje jezikovnih virov sicer ne definira, iz primera za navajanje spletnih strani pa lahko sklepamo, da jezikovne vire v elektronski obliki enači s spletnimi stranmi, saj kot primer navajanja spletnih strani navaja korpus *FidaPLUS*:

- Korpus slovenskega jezika FidaPLUS: <http://www.fidaplus.net>. (Dostop dan. mesec. leto.)

Na podoben način jezikovne vire obravnava revija *Linguistica*[13]:

- Le dictionnaire de la zone. 20 May 2010. http://www.dictionnairedelazone.fr/.

Slavistična revija[14] v navodilih za oblikovanje seznama literature uvaja zelo neeksaktno navajanje spletnih virov, brez navedbe spletnih povezav, verzij oz. datuma dostopa:

- Lemma (Lexikographie). Wikipedia: Die freie Enzyklopädie.
- Primož JAKOPIN, 1980: Zgornja meja entropije pri leposlovnih besedilih v slovenskem jeziku: Doktorska disertacija. Ljubljana. Na spletu.

Pri prvem primeru ni jasno, na katero različico se referenca nanaša, saj je Wikipedija kolaborativen projekt, kjer uredniki gesla lahko ves čas spreminjajo, bi bilo nujno treba dodati datum dostopa. Pri drugem primeru pa ni jasno, ali gre za referenco na doktorsko disertacijo kot publikacijo al za jezikovni vir, ki je bil v okviru disertacije razvit. Prav tako referenca ne vsebuje spletne povezave, zato bralec do vira ne more dostopati. Tovrstna praksa ne spodbuja preverljivosti in ponovljivosti raziskav ter priznavanja zaslug avtorjem virov, zato bi jo bilo pomembno čim prej izboljšati, še posebej, ker gre za jezikoslovno revijo, ki se v sistemu vrednotenja znanstvenih objav uvršča v sam vrh.

Revija *Slovene Linguistic Studies* posebej za navajanje elektronskih virov ne podaja navodil.

Podobno zbornik *JTDH*[15] v predlogi prispevkov sicer vsebuje primer dodajanja hiperpovezav v opombe in navaja načine navajanja različnih tipov enot bibliografije, a med njimi ni primerov za citiranje jezikovnih virov. Glede na to, da gre za vodilno konferenco za področje jezikovnih virov in tehnologij, bi konferenca nujno morala posvečati več pozornosti ozaveščanju in usmerjanju avtorjev prispevkov za ustrezno citiranje jezikovnih virov.

3.3. Kvantitativna analiza

Glede na podatke v Tabeli 1 vsebuje 17.7 % vseh pregledanih objav (vsaj eno) navedbo jezikovnega vira, načini navajanja pa so zelo raznoliki in razpršeni. Izrazito prevladuje navajanje povezave na vir v bibliografiji, česar se poslužuje četrtina vseh prispevkov, v katerih so bili viri uporabljeni. Dvakrat redkejša je praksa navajanja ključne publikacije o uporabljenem viru, ki je v trenutno veljavnem sistemu, ki seveda ni popoln in ni (primarni) cilj znanstvenega udejstvovanja, je pa kljub vsemu zelo pomemben za pridobivanje zaposlitev in projektov, za vrednotenje znanstvene uspešnosti edini način citiranja, ki avtorjem vira prinaša točke. Precej pogosto je kombiniranje več različnih načinov navajanja virov v istem prispevku (19 %), kar kaže na neupoštevanje navodil avtorjem oz. na pomanjkljiva navodila.

V Tabeli 2 navajamo rezultate analize za posamezne revije, ki smo jih vključili v raziskavo. Najvišji delež prispevkov, ki omenjajo jezikovne vire, vsebuje revija *Slovenščina 2.0* (97 %), najnižjega pa revija *Linguistica* (4 %), kar posredno tudi odraža programsko usmeritev revij. Po nenavajanju uporabljenih virov izrazito izstopa *Slavistična revija*, v kateri pri več kot treh četrtinah (78 %) prispevkov, ki rabo virov omenjajo, teh virov nikjer ne

[10] http://www.slovenscina.eu/tehnologije/ucni-korpus

[11] http://slovenscina2.0.trojina.si/si/oddaja-prispevkov/

[12] https://www.jezikinslovstvo.com/02.php

[13] https://revije.ff.uni-lj.si/linguistica/about/submissions#authorGuidelines

[14] https://srl.si/navodila_guidelines.pdf

[15] http://www.sdjt.si/wp/dogodki/konference/jtdh-2018/#navodila

citirajo. Glede na to, da gre za vodilno jezikoslovno revijo v našem prostoru, ki je uvrščena tudi na seznam ARRS revij posebnega pomena, bi še posebej to uredništvo revije moralo skrbeti za visok nivo raziskovalne kulture v slovenskem jezikoslovju in ustrezno citiranje raziskovalnih podatkov od avtorjev izrecno zahtevati v navodilih za avtorje.

Vseh objav	751	100,0 %
Objave z omembo vira	133	17,7 %
Hiperpovezava na vir v besedilu prispevka	13	9,7 %
Hiperpovezava na vir v bibliografiji	33	24,8 %
Hiperpovezava na vir v besedilu in v bibliografiji	14	10,5 %
Publikacija o viru	16	12,0 %
Hiperpovezava na vir v besedilu prispevka in publikacija o viru	8	6,0 %
Kombinirano	25	18,8 %
Brez	25	18,8 %

Tabela 1: Pregled distribucije različnih načinov navajanja virov v analiziranih publikacijah.

Najbolj homogeno navajanje virov je v *Slavistični reviji*, kjer smo identificirali le dva različna načina (povezava na vir v besedilu ali v bibliografiji), najbolj heterogeno pa v *Jezikoslovnih zapiskih*, kjer najdemo vse načine navajanja virov, razen kombiniranega. Najvišji delež navedbe vira v obliki hiperpovezave na spletno stran vira najdemo v reviji *Linguistica* (67 %), najvišji delež citiranja ključnega prispevka o viru pa pripada reviji *Slovenščina 2.0* (18 %).

Od posameznih načinov navajanja virov je navajanje povezav na vir v besedilu prispevka (največkrat v opombah) najpogostejši način navajanja virov v vseh revijah, razen v reviji *Slovenščina 2.0*, kjer je nekoliko pogostejše citiranje ključne publikacije o viru. Tega načina se sicer v manjšem številu prispevkov poslužujejo samo še v revijah *Jezik in slovstvo* in *Jezikoslovni zapiski*.

V Tabeli 3 navajamo rezultate za konferenci, ki smo ju vključili v raziskavo. V zborniku *JTDH 2016* so viri omenjeni v 93 % vključenih prispevkov, kar je glede na področje konference razumljivo. V tem zborniku naletimo na izrazito velik delež prispevkov (46 %), v katerih avtorji uporabljajo različne kombinacije navajanja virov. To je verjetno odraz heterogene raziskovalne skupnosti, ki se predstavlja na tej konferenci, in pomanjkljivih navodil avtorjem ter manj rigoroznega uredniškega in tehničnega pregleda končnih različic oddanih prispevkov.

V zborniku *Obdobja*, ki je bil posvečen Jožetu Toporišiču, je tovrstnih prispevkov 19 %. Glede na to, da je bila ta edicija simpozija tematsko vezana na jezikovni opis slovenščine, se zdi ta rezultat nizek. Vendar je po drugi strani občutno višji kot v programsko sorodnih revijah, predstavljenih v Tabeli 2, kar morda nakazuje spremembe sestave oz. praks tudi v tej skupnosti, saj so revije tradicionalno konzervativnejše in spremembe, do katerih v raziskovalni skupnosti prihaja, absorbirajo nekoliko kasneje od konferenc.

3.4. Kvalitativna analiza

V tem razdelku navajamo zanimivejše pojave, na katere smo naleteli pri kvalitativnem pregledu gradiva. Najprej predstavljamo nekatere primere dobrih praks, nato pa analiziramo identificirane problematične primere navajanja virov. Kot zgleden primer citiranja virov navajamo Logar et al. (2014) v reviji *Slovenščina 2.0*, ki za isti vir navaja tako ključno publikacijo o viru v bibliografiji kot tudi povezavo na vir v besedilu prispevka v sprotnih opombah, ki so prikazane na dnu relevantne strani prispevka. Na ta način bralcu omogočimo, da neposredno dostopa tako do vira kot tudi do publikacije o njem, prav tako pa avtorjem vira ustrezno priznamo zasluge in avtorstvo ter zagotovimo citiranost.

Naslednji zgleden primer citiranje virov, ki prav tako prihaja iz revije *Slovenščina 2.0*, je Arhar Holdt in Dobrovoljc (2016), ki v bibliografiji za vir navede stalni spletni identifikator handle v repozitoriju CLARIN.SI:

- Krek, S., Erjavec, T., Dobrovoljc, K., Može, S., Ledinek, N. in Holz, N. (2015): Training corpus ssj500k 1.4. Dostopno prek: http://hdl.handle.net/11356/1052.

Navajanje handlov je pomembno, ker bralcu zagotavlja, da bo lahko dostopal do vira, četudi se sam naslov spletne strani spremeni. Prav tako pa handle bralcu omogoča dostop do podrobnejšega opisa jezikovnega vira, ki je bil uporabljen v raziskavi, do njegovih metapodatkov, za prosto dostopne vire pa tudi do vira samega. S tem je močno izboljšana preverljivost in ponovljivost raziskav, spodbuja pa tudi nadaljnje razširitve in izboljšave raziskav ter maksimizira izrabo jezikovnega vira, izdelava katerega je zahtevala finančni in časovni vložek.

Pri pregledu smo naleteli tudi na problematične načine citiranja, ki jih uvrščamo v naslednje kategorije:

- Nekonsistentno navajanje istega vira: V *Slavistični reviji* je isti vir navajan zelo različno. Npr. v Meterc (2013) in Jakop (2014):
  - Gigafida, korpus slovenskega jezika. Ur. Filozofska fakulteta Univerze v Ljubljani. Ljubljana: FF. Splet.
  - Korpus GigaFida. Na spletu.
- Nekonsistentno navajanje različnih virov istega tipa: V reviji *Slovene Linguistic Studies* je v Štumberger (2015) za *Sloleks* navedena hiperpovezava v opombi, nemška leksikalna vira pa sta vključena v bibliografijo:

	Jezik in slovstvo		*Slavistična revija*		*Jezikoslovni zapiski*		*SLS*		*Slo 2.0*		*Linguistica*	
Vse objave	157	100%	180	100%	115	100%	26	100%	45	100%	134	100%
Objave z omembo vira	11	7%	14	8%	20	17%	8	31%	34	76%	6	4%
Povezava na vir v besedilu prispevka	0	0%	1	7%	3	15%	2	25%	5	15%	0	0%
Povezava na vir v bibliografiji	4	36%	2	14%	7	35%	2	25%	5	15%	4	67%
Povezava na vir v besedilu in bibliografiji	2	18%	0	0%	2	10%	3	38%	3	9%	1	17%
Publikacija o viru	2	18%	0	0%	1	5%	0	0%	6	18%	0	0%
Povezava na vir in publikacija o viru	0	0%	0	0%	1	5%	0	0%	6	18%	0	0%
Kombinirano	0	0%	0	0%	0	0%	1	13%	8	24%	0	0%
Brez	3	27%	11	79%	6	30%	0	0%	1	3%	1	17%

Tabela 2: Pregled praks navajanja jezikovnih virov v ključnih slovenskih znanstvenih revijah za področje jezikoslovja za obdobje 2013-2017. SLS je okrajšava za revijo *Slovene Linguistic Studies*, Slo 2.0 pa za revijo *Slovenščina 2.0*.

	Zbornik JTDH		***Zbornik Obdobja***	
Vse objave	30	100%	64	100%
Objave z omembo vira	28	93%	12	19%
Povezava na vir v besedilu prispevka	2	7%	0	0%
Povezava na vir v bibliografiji	2	7%	7	58%
Povezava na vir v besedilu in bibliografiji	3	11%	0	0%
Publikacija o viru	6	21%	1	8%
Povezava na vir in publikacija o viru	1	4%	0	0%
Kombinirano	13	46%	2	17%
Brez	1	4%	2	17%

Tabela 3: Pregled praks navajanja jezikovnih virov v ključnih konferenčnih zbornikih za področje jezikoslovja za leto 2016.

- OWID, Online-Wortschatz-Informationssystem Deutsch des Instituts für deutsche Sprache, Mannheim, (13. 3. 2008)
- Klappenbach, Ruth, Steinitz, Wolfgang (ur.). 1967 (1964). Wörterbuch der deutschen Gegenwartssprache. 1. Band. Berlin: Akademie-Verlag. http://www.dwds.de/ (1. 7. 2008, 27. 3. 2015).

• Neustrezne hiperpovezave: V reviji *Jezik in slovstvo* smo opazili neustrezno navajanje povezav na vire. Npr. v Polajnar (2013) ni hiperpovezave na osnovno stran vira, ampak na podstran:

- Gigafida: http://www.Gigafida.net/Support/About

V *Slavistični reviji* smo opazili neustrezno navajanje povezav na vire. Npr. v Fabčič (2014) je korpus *FidaPLUS* v bibliografiji naveden brez povezave:

- FidaPLUS – Korpus slovenskega jezika. Na spletu.

Ker je korpuse mogoče naložiti na različne konkordančnike, kar lahko privede tudi do razlik v rezultatih, je za zagotavljanje preverljivosti in ponovljivosti raziskav v referenci nujno potrebno navesti natančno povezavo, ki je bila v raziskavi uporabljena.[16]

V reviji *Slovenščina 2.0* smo opazili nenatančo navajanje hiperpovezave do korpusa. Npr. v Arias-Badia et al. (2014), kjer je za španski korpus navedena generična povezava na konkordančnih SketchEngine:

- SWC = Spanish Web Corpus. Available at: www.sketchengine.co.uk (20 October 2014).

Tovrstno navajanje referenc na korpuse je med jezikoslovci precej razširjeno, je pa problematično iz več razlogov. Ne samo, da ne priznava avtorstva korpusa, temveč resno zavira preverljivost in ponovljivost raziskav, saj iz reference sploh ni razvidno, za katero različico korpusa konkretno gre, saj po eni strani obstaja več spletnih korpusov španščine, ki so jih ustvarili različni avtorji, po drugi pa so bili številni med njimi bili izdelani v več različicah in vsebujejo različno gradivo. Ko smo korpus želeli preveriti v konkordančniku SketchEngine, na katerega nas raferenca napoti, ga nismo našli, saj konkordančnik na dan preverjanja[17] ponuja dva španska korpusa tega tipa: Spanish Web Corpus oz. SpanishWaC (Sharoff 2006) in Spanish Web 2011 oz. esTenTen11 (Kilgarriff in Renau 2013). Tu je potrebno poudariti, da odgovornost za ustrezno navajanje virov ne leži samo na strani avtorjev prispevkov, temveč tudi avtorjev virov, ki bi vsem uporabnikom prvi morali zagotoviti ustrezno spremno dokumentacijo o korpusu, vključno z navodili za citiranje, tako ključnega prispevka o viru kot tudi navajanje korpusa v konkordančniku in korpusa kot podatkovno zbirko. Veliko razvijalcev virov tega še vedno ne omogoča, zato je ozaveščanje nujno potrebno tudi pri tej ciljni skupini.

4. Diskusija

Kot je pokazala analiza, je trenutno stanje na področju navajanja virov v slovenskem jezikoslovju vse prej kot idealno, saj so navodila avtorjem za področje elektronskih jezikovnih virov zelo raznolika, ponekod zastarela, pri precejšnjem številu revij in zbornikov pa celo manjkajo. Posledično so tudi prakse navajanja virov tako med kot tudi znotraj posameznih znanstvenih publikacij zalo heterogene. Še bolj pa je zaskrbljujoč podatek, da skoraj petina objavljenih prispevkov v uglednih znanstvenih revijah in zbornikih uporabljenih virov sploh ne navaja.

Da bi skušali prispevati k izboljšanju stanja, v nadaljevanju prispevka oblikujemo priporočila, ki temeljijo na mednarodnih iniciativah in predlogih, kako izboljšati citiranost raziskovalnih podatkov. Konkretno sledimo osmim načelom »austinskih principov« citiranja podatkov v jezikoslovju (Berez-Kroeker et al., 2017). Za vsako od načel podamo ime in prevod definicije, nakar ga umestimo v Slovenijo z analizo stanja in predlogi za ukrepe, kako jih realizirati.

4.1. Pomembnost

Podatki bi morali biti legitimen rezultat raziskav in jih je obvezno citirati. Citati podatkov bi za merjenje raziskovalčeve znanstvene uspešnosti morali biti enako pomembni, kot so to citati objav.

Rezultati analize so pokazali, da je to načelo v Sloveniji z manjšimi izjemami zelo slabo zastopano. Za njegovo udejanjanje sta ključna dva ukrepa. Prvi je izobraževanje, predvsem študentov, kjer njihovi profesorji oz. mentorji vztrajajo pri korektnem citiranju podatkov v seminarskih nalogah, zaključnih delih in znanstvenih objavah. Drugi ukrep bi, kot predlagajo Berez-Kroeker et al. (2018), morali izvesti uredniški odbori revij in programski odbori konference tako, da bi v navodila za avtorje dodali navodila za ustrezno citiranje jezikoslovnih podatkov, tako kot so jih predhodno za spletne vire. Posebej poudarjamo, da je dobrim praksam navajanja raziskovalnih podatkov v slovenskih publikacijah že posvečen priročnik *Priprava raziskovalnih podatkov za odprt dostop* (Štebe et al., 2015). Avtorji priporočajo, »da se v seznamu uporabljene literature podatke navaja s *polno navedbo avtorja oz. avtorjev, naslova, mesta dostopa do podatkov in stalnega identifikatorja*, skladno z oblikovnimi zahtevami znanstvene revije« (2015: 13; naš poudarek).

[16] Repozitorij *CLARIN.SI* rešuje ta problem tako, da je v navodilih za navajanje virov, ki so podani kot prva informacija v glavi vnosa za posamezen vir, jasno izpostavljeno, za katero različico vira gre in ali je ta različica dostopna preko konkordančnika. Za starejše različice repozitorij opozori o morebitni zastarelosti podatkov. Primerjaj npr. vnos za drugo različico korpusa *Gos VideoLectures (Transcriptions)* (Verdonik et al., 2017), ki je dostopna preko konkordančnika *KonText*, s prvo (Verdonik et al., 2016), ki preko taistega konkordančnika ni dostopna.

[17] https://www.sketchengine.eu [15. 4. 2018]

Tu so ključni naslovniki *Slavistična revija* kot revija s posebnega seznama ARRS, konferenca oz. monografija *Obzorja*, kot tudi pričujoča konferenca *Jezikovne tehnologije in digitalna humanistika*, ki bi na tem področju morala orati ledino. Ni odveč omeniti, da je citiranje obvezna sestavina v uvodu omenjeni Nacionalni strategiji odprtega dostopa in njenem Akcijskem načrtu, izpostavljeno pa je tudi v raznih drugih razpravah o odprtih podatkih v Sloveniji, vključno z nalogami financerja in uredništev revij.

4.2. Priznanje zaslug in avtorstva

Citiranje podatkov bi moralo služiti priznavanju znanstvenih zaslug, normativnega ter pravnega avtorstva vsem, ki so prispevali k njihovi izdelavi.

Za priznavanje znanstvenih zaslug je v Sloveniji merodajen SICRIS, ki se za štetje citatov zanaša na Web of Science in SCOPUS. Vplivanje na štetje citatov znanstvenih podatkov je tako izven dometa pričujočega članka.

Lahko pa v Sloveniji vplivamo na to, kako se točkujejo objave znanstvenih podatkov. Trenutno v sistemu COBISS že obstaja rubrika »2.20 Zaključena znanstvena zbirka podatkov ali korpus«, vendar ima takšen vnos priznanih samo 5 točk. Bistveno bolje so lahko točkovane objave pod to rubriko v primerih, ko je vir podatkov naveden v seznamu »Zaključene znanstvene zbirke podatkov, ki se upoštevajo pri kategorizaciji znanstvenih publikacij (BIBLIO-D)«.[18] Trenutno je na tem seznamu samo Arhiv družboslovnih podatkov (ADP). Pomembne objave v ADP tako privzeto dobijo 30 točk (Vončina, 2016), če so deponirani podatki s strani komisije ADP ocenjeni kot zelo pomembni.

Za jezikoslovne podatke bi bilo potrebno tudi repozitorij CLARIN.SI uvrstiti na seznam BIBLO-D, kar pa bi poleg samega predloga komisiji ARRS zahtevalo tudi bolj podrobna navodila za vnašanje virov, kot tudi ustanovitev komisije za vrednotenje vnesenih virov. Vse to pa seveda tudi zahteva precejšen vložek dela in s tem financiranja CLARIN.SI.

4.3. Dokazi

V znanstvenih objavah, kadarkoli in kjerkoli neka trditev sloni na podatkih, bi morali biti ti podatki ustrezno citirani.

Podobno kot za 1. načelo (pomembnost) je tudi tu ključno izobraževanje, navodila za avtorje in uredniška politika publikacij.

4.4. Nedvoumna identifikacija

Citiranje podatkov naj bi vsebovalo trajno metodo identifikacije, primerno za strojno obdelavo, mednarodno edinstveno in široko sprejeto v skupnosti.

Ta pogoj je v veliki meri že realiziran v sklopu repozitorija CLARIN.SI. Vsak vir ima trajni identifikator PID (*persistent identifier*) po sistemu »handle«, na vrhu strani pa je jasno napisano, kako naj se vir citira, pri čemer navedek vsebuje tudi identifikator handle. Repozitorij CLARIN.SI tudi podpira izvoz metapodatkov po shemi Dublin Core, ki jih žanje več agregatorjev: CLARIN VLO,[19] OpenAIRE,[20] re3data[21] in OAI.[22]

V jezikoslovju je poleg navajanja vira kot podatkovne zbirke pomembno tudi navajanje poizvedbe v konkordančniku. Konkordančniki CLARIN.SI, kot tudi konkordančniki projekta Sporazumevanja v slovenskem jeziku (torej konkordančnik za Gigafido,[23] Kres[24], itd.) so vsi narejeni po principu REST, da torej URL poizvedbe zadošča za ponovno in enako poizvedbo.[25] Z drugimi besedami, če po poizvedbi in prikazu rezultatov shranimo URL rezultata, je možno ta URL shraniti, in prek njega ponovno dobiti iste rezultate. Tu velja še opomba, da so takšni URL-ji tipično zelo dolgi in zato neprimerni ali vsaj težavni za citiranje. Vendar pa za krajšanje URL-je obstaja več spletnih storitev, od katerih je posebej zanimiva shortref.org[26], ki jo ponuja češki LINAT/CLARIN. Za razliko od drugih krajševalnikov ponuja shortref.org opis poizvedbe, vrne pa trajni identifikator po sistemu *handle*.

4.5. Dostop

Citiranje podatkov naj bi pripomoglo k dostopu do samih podatkov in do povezanih metapodatkov, dokumentacije, programske opreme in drugih materialov, ki so potrebni, da tako ljudje kot računalniki te podatke informirano uporabljajo.

Ta zahteva je tudi že v veliki meri realizirana v sklopu repozitorija CLARIN.SI, saj vsak vnos vsebuje tako metapodatke kot tudi same podatke, ki so pred vključitvijo v repozitorij preverjeni s strani urednikov.

4.6. Trajnost

Enoznačni identifikatorji in metapodatki, ki opisujejo podatke, morajo biti trajni, celo bolj kot sami podatki.

Repozitorij CLARIN.SI je del slovenske in evropske infrastrukture, domuje pa na Institutu »Jožef Stefan«, ki ima visoko razvito računalniško infrastrukturo. Oboje v največji možni meri ponuja garancijo za dolgotrajnost (meta)podatkov, deponiranih v repozitoriju. K trajnosti metapodatkov pa prispeva tudi že omenjeno dejstvo, da se le-ti redno izvažajo v več spletnih agregatorjev. CLARIN.SI izvaja tudi redne testiranje skladnosti in povezljivosti podatkov.

4.7. Specifičnost in preverljivost

Citiranje podatkov naj bi pripomoglo identifikaciji, dostopu in preverjanju specifičnih podatkov, ki podpirajo neko trditev. Citiranje ali metapodatki citiranja naj bi vsebovali podatke o izvoru in stabilnosti v zadostni meri, da omogočijo preverbo, da je specifičen časovni okvir, različica ali del podatkov, ki so bili naknadno prevzeti, enak kot podatki, ki so bili izvorno citirani.

Tudi tu CLARIN.SI v veliki meri zadošča temu načelu. Vnosi v repozitorij se ne spreminjajo, v primeru dopolnjenih ali popravljeni podatkov se ti vpišejo v nov vnos, vendar z medsebojno povezavo med starim in novim vnosom. Velja posebej poudariti, da je nadzor nad različnimi verzijami, ki ga repozitorij CLARIN.SI

[18] http://home.izum.si/COBISS/bibliografije/Kateg-znan-zbirke.html
[19] https://vlo.clarin.eu/
[20] https://www.openaire.eu/
[21] https://www.re3data.org/repository/r3d100011922
[22] http://www.language-archives.org/archive/clarin.si
[23] http://www.gigafida.net/
[24] http://www.korpus-kres.net/
[25] Seveda, če se medtem ni spremenil korpus.
[26] http://shortref.org/

omogoča, eno izmed priporočil ustreznega digitalnega skrbništva jezikovnih podatkov (npr. Štebe et al., 2015: 6). Mnogo virov je zapisanih po priporočilih TEI, ki tipično vsebujejo bogate metapodatke, s katerimi je mogoče podrobno določiti želene izseke virov.

4.8. Interoperabilnost in fleksibilnost

Metode za citiranje podatkov naj bi bile fleksibilne v zadostni meri, da omogočajo različne prakse med skupnostmi, vendar se ne smejo razlikovati v tolikšni meri, da ogrozijo interoperabilnost praks citiranja podatkov med skupnostmi.

Repozitorij mdr. navaja naslov in avtorje vsakega vira ter na vrhu dostopne strani vira točno definira, kako je vir potrebno citirati.

5. Zaključki

V prispevku smo predstavili rezultate študije, s katero smo preverjali stanje citiranja jezikoslovnih podatkov, predvsem korpusov, v najpomembnejših slovenskih znanstvenih revijah in zbornikih, ki so bili objavljeni v zadnjih petih letih. Izvedli smo pregled navodil za avtorje ter kvantitativno in kvalitativno analizirali obseg in način navajanja virov, s katerim smo pokazali, da stanje ni zavidljivo in si je zato potrebno prizadevati za ozaveščanje, izobraževanje in podporo v skupnosti.

Po opravljeni analizi ugotavljamo, da na jezikovnih virih temelji manj kot petina vseh objavljenih prispevkov, kar je glede na stopnjo razvitosti in razpoložljivosti jezikovnih virov za slovenščino malo in kaže na izključenost skupnosti, ki vire razvija, iz »mainstream« jezikoslovne raziskovalne skupnosti pri nas. Kjer pa so bili v raziskavi uporabljeni, pa jih v skoraj petini prispevkov avtorji sploh ne navajajo. To kaže na pomanjkanje ozaveščenosti jezikoslovcev o pomenu navajanja vseh virov v znanstvenem publiciranju.

S prispevkom, v katerem smo predlagali načela za ustrezno citiranje digitalnih jezikovnih virov, ki temeljijo na mednarodnih poročilih, smo storili prvi korak v tej smeri. Brez tega onemogočamo preverljivost, ponovljivost in nadgrajevanje prejšnjih raziskav, ki so osnovni temelji odprte znanosti. Korektno citiranje jezikovnih virov pa je pomembno tudi zato, ker je v njihov razvoj potrebno vložiti izjemno veliko truda in časa, znanstveni citati pa so najpomembnejši indikator znanstvene uspešnosti.

Ozaveščanje in izobraževanje bi bilo potrebno začeti že v okviru univerzitetnih študijskih programov in poskrbeti za ustrezne smernice za navajanje virov tudi v tem kontekstu. Na področju ozaveščanja skupnosti aktivnih raziskovalcev pa bi z izobraževalnimi dogodki in spletnimi gradivi veliko lahko pripomogla nacionalna raziskovalna infrastruktura CLARIN.SI.

Čim prej bi bilo potrebno vzpostaviti dialog s knjižničarji in uredništvi, ki imajo neposreden stik z raziskovalci in tako tudi veliko moč pri promoviranju dobrih praks citiranja jezikovnih virov, zaradi česar so eni najpomembnejših akterjev pri vzpostavljanju in zagotavljanju dobrih praks za citiranje.

Prav tako pa je nujno potrebno poskrbeti tudi za ozaveščanje razvijalcev virov, ki lahko k ustreznemu citiranju veliko pripomorejo tako, da ustrezno deponirajo in dokumentirajo svoje vire. Za odprto znanost namreč še zdaleč ni dovolj, da nek vir obstaja in je dostopen, temveč mora biti tudi opremljen z vso potrebno spremno dokumentacijo, med katero vključujemo tudi navodila za citiranje. Opuščanje teh praks že na prvem koraku zavira ustrezno citiranje, avtorji pa pri tem pogosto ostajajo nemočni. K temu bi lahko z nudenjem ustrezne dokumentacije, izobraževanj in tehnične podpore veliko doprinesla nacionalna raziskovalna infrastruktura CLARIN.SI.

V prihodnje bi bilo zanimivo raziskavo razširiti na jezikovne vire s področja eksperimentalnega in računalniškega jezikoslovja, ki jezikovne vire uporabljajo kot podatkovne množice, zaradi česar se njihovi interesi, pa tudi potrebe razlikujejo od skupnosti, ki smo se jim posvetili v tej raziskavi. Prav tako načrtujemo dodatno analizo praks navajanja primerov v slovenskih znanstvenih publikacijah s področja jezikoslovja.

Zahvala

Avtorji se anonimnim recenzentom zahvaljujejo za zelo koristne pripombe. Raziskava, opisana v prispevku, je bila opravljena v okviru raziskovalnih infrastruktur za jezikovne vire in orodja CLARIN.SI in CLARIN ERIC.

6. Literatura

Špela Arhar Holdt in Kaja Dobrovoljc. 2016. Vrednost korpusa *Janes* za slovensko normativistiko. *Slovenščina 2.0*, 2: 1–37. http://slovenscina2.0.trojina.si/arhiv/2016/2/Slo2.0_2016_2_02.pdf. Zadnji dostop 10.4.2018.

Špela Arhar Holdt, Kaja Dobrovoljc in Iztok Kosem. 2016. Predstavitveni portal spletnih jezikovnih virov za slovenščino. V: T. Erjavec in D. Fišer, ur., *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, str. 27–31. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Arhar-et-al_Predstavitveni-portal-spletnih-jezikovnih-virov-za-slo.pdf. Zadnji dostop 20.4.2018.

Blanca Arias-Badia, Elisenda Bernal in Araceli Alonso. 2014. An online Spanish Learners' dictionary: the Daele project. *Slovenščina 2.0*, 2: 53–71. http://slovenscina2.0.trojina.si/arhiv/2014/2/Slo2.0_2014_2_05.pdf. Zadnji dostop 10.4.2018.

Júlia Bálint Čeh in Iztok Kosem. 2017. Prvi koraki do novega velikega slovensko-madžarskega slovarja: analiza relevantnih dvojezičnih virov. *Slovenščina 2.0*, 2. http://slovenscina2.0.trojina.si/arhiv/2017/2/Slo2.0_2017_2_06.pdf. Zadnji dostop 17.8.2018.

Andrea L. Berez-Kroeker, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer in Lauren B. Collister. The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest Group. 2017. The Austin Principles of Data Citation in Linguistics (Različica 0.1). http://site.uit.no/linguisticsdatacitation/austinprinciples/. Dostop 15.4.2018.

Andrea L. Berez-Kroeker, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice in Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1): 1–18. https://doi.org/10.1515/ling-2017-0032.

European Commission. 2012. Towards better access to scientific information: Boosting the benefits of public investments in research. http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf. Dostop 13.8.2018.

Tomaž Erjavec. 2009. Odprtost jezikovnih virov za slovenščino. V: M. Stabej, ur., *Simpozij OBDOBJA 28*. http://centerslo.si/wp-content/uploads/2015/10/28-Erjavec.pdf. Dostop 13.8.2018.

Tomaž Erjavec, Darja Fišer, Simon Krek in Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. V: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. http://www.lrec-conf.org/proceedings/lrec2010/summaries/139.html. Zadnji dostop 17.8.2018.

Melanija Larisa Fabčič. 2014. Mentalna podoba človeka v slovenskih, nemških in madžarskih primerjalnih frazemih. *Slavistična revija*, 62(2): 195–215. https://srl.si/sql_pdf/SRL_2014_2_05.pdf. Zadnji dostop 10.4.2018.

Martin Haspelmath. 2014. The Generic Style Rules for Linguistics. *Zenodo*. https://doi.org/10.5281/zenodo.253501.

Nataša Jakop. 2014. Leksikalizacija prostorskih razmerij v slovenščini: jezikovnopragmatični vidik. *Slavistična revija*, 62(3): 353–362. https://srl.si/sql_pdf/SRL_2014_3_08.pdf. Zadnji dostop 20.4.2018.

Adam Kilgarriff in Irene Renau. 2013. esTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences*, 95, 12-19. https://doi.org/10.1016/j.sbspro.2013.10.617.

Nikola Ljubešić, Marija Stupar, Tereza Jurić in Željko Agić. 2013. Combining Available Datasets for Building Named Entity Recognition Models of Croatian And Slovene. *Slovenščina 2.0*, 2. http://slovenscina2.0.trojina.si/arhiv/2013/2/Slo2.0_2013_2_03.pdf. Dostop 13.8.2018.

Nataša Logar Berginc, 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. https://www.fdv.uni-lj.si/docs/default-source/zalozba/pages-from-logar-et-al---korpusi.pdf?sfvrsn=2. Zadnji dostop 17.8.2018.

Nataša Logar, Polona Gantar in Iztok Kosem. 2014. Collocations and examples of use: a lexical-semantic approach to terminology. *Slovenščina 2.0*, 1: 41–61. http://slovenscina2.0.trojina.si/arhiv/2014/1/Slo2.0_2014_1_03.pdf. Zadnji dostop 10.4.2018.

Matej Meterc. 2013. Antonimija enako motiviranih paremioloških enot (primeri iz slovenščine in slovaščine). *Slavistična revija*, 61(2): 361–376. https://srl.si/sql_pdf/SRL_2013_2_02.pdf. Zadnji dostop 10.4.2018.

Janja Polajnar. 2013. Neprodani in trdni. Ja, seveda, potem pa svizec ... Osamosvajanje oglasnih sloganov v slovenskem jeziku. *Jezik in slovstvo*, 58(3): 3–19. https://www.jezikinslovstvo.com/pdf.php?part=2013|3|3%E2%80%9319. Zadnji dostop 10.4.2018.

Janja Ribič. 2016. Ujemanje med povedkom in osebkom v kopulativnih stavkih. *Jezik in slovstvo*, 61(2): 139–147. https://www.jezikinslovstvo.com/pdf.php?part=2016|2|139%E2%80%93147. Zadnji dostop 10.4.2018.

Jörg Tiedemann. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. V: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (ur.): *Recent Advances in Natural Language Processing*, 5: 237–248. Amsterdam, Philadelphia: John Benjamins. http://stp.lingfil.uu.se/~joerg/published/ranlp-V.pdf. Zadnji dostop 17.8.2018.

Ada Vidovič Muha. 2015. Propozicija v funkcijski strukturi stavčne povedi – vprašanje besednih vrst (poudarek na povedkovniku in členku). *Slavistična revija*, 63(4): 389–406. https://srl.si/sql_pdf/SRL_2015_4_04.pdf. Zadnji dostop 10.4.2018.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna, http://wackybook.sslmit.unibo.it/pdfs/sharoff.pdf. Dostop 13.8.2018.

Janez Štebe, Sonja Bezjak in Irena Vipavc Brvar. 2015. Priprava raziskovalnih podatkov za odprt dostop. Priročnik za raziskovalce. Fakulteta za družbene vede, Založba FDV. https://www.dlib.si/details/URN:NBN:SI:DOC-06SLBVXX.

Saška Stumberger. 2015. Slovaropisna obravnava novejše leksike. *Slovene Linguistic Studies*, 10: 153–166. https://kuscholarworks.ku.edu/bitstream/handle/1808/18316/08_Stumberger.pdf. Zadnji dostop 15.4.2018.

Darinka Verdonik in Mirjam Sepesy Maučec. 2013. *Slovenščina 2.0*, 1. http://slovenscina2.0.trojina.si/arhiv/2013/1/Slo2.0_2013_1_06.pdf. Dostop 13.8.2018.

Darinka Verdonik, Tomaž Potočnik, Mirjam Sepesy Maučec in Tomaž Erjavec. 2016. *Spoken corpus Gos VideoLectures 1.0 (transcription)*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1069.

Darinka Verdonik, Tomaž Potočnik, Mirjam Sepesy Maučec in Tomaž Erjavec. 2017. *Spoken corpus Gos VideoLectures 2.0 (transcription)*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1158.

Mira Vončina. 2016. Zaključena znanstvena zbirka podatkov – primeri katalogizacije in Sicris vrednotenja. Delavnica ADP, 26.10. https://www.adp.fdv.uni-lj.si/adp_delavnica_okt2016/presentations/2016_MiraVoncina_Znanstvena_zbirka_podatkov.pdf. Zadnji dostop 7.9.2018.

Andreja Žele. 2014. Členki tudi kot vnašalniki novih prostorskih razmerij v obstoječe sporočilo. *Slavistična revija*, 62(3): 321–330. https://srl.si/sql_pdf/SRL_2014_3_05.pdf. Zadnji dostop 10.4.2018.

Andreja Žele. 2015. Konverzija v slovenščini. *Jezik in slovstvo*, 60(2): 65–77. https://www.jezikinslovstvo.com/pdf.php?part=2015|2|65%E2%80%9377. Zadnji dostop 15.4.2018.

Glagolske večbesedne enote v učnem korpusu ssj500k 2.1

Polona Gantar,* Špela Arhar Holdt,† Jaka Čibej,‡ Taja Kuzman,♣ Teja Kavčič♦

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 12, 1000 Ljubljana
apolonija.gantar@guest.arnes.si
† CJVT (Fakulteta za računalništvo in informatiko, Filozofska fakulteta), Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
spela.arhar@cjvt.si
‡ Laboratorij za umetno inteligenco, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
jaka.cibej@ijs.si
♣kuzman.taja@gmail.com
♦teya10teja@gmail.com

Povzetek

V prispevku predstavljamo kategorije glagolskih večbesednih enot, kot so bile oblikovane v okviru mednarodne COST akcije PARSEME Shared Task 1.1. za 26 različnih jezikov, in izdelavo učnega korpusa glagolskih večbesednih enot za slovenščino. Osnovni namen prispevka je opisati prve kvantitativne in kvalitativne analize, ki bodo predstavljale izhodišča za izdelavo modela za strojno luščenje večbesednih enot iz korpusnih besedil in za izdelavo leksikona večbesednih enot za slovenščino. V prvem delu prispevka predstavimo postopek označevanja ter označevalne smernice s prilagoditvami za slovenščino, korpusno gradivo ter označevalni program. V drugem delu prispevka natančneje predstavimo 3.364 večbesednih glagolskih enot (iz skupno 13.511 ročno pregledanih povedi) po pripisanih kategorijah: inherentno povratni glagoli, zveze z glagoli v pomensko oslabljeni rabi, predložnomorfemski glagoli in glagolski idiomi. Prispevek sklenemo z razpravo in načrti za prihodnje delo

Verbal Multi-Word Expressions in the Slovene Training Corpus SSJ500k 2.1

The paper presents the categories of verbal multi-word expressions (VMWEs) as developed within the international PARSEME COST Action Shared Task 1.1 for 26 different languages, and the annotation of a Slovene training corpus of VMWEs. The main goal of the paper is to describe the first quantitative and qualitative analyses of VMWEs that will serve as a basis for building a model for the automatic extraction of VMWEs from corpus texts, as well as for the compilation of a lexicon of Slovene MWEs. We begin the paper by presenting the annotation process, the annotation guidelines adapted to Slovene, the corpus, and the annotation tool used. This is followed by a detailed analysis of 3,364 VMWEs (from a total of 13,511 manually annotated sentences) divided into four categories: inherently reflexive verbs, light-verb constructions, inherently adpositional verbs, and verbal idioms. We conclude the paper with a discussion and the description of our plans for future work.

1. Uvod

Večbesedne enote (VE) so prepoznane kot obsežen del mentalnega leksikona govorcev določenega jezika, zato so pomembne tako za jezikoslovne raziskave kot za izgradnjo računalniško procesljivih jezikovnih virov, ki omogočajo izdelavo elektronskih leksikonov VE in razvoj orodij za njihovo procesiranje.

Obstaja več definicij VE, ki se razlikujejo glede na metodološko-teoretična izhodišča in raziskovalne cilje. Jezikoslovni, ali natančneje slovarski vidik, postavlja v ospredje semantične lastnosti VE in jih opredeljuje kot različne tipe zvez, ki izkazujejo določeno stopnjo idiomatičnega pomena (Atkins in Rundell, 2008: 166) ali z drugimi besedami, kot zveze, katerih celostni pomen ni vsota pomenov posameznih sestavin. Definicija VE, oblikovana za namene strojnega procesiranja, na drugi strani izpostavlja (ne)zmožnost njihove razstavljivosti na samostojne lekseme ob ohranitvi pomenskih lastnosti in skladenjske funkcije ter izražanje t. i. leksikalne, skladenjske, pomenske, pragmatične in statistične idiomatičnosti (Baldwin in Kim, 2010: 3). Čeprav ne obstaja splošno sprejeta definicija VE, se tako jezikoslovna kot NLP skupnost strinjata, da je osnovna lastnost, ki ločuje VE od prostih zvez, specifično razmerje, ki obstaja med elementi VE. To razmerje se navadno obravnava v okviru konceptov, kot so kolokabilnost (ali statistična idiomatičnost), idiomatičnost (ali semantična nerazstavljivost), sintaktična (ne)fleksibilnost, ki vključuje tudi možnost notranje modifikacije zveze in nezaporednost leksikaliziranih elementov, ter leksikalna variantnost. Zaradi naštetega predstavljajo VE problem ne samo pri jezikovnih analizah, ampak tudi pri strojnem procesiranju in avtomatski prepoznavi v besedilu.

Eden od načinov za izboljšanje jezikovnotehnoloških nalog, ki vključujejo obravnavo VE, je poznavanje njihovih temeljnih jezikovnih lastnosti ter – na tej osnovi – razvoj metode in standardov za prepoznavanje različnih tipov VE v tekočem besedilu. Če želimo omogočiti, da bo čim večji nabor novorazvitih postopkov dobro deloval tudi za slovenščino, je treba izdelati jezikoslovne analize, ki upoštevajo specifike slovenščine in so hkrati kompatibilne na medjezikovni ravni.

Okvir analize, katere rezultate opisujemo v prispevku, določa sodelovanje v okviru COST akcije PARSEME Shared Task 1.1, rezultati pa bodo uporabni pri izdelavi elektronskega leksikona VE za slovenščino, ki je ena od aktivnosti projekta *Nova slovnica slovenskega jezika: viri in metode*, posredno pa tudi za izdelavo jezikovnih priročnikov, kot sta npr. Slovar sodobnega slovenskega jezika (Gorjanc et al., 2015) in na korpusu temelječa znanstvena slovnica.

V prispevku najprej opišemo postopek prepoznavanja potencialnih glagolskih večbesednih enot (GVE) in posamezne kategorije, kot so definirane v smernicah PARSEME Shared Task 1.1. Nato opišemo postopek označevanja GVE v učnem korpusu ter orodje za označevanje. V nadaljevanju opišemo prve rezultate analiz ročno označenih primerov, in sicer tako s kvantitativnega kot kvalitativnega vidika. V prvem primeru nas je zanimala zastopanost posamezne kategorije v korpusu, frekventnejši posamezni primeri ter zastopanost posameznih elementov znotraj zveze. Jedro prispevka je namenjeno opisu strukturnih, skladenjskih in pomenskih lastnosti prepoznanih GVE.

2. Kategorije glagolskih večbesednih enot in postopek prepoznavanja

Kategorizacija GVE temelji na smernicah, izdelanih v okviru PARSEME Shared task 1.1 (Bathia et al., 2017), definicija posamezne kategorije pa se opira na pomenske in skladenjske lastnosti glagolske zveze, ki so opisane v obliki odločitvenih drevesnic. Prepoznavanje in kategorizacija sta potekala v treh korakih. V prvem koraku smo identificirali zveze glagola z vsaj še eno besedo, ki predstavljajo potencialne GVE. V drugem koraku smo prepoznavali leksikalizirane elemente zveze, tj. elemente brez katerih GVE ne more obstajati, v tretjem koraku pa smo se na podlagi podrobnih jezikovnih testov v obliki generičnih in specifičnih jezikovnih meril odločali, v katero kategorijo sodi prepoznana GVE.

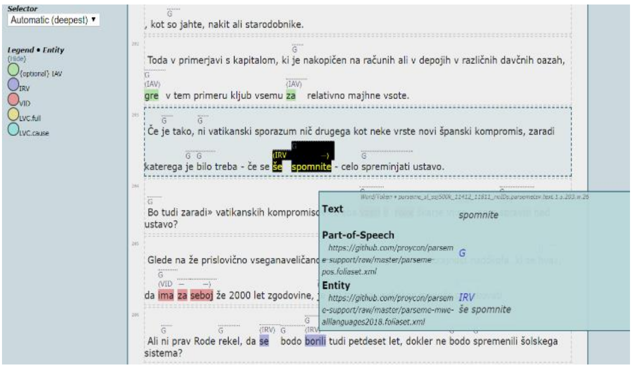
Na podlagi smernic so prepoznane GVE razdeljene na kategorije znotraj dveh razredov, določenih glede na to, ali je kategorijo mogoče aplicirati na večino jezikov, vključenih v raziskavo, ali pa je značilna samo za posamezne jezike. Univerzalne kategorije vključujejo *zveze z glagoli v pomensko oslabljeni rabi* (ang. LVC: Light Verb Constructions), ki so nadalje ločne na celostne (ang. LVC.full) in na kavzativne oz. vzročnostne (ang. LVC.cause), ter na *glagolske idiome* (ang. VID: Verbal Idioms). T. i. kvaziuniverzalne kategorije, ki so vezane na posamezne jezikovne skupine, vključujejo *inherentno povratne glagole* (ang. IRV: Inherently Reflexive Verbs), značilne za večino slovanskih jezikov, ter zveze glagola z izpredložnim morfemom (ang. VPC: Verb-Particle Constructions), značilne predvsem za germanske jezike. Zadnja kategorija je bila v drugi verziji Smernic dopolnjena s tipom predložnih glagolskih zvez (ang. IAV: Inherently Adpositional Verbs), ki predvidevajo odprto skladenjsko mesto in so značilne tudi za slovenščino in nekatere druge slovanske jezike. V prispevku jih imenujemo *predložnomorfemski glagoli* z leksikaliziranim predložnim morfemom.

Za slovenščino je bilo mogoče registrirati GVE za vse predvidene kategorije, razen za VPC, pri čemer obstajajo za posamezne kategorije posebnosti, povezane bodisi s skladenjskimi in morfološkimi lastnostmi slovenščine bodisi s slovničnimi kategorijami, ki so splošno uveljavljene v jeziku in se deloma razlikujejo od drugih jezikov. Na slovenske posebnosti bomo v nadaljevanju opozorili ob posameznih tipih GVE.

3. Korpus in označevalnik

Za označevanje GVE smo uporabili učni korpus ssj500k 2.0 (Krek et al., 2017), ki vključuje približno 500,000 pojavnic in nekaj manj kot 28.000 stavkov iz vzorčenih odstavkov korpusa FidaPLUS (Arhar Holdt in Gorjanc, 2007). Korpus je v celoti označen na oblikoskladenjski ravni (Grčar et al., 2012), v posameznih deležih pa še na ravni lastnoimenskih entitet in skladenjskih razčlemb (Dobrovoljc et al., 2012). Naslednja različica korpusa ssj500k 2.1 vključuje tudi semantične oznake v obsegu 5.500 stavkov (Krek et al., 2018). V prvi fazi označevanja je bilo s kategorijami GVE, kot jih je določala prva verzija Smernic (Candito et al., 2016), označenih 11.411 stavkov s strani dveh označevalcev, nestrinjanja v odločitvah pa so bila prediskutirana in ustrezno popravljena. V drugi fazi so bile kategorije avtomatsko preoznačene na podlagi druge verzije Smernic ter ročno pregledane. S posodobljenimi kategorijami je bilo v drugi fazi dodatno označenih še 2.100 stavkov s strani enega označevalca, pri čemer so bili problematični primeri prav tako prediskutirani in ustrezno popravljeni. Celotni izplen vseh pojavitev GVE v učnem korpusu, kot je razvidno iz Tabele 1, je 3.364 enot.

Za označevanje smo v prvi fazi uporabili orodje SentenceMarkup System, ki je bilo primarno razvito za skladenjsko označevanje slovenščine (Dobrovoljc et al., 2012). Orodje smo prilagodili za namene označevanja GVE tako, da smo mu dodali neodvisen in hkrati medsebojno povezljiv nivo (prim. Gantar et al., 2017). V drugi fazi je označevanje potekalo v spletni anotacijski platformi FLAT (FoLiA Linguistic Annotation Tool), ki je bila prilagojena za namene PARSEME Shared Task in preizkušena na 13 sodelujočih jezikih (Slika 1).



, kot so jahte, nakit ali starodobnike.

Toda v primerjavi s kapitalom, ki je nakopičen na računih ali v depojih v različnih davčnih oazah, gre v tem primeru kljub vsemu za relativno majhne vsote.

Če je tako, ni vatikanski sporazum nič drugega kot neke vrste novi španski kompromis, zaradi katerega je bilo treba - če se še spomnite - celo spreminjati ustavo.

Bo tudi zaradi» vatikanskih kompromiso ustavo?

Glede na že pregovorno vseganaveličano da ima za seboj že 2000 let zgodovine,

Ali ni prav Rode rekel, da se bodo borili tudi petdeset let, dokler ne bodo spremenili šolskega sistema?

Text
spomnite

Part-of-Speech
https://github.com/proycon/parseme-support/raw/master/parseme-pos.foliaset.xml G

Entity
https://github.com/proycon/parseme-support/raw/master/parseme-mwe-alllanguages2018.foliaset.xml IRV še spomnite

Slika 1: Označevalnik FLAT.

Platforma FLAT omogoča označevanje nizov besedila z vnaprej določenimi kategorijami in dodeljevanje datotek različnim označevalcem. Pri uvozu podpira formata XML in TSV, izvoz končnih datotek pa je v formatu XML. Vnesene oznake se med označevanjem shranjujejo samodejno. Vmesnik omogoča tudi iskanje po besedilih s pomočjo iskalnih pogojev v jeziku CQL.

4. Kvantitativna analiza

Označene GVE so bile po koncu označevanja uvožene v učni korpus ssj500k 2.1 (Krek et al., 2018). Od 13.511 stavkov, pregledanih med prvo in drugo fazo označevanja, jih vsaj eno GVE vsebuje 2.920, kar znaša približno 22 %. Vsak od teh stavkov v povprečju vsebuje 1,15 GVE, glede na celoten pregledani nabor stavkov pa je količina GVE na stavek približno 0,25, kar pomeni, da na GVE v povprečju naletimo v vsakem četrtem stavku.

Tabela 1 prikazuje razpored označenih GVE glede na kategorije. Vseh različnih GVE (brez večkratnih pojavitev ene same enote) je bilo slabih 1.100. Po absolutni frekvenci največji delež zajema kategorija IRV (48 %), najmanj enot pa je v kategoriji LVC.cause (2 %). Po visokem številu različnih GVE izstopata kategoriji VID in IAV, najmanj raznoliki pa sta kategoriji LVC.full in LVC.cause.

Kategorija	Vse GVE	Delež	Različne GVE
IRV	1.627	48 %	345
IAV	710	21 %	154
VID	724	22 %	457
LVC.cause	64	2 %	27
LVC.full	239	7 %	103
Skupaj	3.364	100 %	1.086

Tabela 1: Razporeditev označenih GVE glede na kategorije.

Pregledani stavki so večinoma vzeti iz besedil s pisnim prenosnikom (13.277 stavkov oz. 98 %), iz govornega prenosnika pa je le 234 stavkov (2 %). Glede na besedilno zvrst je največ stavkov (9.017 oz. 67 %) iz periodičnih publikacij (časopisi in revije), 3.968 stavkov (29 %) je iz knjižnih besedil, preostali 4 % pa so uvrščeni pod drugo. Glede na čas objave besedila pregledani stavki zajemajo obdobje med letoma 1991 in 2006: 3.616 stavkov (27 %) je bilo objavljenih pred letom 2000, 9.375 (69 %) pa med letoma 2000 in 2006. Pri 520 stavkih (4 %) čas objave ni znan. Večina stavkov (10.859 oz. 80 %) je vzeta iz lektoriranih besedil. Pri 2.459 stavkih (18 %) metapodatek o lektoriranosti ni na voljo, pri 193 stavkih (2 %) pa besedilo ni bilo lektorirano.

Tabela 2 prikazuje najpogostejše strukture GVE glede na besedno vrsto sestavine (G – glagol, S – samostalnik, P – pridevnik, R – prislov, D – predlog, Z – zaimek). Strukture, ki so se v korpusu pojavile manj kot 10-krat, so združene v kategorijo Drugo. Najpogostejše strukture so G + Z, G + D, G + S in G + D + S, ki skupaj zajemajo kar 85 % vseh označenih GVE.

Struktura	Primer	Frekvenca	Delež
G + Z	*bati se*	1.663	49 %
G + D	*priti do*	535	16 %
G + S	*imeti odnos*	372	11 %
G + D + S	*biti pod vtisom*	303	9 %
G + Z + P	*biti si edini*	146	4 %
G + R	*biti res*	136	4 %
G + Z + D + S	*ujeti se v past*	24	1 %
G + P	*biti jasno*	20	1 %
G + P + S	*imeti glavno besedo*	19	1 %
S + G + D + S	*biti na robu propada*	12	<1 %
G + Z + S	*vzeti si čas*	11	<1 %
Drugo	-	123	4 %
Skupaj	-	3.364	100%

Tabela 2: Razporeditev označenih GVE glede na besednovrstno strukturo.

5. Kvalitativne analize

Kvalitativna analiza označenih primerov GVE v učnem korpusu zajema njihove strukturne in pomenske lastnosti. Hkrati so bile na podlagi Smernic, ki definirajo posamezno kategorijo znotraj PARSEME Shared Task, prepoznane specifike slovenščine, ki se kažejo na ravni strukturnih in pomenskih testov. Pri tem nas je zanimala vzorčenost strukture znotraj posamezne kategorije, skladenjsko okolje zveze kot celote ter leksikalne zapolnitve na predvidenih udeleženskih mestih. Na podlagi korpusnih primerov smo želeli prepoznati tudi kazalce pomenske celovitosti, ki so uporabni pri avtomatskem prepoznavanju v besedilu.

5.1. Inherentno povratni glagoli (IRV)

Smernice PARSEME Shared Task 1.1 kot samostojne GVE obravnavajo glagole s prostim morfemom *se/si*, imenovali jih bomo inherentno povratni glagoli. Gre za jezikovnospecifično kategorijo, kjer so kot IRV prepoznane samo zveze, kjer glagol brez *se/si* bodisi ne obstaja (*zdeti se*) bodisi prisotnost *se/si* glagolu spreminja pomen in/ali funkcijo (*dati se* – 'moči'). Za preizkus leksikalne trdnosti zveze kot celote in za ločevanje od vseh drugih zvez glagola + *se/si*,[1] je mogoče uporabiti več pomensko-skladenjskih testov, ki preverjajo obnašanje glagola z vidika odpiranja skladenjskih položajev zveze kot celote, pri čemer so določene spremembe v vzorcu in vlogi udeležencev, ki jih taka glagolska zveza predvideva, lahko tudi znanilec pomenskih sprememb.

V učnem korpusu predstavljajo IRV največji delež znotraj obravnavanih kategorij (gl. Tabelo 1). Med 1.627 označenimi primeri so bili štirje primeri napačno kategorizirani, v dveh primerih pa elementi zveze niso bili ustrezno označeni. Ti primeri so bili iz nadaljnje obravnave izločeni. Med pravilno označenimi primeri (1.621) je bilo mogoče identificirati 339 različnih IRV, med katerimi se v korpusu zveze *bati se, dati se, dogajati se, izkazati se, lotiti se, odločiti se, počutiti se, pogovarjati se, pojaviti se, spominjati se, spomniti se, strinjati se, udeležiti se, vrniti se, zavedati se, zdeti se* in *zgoditi se* pojavijo več kot 20-krat. Variantnost morfema *se/si* se kaže pri manjšem deležu primerov (*premisliti se/si, prizadevati se/si, upati se/si, zapomniti se/si*), v drugih primerih je morfem ustaljen, npr. *bati se, zdeti se; zamišljati si, zaželeti si*.

Najbolj prepoznavna lastnost IRV, za katere je značilno, da brez *se/si* ne obstajajo, je, da glagolska zveza kot celota ne prenese neposrednega predmetnega določila, ki nastopa v udeleženski vlogi prizadetega. Znotraj te skupine je mogoče ločiti dve tipični situaciji: (a) glagolska zveza ne predvideva predmetnega določila v svojem širšem stavčnem vzorcu, npr. *dreti se* : **dreti (se) koga, drstiti se* : **drstiti (se) koga,* lahko pa je (b) predmetno določilo del stavčnega vzorca ob prisotnosti morfema *se/si*, npr. *bati se koga* : **bati koga, izogibati se koga/komu* : **izogibati koga,* zlasti pogosto s predložnim ali dajalniškim predmetnim določilom, npr. *strinjati se s kom, pogovarjati se s kom, odzvati se na kaj, odpovedati se komu/čemu* ipd. Lahko bi rekli, da gre v prvem primeru za neprehodne IRV, kamor poleg naštetih sodijo tudi *zvečeriti se, mračiti se* ipd., ter prehodnimi IRV, kjer je (zlasti predložno) predmetno določilo pričakovani del širšega stavčnega vzorca, ki ga napoveduje glagolska zveza.

[1] Upoštevajoč večfunkcijskost *se/si* so iz obravnave IRV izločeni primeri, kjer gre bodisi za pasivne zgradbe (npr. *kazati se*), povratnost (*umivati se, zlomiti si (roko)*) ali vzajemnost (*poljubiti se*) (Gantar et al., 2017).

Pri glagolih, ki lahko obstajajo tudi brez morfema *se/si*, so kot IRV določene samo tiste zveze, kjer morfem glagolu spreminja pomen. Tudi pomensko celovitost zveze glagola in morfema je mogoče prepoznati na podlagi skladenjsko-pomenskih lastnosti, ki jih posamezni pomen zveze definira v svojem stavčnem vzorcu. V prvi skupini nastopajo (prehodni) glagoli, pri katerih prisotnost oz. odsotnost *se/si* povzroči očitno pomensko spremembo, ta pa je pogosto vezana na pomenske lastnosti osebka, ki je v primeru IRV navadno človeško+, hkrati pa glagoli v taki zvezi pogosto nastopajo v svojem prenesenem pomenu, ki ima tudi prepoznavno dobesedno ustreznico, npr. *delati se* – 'pretvarjati se' : *delati kaj* – 'početi, izdelovati'; *pobrati se* – 'opomoči si' : *pobrati kaj* – 'dvigniti s tal'.

V drugi skupini prehodnih glagolov (tj. glagolov, ki dovoljujejo neposredno predmetno določilo, *si/se*, če se pojavlja ob njih, pa lahko izraža pravo povratnost, zaradi česar zveza ni prepoznana kot IRV, npr. *umivati se – umivati koga*) je za ločevanje povratnih zvez od IRV treba upoštevati predvsem primere, kjer prisotnost *se/si*, kljub temu da vrača dejanje/stanje na osebek, zagotavlja zadosten pomenski prenos, da je zvezo mogoče obravnavati kot celoto, npr. *dokazovati se* : *dokazovati kaj, gristi se* : *gristi kaj, naslikati se* : *naslikati koga/kaj*, in sicer tudi v primerih, kjer osebek ni konkretiziran in izraža splošnost, npr. *izplačati se* : *izplačati koga/kaj, vleči se* : *vleči koga/kaj*, ali vzajemnost dejanja, npr. *ljubiti se* : *ljubiti koga/kaj.*

Pri prepoznavanju zveze glagola s prostim morfemom *se/si* kot IRV je treba upoštevati tudi številne primere, kjer *se/si* izraža pasiv, npr. *ponavljati kaj – kaj se ponavlja, zagotavljati kaj – kaj se zagotavlja*. V takih primerih *se/si* ni del glagola, pač pa samo ena od skladenjskih možnosti umikanja osebka iz stavčnega vzorca.

Leksikalizirane zveze glagola in morfema *se/si* v slovenistični literaturi niso bile obravnavane (izključno) z vidika leksikalne celovitosti, npr. kot samostojna kategorija stalnih besednih zvez, pač pa predvsem z vidika funkcije morfema oz. povratnega zaimka (Toporišič, 2000: 503, 579; Žele, 2012). Pri tem se ugotavlja vloga *se/si* z vidika izražanja različnih stopenj vršilskosti oz. osebkove (ne)udeleženosti, kot npr. v primeru needninskega (*bratiti se* ali *zbrati se*) ali splošnega vršilca dejanja (*tiskati se*) (Žele, 2012: 44, Toporišič, 1982: 244). Njihova ustrezna prepoznava v besedilu z vidika pomenskoskladenjske celovitosti, kot jo opisujemo v prispevku, je pomembna predvsem za strojno prepoznavanje večbesednih enot in njihovo razlikovanje od »prostih« zvez tega tipa. Posledično gre v primeru IRV za enote leksikona, ki jih je kot take smiselno obravnavati v slovarju, bodisi kot samostojne iztočnice bodisi v okviru večpomenskosti.

5.2. Zveze z glagoli v pomensko oslabljeni rabi (LVC)

Da je znotraj sistema Parseme večbesedna enota označena kot LVC, mora ustrezati naslednjim pogojem: sestavljena mora biti iz glagola in samostalnika oz. samostalniške besedne zveze, ki je lahko v obliki predložne zveze, npr. *imeti mnenje, biti v dvomih*, in odpirati mora lastna vezljivostna mesta (npr. *kdo ima predavanje za koga*). Pomensko mora biti povezana z dogajanjem (*imeti predavanje*) ali stanjem (*biti v dvomih*). Glagolski del je lahko dveh tipov: (a) če je glagol pomensko oslabljen oz. k pomenu prispeva pretežno na kategorialni ravni, zvezo uvrstimo v podkategorijo LVC.full, npr. *biti v pomoč*; (b) če lahko osebek razumemo kot vzrok ali vir izraženega dejanja/stanja, zvezo uvrstimo v podkategorijo LVC.cause, npr. *imeti učinek, spraviti v smeh*. V algoritmu za presojanje, ali je določena zveza kandidatka za označevanje ali ne, se upošteva še abstraktnost samostalnika (tip *imeti avto* se ne uvršča med večbesedne enote, idiomatične zveze tipa *imeti mačka* v pomenu 'slabo počutje po uživanju alkohola' pa se uvrščajo v kategorijo VID) in pri uvrščanju v LVC.full zmožnost pretvorbe z izpustom glagola *Janez ima predavanje* → *Janezovo predavanje* (glede slednjih gl. tudi 5.4).

Različne možnosti opredeljevanja zvez s pomensko oslabljenimi glagoli v slovenskem in širšem prostoru pregledno predstavi Soršak (2013), po kateri povzemamo tudi poimenovanje *oslabljenopomenski glagol* (nasproti *polnopomenskemu*), skupaj z opozorilom, da je najustrezneje govoriti o glagolih v pomensko oslabljeni rabi. Kot glavno razliko sistema Parseme v primerjavi z dosedanjimi slovenskimi opredelitvami je mogoče izpostaviti delitev na LVC.full in LVC.cause ter dejstvo, da je pretvorljivost oz. zamenljivosti s polnopomenskim glagolom omenjena le med dodatnimi pogoji v odločevalnih drevesnicah, ni pa med osnovnimi določevalnimi parametri.

Med skupno 303 primeri, označenimi kot LVC (en primer je bil označen napačno), jih je v kategoriji LVC.full 238 (78,8 %) in v LVC.cause 64 (21,2 %). V podatkih se pojavljata dve vrsti struktur: (a) kombinacije glagola in samostalnika (39 oz. 87,1 %) in (b) zveza glagola in predložne samostalniške zveze (39 oz. 12,9 %). Močno prevladujejo zveze z glagolom *imeti* (65,6 %), nekoliko pogosteje se pojavljata tudi *biti* (13,6 %) in *da(ja)ti* (skupaj 9,6 %). Drugi glagoli (*narediti, postaviti, postavljati, ostati, voditi, namenjati, delati, storiti, vzbujati, zbujati, dobiti, zastaviti, spraviti, doseči* in *nositi*) se pojavljajo redkeje in so pogosto vezani na eno samo identificirano zvezo (npr. *ostati v spominu, namenjati pozornost, (v)zbujati vtis*).

Zveze glagola in predložne zveze so glede na razmerja nekoliko več zastopane v kategoriji LVC.cause. V podatkih se pojavljajo izključno zveze s predlogoma *v* (33 oz. 84,6 %) in *na* (6 oz. 15,4 %). Velika večina podatkov (24 oz. 61,5 %) vsebuje *biti v* (*biti v pomoč, biti v podporo, biti v navadi, biti v interesu, biti v dvomih, biti v korist, biti v prednosti, biti v težavah, biti v užitek, biti v sporu, biti v skrbeh*). S 6 pojavitvami sledijo zveze z *imeti v* (*imeti v lasti, imeti v načrtu, imeti v spominu*), nato s po 3 pojavitvami *ostati v* (*ostati v spominu*) ter *biti na* (*biti na voljo*), samo po 1 pojavitev pa imajo *dati na* (*dati na voljo*), *imeti na* (*imeti na izbiro*) in *spraviti v* (*spraviti v smeh*).

V označenih večbesednih enotah nastopa relativno omejen nabor samostalnikov, skupno jih je 97. Najpogostejša sta *težava* (21) in *pravica* (20), sledijo *možnost, mnenje, učinek, vloga, vpliv, vtis, pomoč, občutek, prednost, sreča, korist, vprašanje, volja*, *posledica*. Po pričakovanjih se nekateri od teh samostalnikov pojavljajo izključno v zvezah LVC.full (*pravica, možnost, mnenje, vloga*), drugi v LVC.cause (*učinek, vpliv, vtis, pomoč*), ponekod pa je pripis kategorije vezan na pomen glagola, npr. *dati prednost* → LVC.cause ter *imeti prednost* → LVC.full.

Pri večini primerov (79 oz. 81,4 %) se samostalnik v podatkih pojavlja z enim samim glagolom, npr. *imeti pravico, biti v pomoč, dati predlog*. Dodatni 3 primeri se pojavljajo z vidskimi pari, npr. *dati/dajati soglasje*. Ločeno skupino predstavlja 5 samostalnikov (*težava, mnenje,* *korist, interes, vzrok*), ki se pojavljajo z glagoloma *biti* in *imeti*, npr. *biti v težavah* vs. *imeti težave* (prim. Vidovič Muha, 1998: 307–308, ki utemeljuje povezavo med glagoloma prek izražanja prostorske umeščenosti). Preostalih 10 samostalnikov se pojavlja z različnimi glagoli: *dati, dajati, doseči, imeti učinek*; *dajati, narediti, vzbujati, zbujati vtis*; *postaviti, postavljati, zastaviti vprašanje; biti na voljo, dati na voljo, imeti na voljo*; *imeti v spominu, ostati v spominu*; *dati ime, nositi ime*; *biti v podporo, dati podporo, imeti podporo*; *biti v skrbeh, delati skrbi, imeti skrbi*.

Kot omenjeno (Soršak, 2013), pomenska oslabljenost glagolov ne pomeni pomenske praznosti, kar potrjujejo tudi označeni podatki. Tako v skupini LVC.full kot LVC.cause se pojavljajo glagoli, ki so v rabi prisotni tudi s polnim pomenom, pomensko oslabljenost pa v zvezah LVC dopolnjuje samostalniški del (npr. *imeti* v pomenu 'posedovati' nasproti *imeti posledice* v pomenu 'sprožiti, povzročiti, voditi v posledice'). V pomenskem smislu skupine samostalnikov, ki se pojavljajo v LVC.cause po pričakovanjih opisujejo rezultat določenega dejanja, naj bo to opredelitev vrste rezultata (*učinek, vpliv, vtis*) oz. pozitivna (*korist, užitek*) ali negativna posledica (*muka, preglavica*). Pomensko oslabljeni glagol veže rezultat na stavčni osebek (*nekdo* oz. *nekaj daje, naredi, vzbuja vtis*, je torej povzročitelj dejanja). V določenih primerih so zveze LVC pretvorljive v polnopomenski glagol s sorodno morfološko podobo (npr. *imeti, dajati, dosegati učinek – učinkovati*; *imeti vpliv – vplivati*), ne pa vedno (npr. *vzbujati vtis*; *imeti posledice*).

Na drugi strani je skupina samostalnikov pri LVC.full pomensko bolj heterogena. Poskus delitve v pomenske skupine razkrije, da bi kot stično točko številnih zvez morda lahko izpostavili načrtovanje in ocenjevanje uspeha. Pojavljajo se npr. zveze s samostalniki, ki so vezani na (a) komunikacijo (*mnenje, predlog, vprašanje, izjava, soglasje*), opisujejo (b) potencial za uspeh (*možnost, prednost, priložnost, naskok*), (c) začetne korake (*obljuba, napoved, načrt, pobuda*) ali (č) potencialne razloge za neuspeh (*napaka, pomanjkljivost*). Pojavljajo se tudi skupine, ki opredeljujejo (d) negativno stanje (*težava, strah, dvom, zamera*), (e) pozitivne lastnosti (*moč, pogum, potrpljenje*), (f) dosežene rezultate (*izobrazba, status, posel, mir*) in (g) odnos do še nerealiziranih ciljev (*želja, ambicija, vizija, interes*). Tudi pri tej skupini zvez velja, da so pretvorljive v polnopomenski glagol (npr. *imeti mnenje – meniti*; *dati soglasje – soglašati*) ali ne (*imeti ambicije*; *dati priložnost*).

Kot je razvidno iz navedenih primerov, pokriva kategorija LVC pomensko različne večbesedne enote, ki ponujajo možnost za nadaljnje premisleke in popravke označevanja. Z vidika natančnosti je mogoče premisliti in natančneje opredeliti mejo med zvezami LVC in kolokacijami, pri čemer je lahko vodilo pojavljanje z več glagoli (npr. *postaviti, postavljati, zastaviti vprašanje*). Z vidika priklica pa je treba preveriti, ali so bile v korpusu v resnici označene vse relevantne enote, v prvem koraku morda s pomočjo preverbe besednih skic za identificirane glagole in samostalnike. S slovenističnega vidika bi kazalo na podatkih preveriti še ugotovitve (Žele, 2012: 227–28), da je abstraktni samostalnik navadno v obliki za ednino in v tožilniku.

5.3. Predložnomorfemski glagoli (IAV)

Predložnomorfemski glagoli, imenovani tudi glagoli z leksikaliziranim predložnim morfemom (prim. Žele, 2002), so bili v drugi fazi označevanja vključeni kot neobvezna poskusna kategorija. V Smernicah so kot IAV definirani glagoli, ki brez predložnega morfema ne obstajajo, npr. *simpatizirati z, sprevreči se v, sklicevati se na, apelirati na*, in glagoli, ki jim predložni morfem občutno spremeni pomen, npr. *biti za* – 'strinjati se', *priti do* – 'zgoditi se', *hoditi v/na* – 'obiskovati'. Pri tem je pomembno upoštevati, da udeleženci, ki jih predvideva glagolska zveza kot celota, niso del glagolske večbesedne enote, za razliko od npr. *stati na + trdnih tleh*, so pa bodisi skladenjsko obvezni ali neobvezni (*gre za : *kdo/kaj gre za, vendar: gre za koga/kaj*) in omejeni s slovničnimi, npr. sklon (*simpatizirati s (kom)*), in pomenskimi kategorijami, npr. *priti do (nesreče)* : *priti do (cilja)*.

Na obravnavo predlogov kot prostih glagolskih morfemov naletimo že v Metelkovi slovnici (1825: 247-256, cit. po Žele, 2002: 99), podrobneje jih obravnava tudi Breznik (1916: 250; 1934: 225, cit. po ibid.), izraz »prosti predložni glagolski morfem« pa se v slovenščini ustali v šestdesetih letih (Toporišič, 1967: 111). Podrobneje sta glagole z leksikaliziranim predložnim morfemom v slovenski literaturi obravnavali Žele (2002) in Kržišnik (1994). Prva z vidika stopnje leksikaliziranosti predloga (t. i. leksikalizirani, neleksikalizirani in vezavnodružljivi morfemi), druga pa z vidika frazne trdnosti, tj. bodisi kot frazeološke enote, kjer gre zgolj za strukturno ustaljenost, npr. *biti ob (čem)* – 'nahajati se (ob čem)', ali kot frazeme, kjer gre za leksikalno ustaljenost, npr. *biti ob (kaj)* – 'izgubiti; ne imeti več'.

V učnem korpusu predstavljajo IAV približno petino označenih primerov GVE (gl. Tabelo 1). Med 710 primeri vseh pojavitev, je bilo mogoče identificirati 154 različnih IAV, med katerimi se v korpusu vsaj dvajsetkrat pojavijo zveze *iti za* (vedno z glagolom v tretji osebi ednine – *gre za*), *priti do, vplivati na, skrbeti za, temeljiti na, naleteti na, veljati za* in *biti proti*. V skladu s smernicami smo kot IAV označevali tudi glagolske zveze, sestavljene iz inherentno povratnega glagola (gl. pogl. 5.1) in leksikaliziranega predložnega morfema, kot npr. *ukvarjati se z, nanašati se na, zavzemati se za* ipd.

Pri IAV leksikalizirani predložni morfem običajno sledi glagolu, kar potrjuje 86 % označenih primerov, in sicer se v veliki večini primerov nahaja neposredno za glagolom oz. v njegovi neposredni bližini (+ 3 besede). Izjemo predstavlja *gre za*, kjer je vrivanje služi referiranju na predhodno ubeseditev, npr. *gre (v tem primeru) za*.

Primeri, kjer se predložni morfem z vidika izbire besednega reda nahaja pred glagolom, so v učnem korpusu veliko redkejši. V teh primerih glagol nikoli ne sledi neposredno predložnemu morfemu, razdalja med njima pa je občutno večja, in sicer v petini primerov znaša tri besede ali več. Ta tendenca se zdi zanimiva za strojno prepoznavanje IAV, kjer besednoredna distribucija predloga pred glagolom predvideva upoštevanje razmeroma široke okolice glagola.

Glagole z leksikaliziranim predložnim morfemom je mogoče prepoznati tudi glede nekaterih skupnih pomenskih lastnosti, npr. za izražanje (a) funkcije ali lastnosti, *veljati* *za (favorita, človeka)*,[2] *imenovati za (direktorja), šteti za, (uspeh), označiti za (laž), narediti za (politika), razglasiti za (svetnika), spoznati za (nevarnega), smatrati za (sovražnika)*, b) (ne)strinjanje, npr. *biti za (globalizacijo), govoriti za (združevanje), biti proti (vojni), imeti (kaj) proti*, c) izhajanja, upoštevanja, npr. *temeljiti na (dejstvu), graditi na (zaupanju), nanašati se na (podatke), nasloniti se na (tradicijo), navezovati/navezati se na (besede), opirati se na (izkušnje), izhajati iz (predpostavke)*, č) začetek ali spremembo dejanja/stanja, npr. *pasti v (komo), spuščati se v (polemiko), pahniti v (obup), priti v (formo), priti/prihajati do (spremembe), pasti pod (vpliv), pripeljati do (spoznanja), prerasti v (ljubezen), sprevreči se v (nasprotje)*, d) spremembo lastnosti, oblike, npr. *pretvoriti v (energijo)*, e) preživljanje, prestajanje, npr. *iti skozi (proces), (morati) dati skozi, (ceneje) priti skozi*, f) aktivno delovanje, npr. *ukvarjati se z, baviti se z, ubadati se z, skrbeti za, poskrbeti za, zavzemati se za, potegovati se za, prizadevati si za* itd.

Z vidika obnašanja v širšem stavčnem vzorcu je za IAV značilno, da prisotnost predložnega morfema pogosto spremeni vezljivostne lastnosti glagola, npr. (a) ko prvotno neprehodni glagol postane prehodni, tipično s predložnim določilom, kot v primerih *živeti od koga/česa, gre za koga/kaj*, (b) ko pride do spremembe sklona predložnega določila, *obrniti se na koga : obrniti se h komu, spoznati se na kaj : spoznati se s kom, klicati po kom/čem : klicati koga/kaj*. Prepoznati je bilo mogoče tudi številne primere glagolov premikanja, ki kot IAV spremenijo pomen v neprostorsko vrednotenje stanja, na primer *priti skozi* – 'preživeti', *hoditi v* – 'obiskovati', *pahniti v* – 'povzročiti, da začne kdo doživljati kaj neprijetnega', *priti/pripeljati/prihajati do* – 'zgoditi se'. Glagolom s širokim pomenskim obsegom predložni morfem tipično zoži pomen, kot v primerih *biti za, govoriti za, imeti proti* ipd. Treba pa je omeniti tudi glagole, ki jim v pomenu znotraj IAV obvezno sledi abstraktni predmet, kot npr. *pasti v (nemilost, depresijo, vrtinec nizkotnosti), dišati po (prevari), pokati od (od veselja), postreči z (zanimivostmi)*.

Prepoznavanje predložnomorfemskih glagolov predstavlja izziv tako za označevalce kot za strojno učenje, saj se med leksikalizirani morfem in glagol lahko vrivajo druge besede, poleg tega pa številne zveze glagola s predlogom niso leksikalizirane, npr. *pasti v luknjo/na tla/pod vlak/čez previs*, lahko izkazujejo dobesedni pomen ob tem da ohranjajo nespremenjen tudi sklon predmetnega določila, npr. *stati za (vrati)* – 'nahajati se' : *stati za (dejanji)* – 'podpirati', in so hkrati lahko tudi večpomenske, npr. *priti do (spremembe)* – 'zgoditi se' in *priti do (denarja)* – 'dobiti'.

Analiza predstavlja izhodišča za strojno prepoznavanje tovrstnih enot, hkrati pa ponuja možnosti za bolj poglobljene raziskave, zlasti na ravni vezljivosti, prepoznavanja stavčnih vzorcev in pomenskih lastnosti udeležencev.

5.4. Glagolski idiomi (VID)

Smernice opredeljujejo glagolske idiome[3] (VID) kot zvezo dveh leksikaliziranih sestavin, pri katerih glagol predstavlja skladenjsko jedro, ki predvideva vsaj enega udeleženca znotraj stavčnega vzorca. Udeleženci imajo lahko različne skladenjske vloge, npr. neposrednega ali predložnega predmetnega določila, *plačati ceno*, *zravnati z zemljo*, osebka, *stara zgodba se ponavlja*, prislovnega določila, *spati kot ubit*, odvisnega stavka, *vedeti, koliko je ura*, itd. Poleg omenjenega, mora taka zveza izkazovati tudi samostojen pomen, kar pomeni, da mora ob določenih spremembah skladenjskih in pomenskih funkcij ohranjati svoj pomen. Kot nabor takih sprememb, ki zvezi ohranjajo pomen, Smernice navajajo možnost pojavljanja sestavin v predvidenih paradigmah (sklanjatveni in spregatveni), tvorjenje časov, tvorjenje aktivnih in pasivnih zgradb, leksikalno variantnost itd.

Definicija znotraj Smernic Parseme se od slovenske razlikuje v tem, da obravnava GVE kot glagolsko jedro stavka, ki predvideva leksikalizirane elemente znotraj svojega stavčnega vzorca – pomenskoskladenjski pristop, medtem ko se v slovenski literaturi izpostavlja predvsem zmožnost opravljanja povedkove funkcije zveze kot celote (Toporišič, 1973/74; Kržišnik, 1994) – funkcijsko-skladenjski pristop. S tega vidika so v slovenščini problematične zveze, ki sicer vključujejo glagol kot ustaljeni del, vendar kot celota ne nastopajo nujno le v vlogi povedka, pač pa tudi v vlogi predmetnega določila, *(ne spodobi se) voditi za nos*, ali v vlogi stavka (*srce se trga (komu)*).

V učnem korpusu je bilo kot GID označenih 724 enot, kar predstavlja 22 % vseh GVE (gl. Tabelo 1). GID z več kot 10 pojavitvami po pričakovanju vključujejo glagol *biti* (tudi *imeti*), določilo pa je glede na besednovrstno opredelitev izmuzljivo, saj se lahko hkrati pojavlja v prislovni in členkovni, pridevniški ali samostalniški funkciji, npr. *biti jasno, biti si na jasnem; biti žal, biti stvar (koga/česa).* Z več kot 5 pojavitvami najdemo še *biti kos, biti prav; priti prav, igrati vlogo, pustiti pri miru, priskočiti na pomoč* in *imeti opravka s/z* ter t. i. ustaljene diskurzne označevalce (prim. Dobrovoljc, 2017): *kot se pravi, se pravi, kdo ve*. Glagoli, ki tipično tvorijo različne VID, so poleg *biti* in *imeti* še *vzeti, postaviti, priti, dati* in *iti* (v 10 ali več različnih VID). Med samostalniškimi sestavinami z več kot 10 pojavitvami izstopata *roka* in *glava* ter *beseda, nič, stran* in *vrata* z vsaj 5 pojavitvami v različnih VID.

Kot omenjeno, po frekventnost strukture izstopajo zveze glagola *biti* in prislova/pridevnika/samostalnika, ki jih je glede na strukturno ustaljenost in pomensko izpraznjenost glagola smiselno obravnavati kot ustaljene glagolske zveze oz. leksikonske enote (*biti všeč/res/mar/prida/prav/kos, biti jasno/žal/narobe/stvar/moč*), manj pa se zdi smiselno na podlagi njihove distribucijske omejenosti na pomožnik odpirati samostojno besedno vrsto – povedkovnik (Toporišič, 2000; Žele, 2011). V to skupino sodijo tudi zveze s pomensko širokim *imeti*: *imeti prav/rad, ne imeti pojma/smisla, imeti smisel za* ipd.

Druga struktura, ki je opazno zastopana v učnem korpusu, je zveza glagola in samostalnika oz. samostalniške besedne zveze. Med glagoli izstopata *delati* (*delati družbo/gužvo/izjeme/preglavice/razlike/sceno/škodo*) in *dati* (*dati košarico, dati polet, dati pečat, dajati videz* ipd.), ki strukturno sovpadajo z LVC, vendar ne prenesejo

[2] Ob IAV navajamo še tipične kolokatorje na podlagi korpusa Gigafida za lažje ustrezno pomensko razdvoumljanje.

[3] S tem izrazom se oddaljujemo od slovenske tradicije, ki bi na tem mestu uporabila izraz *glagolski frazemi*. Razlogi so predvsem v drugačnem dojemanju skladenjske vloge glagolske sestavine, kot je pojasnjeno v prispevku.

določenih pretvorb, ki jim sicer podlegajo LVC, npr. izražanje svojine, ki se pri LVC ohranja: *Miha ima predavanje* → *Mihovo predavanje*, pri VID pa taka pretvorba ob ohranitvi pomena ni mogoča: *Miha dela družbo/gužvo* ipd. → **Mihova družba/preglavice*. Tipično se samostalniška zveza razširja s pridevnikom, ki je bodisi leksikaliziran, *imeti polne roke, zadati smrtni udarec*, varianten: *ubrati drugo/drugačno pot, preteči veliko/dosti vode*, ali zgolj tipičen vrivek v sicer ustaljeno glagolsko zvezo: *služiti si (vsakdanji, nogometni* ipd.*) kruh*. Največji delež predstavljajo v učnem korpusu VID s strukturo glagola in predložne zveze, kjer spet izstopa *biti*, npr. *biti na dosegu roke, biti v konfliktu, biti na preizkušnji, biti na razpolago/voljo, biti na tleh, biti na udaru, biti pod pritiskom, biti pri srcu, biti pri stvari* ipd., sicer pa so zastopani tudi drugi glagoli, npr. *postaviti ob bok, potegniti na dan, priti na dan, dati na izbiro, dati na led, priti do izraza, priti na misel, voditi/vleči za nos, trkati na vrata, stati ob strani, pasti v oči, požirati z očmi, zavijati z očmi, postaviti/postavljati na stranski tir, škrtati z zobmi* ipd., sem pa smo šteli tudi primere kot *vzeti (kaj) nase, postaviti se zase, obdržati (kaj) zase* ipd. Zlasti za zveze glagola s samostalniško zvezo in s predložno samostalniško zvezo je z vidika ustaljenosti treba opozoriti na obvezno ali tipično zanikanje, npr. *ne moči*[4] *(komu) do živega, ne moči si kaj, (kaj) ni po godu (komu), (kaj) ne gre v račun (komu), ni ne duha ne sluha o (kom/čem), ni para (komu), ne gre iz glave (komu)* ipd.

V manjšem deležu so v učnem korpusu zastopane tudi druge strukture, npr. stavčne: *solze stopijo v oči (komu), oči so večje od želodca, noge nesejo (koga), stara zgodba se ponavlja, kamen se odvali od srca (komu), časi se spreminjajo, vrabci že čivkajo*, tudi v obliki pregovorov, npr. *bolje preprečiti kot zdraviti, samo osel gre dvakrat na led*, in primerjav: *igrati se (s kom/čim) kot mačka z mišjo, delati (s kom/čim) kot svinja z mehom, steči kot namazano* ipd., zveze glagola in prislova, npr. *priti skupaj, daleč priti, iti predaleč, narediti svoje, ustreliti mimo, imeti zadosti, dobro iti*, ter zveze glagola in zaimenskega morfema, *zagosti jo (komu), ubrati jo, mahniti jo* ipd.

VID se v stavčni vzorec vključujejo na različne načine. Glagolske zveze odpirajo predvidljiva skladenjska mesta, ki jih zapolnjujejo udeleženci s svojimi tipičnimi pomenskimi vlogami, kot smo nakazali pri posameznih primerih zgoraj. Že ob hitrem pregledu primerov v korpusu je mogoče zaznati tudi ustaljenost ali večjo pogostnost nekaterih glagolskih oblik (npr. 3. oseba, zanikanje) pa tudi predvidljivost zapolnitev udeleženskih mest na leksikalni ravni.

V naši raziskavi stavčni vzorci, ki jih narekujejo VID (in druge kategorije GVE), niso bili sistematično raziskani, je pa v ta namen mogoče uporabiti podatke, ki jih vsebuje učni korpus na skladenjski in semantični ravni. V prvem primeru s formaliziranimi skladenjskimi povezavami, v drugem pa s pripisom semantičnih vlog udeležencem na teh mestih. Na ta način bi bilo mogoče identificirati širše stavčne vzorce za posamezni tip GVE in jih uporabiti pri nadaljnjem strojnem luščenju.

6. Razprava in zaključek

Kategorizacija glagolskih večbesednih enot na podlagi Smernic Parseme 1.1 ima dva osnovna namena: (a) določiti merila za prepoznavanje glagolskih večbesednih enot, ki jih je smiselno pri jezikovnem opisu (slovar, slovnica) in strojnem luščenju obravnavati kot celote, tj. elemente leksikona, ter (b) formalizirati opis v skladu z večjezikovno primerljivimi merili.

Na podlagi označenih glagolskih večbesednih enot v učnem korpusu je bilo mogoče izdelati prve kvantitativne in kvalitativne analize za posamezno kategorijo in na njihovi podlagi prepoznati določene načine vzorčenja na skladenjski in pomenski ravni. Ti vzorci predstavljajo dobro izhodišče za izdelavo pravil pri strojnem luščenju GVE in za nadaljnje jezikoslovne opise. Metodološko gledano gre tudi za preusmeritev fokusa s funkcijskoskladenjskega vidika v opis medsebojno povezanih lastnosti na oblikoskladenjski, skladenjski, pomenski in leksikalni ravni.

Glagoli, ki tipično tvorijo GVE, so po pričakovanju glagoli z zelo širokim pomenskim obsegom, npr. *biti, dati, imeti*, zaradi česar izgubljajo svoje leksikalne, ohranjajo pa morfološke lastnosti, skladenjsko funkcijo in pozicijo v stavčnem vzorcu. Pomenska udeleženost glagola v odnosu do posameznih sestavin v zvezi kot celoti je pogosto težko določljiva zaradi velike pogostnosti glagolskih zvez, med katerimi številne ne izkazujejo idiomatičnega branja, prim. zveze z glagoli v pomensko oslabljeni rabi, inherentno povratne glagole in glagole z leksikaliziranim predložnim morfemom. Zaradi tega jih je v tekočem besedilu težko ločiti od prostih zvez pa tudi od kolokacij, ki so frekventne pomensko smiselne in strukturno pravilne besedne povezave. Seznam GVE, ki smo jih prepoznali v korpusu, tako že predstavlja nabor leksikonskih enot, ki jih je kot take mogoče uporabiti pri avtomatskem prepoznavanju v besedilu in nadaljnjem strojnem učenju mehanizmov.

Na drugi strani so prve strukturne in pomenske analize pokazale, da (a) posamezni tipi GVE tvorijo prepoznavne strukturne vzorce, npr. glagol + samostalniška zveza, zlasti predložna, da (b) leksikalizacija elementov vpliva na spremembe v udeleženskih mestih in njihovih semantičnih vlogah, npr. *vreči se po kom – vreči se v kaj – ven se vreči – vreči koga ven* ipd., (c) da je npr. variantnost glagolov predvidljiva z vidika dovršnih in nedovršnih parov, *plačati/plačevati ceno, dati/dajati si opravka s čim*, (č) da zaporedje glagolskih sestavin v posamezni GVE navadno ni ustaljeno, se pa (d) kažejo določene tendence v besednem redu ter (e) številu in zastopanosti vrinjenih elementov, kot tudi, da je (f) na določene leksikalne zapolnitve mogoče sklepati iz frekvenčnih podatkov in elementov besedilnega okolja, ter da je (g) za lažje strojno prepoznavanje GVE smiselno v formaliziran opis vključiti informacije na vseh ravneh korpusne označenosti. V nadaljevanju raziskav, usmerjenih v prepoznavanje večbesednih enot z glagolskim jedrom, bomo zato upoštevali vse vrste podatkov, ki so v zvezi s posameznimi besedami in zvezami v korpusu že na voljo, tj. oblikoskladenjsko označenost, skladenjsko razčlenjenost in pripis pomenskih vlog stavčnim udeležencem.

Za ustrezno identifikacijo različnih VE v jeziku bomo v nadaljevanju izdelali tudi tipologijo večbesednih enot, ki ne predvidevajo glagolskega jedra, kot npr. ustaljene samostalniške zveze tipa *žlahtna kapljica, kaplja v morje,*

[4] V primeru, da je realizacija izključno vezana na 3. osebo, je to razvidno tudi iz osnovne oblike VID. Nedoločnik v osnovni obliki ohranjamo takrat, ko glagol predvideva tudi prvo- in drugoosebne osebke.

domača tla, kjer predstavlja poseben izziv ugotavljanje trdnosti povezave z glagolom, kot npr. *(kaj) je kaplja v morje*, (*zmagati, odigrati) na domačih tleh, (finale) na domačih tleh*. Poleg tega predstavlja v nadaljevanju izziv tudi prepoznavanje VE s samostojnim, vendar ne metaforičnim pomenom, npr. *formula ena, velika začetnica, druga svetovna vojna*, ki se na eni strani približujejo terminološkim na drugi pa lastnoimenskim enotam.

Pri identifikaciji VE bo pozornost treba nameniti tudi ustaljenim skladenjskim strukturam, ki sicer niso pomensko samostojne, imajo pa predvidljivo zgradbo in opravljajo samostojno skladenjsko vlogo, npr. *v času od do*, kamor sodijo tudi številne ustaljene predložne zveze kot npr. *med drugim, v celoti, v skladu z/s, po besedah* ipd. ter t. i. besedilni povezovalci, ki so opazna sestavina tako pisne kot govorne komunikacije (Dobrovoljc, 2017).

7. Zahvala

Raziskava je potekala v okviru projekta ARRS J6-8256 *Nova slovnica sodobne standardne slovenščine: viri in metode* ter sodelovanja v okviru IC1207 PARSEME COST Action[5] in IS1305 ENeL COST Action.[6]

8. Literatura

Špela Arhar Holdt in Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52(2): 95–110.

Sue B. T. Atkins, in Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography.* New York, Oxford University Press.

Timothy Baldwin in Su Nam Kim. 2010. 'Multiword Expressions' V *Handbook of Natural Language Processing*, Second Edition, str. 267–292, CRC Press, Boca Raton, USA.

Archna Bhatia, Claire Bonial, Marie Candito, Fabienne Cap, Silvio Cordeiro, Vassiliki Foufi, Polona Gantar, Voula Giouli, Carlos Herrero, Uxoa Iñurrieta, Mihaela Ionescu, Alfredo Maldonado, Verginica Mititelu, Johanna Monti, Joakim Nivre, Mihaela Onofrei, Viola Ow, Carla Parra Escartín, Manfred Sailer, Carlos Ramisch, Renata Ramisch, Monica-Mihaela Rizea, Agata Savary, Nathan Schneider, Ivelina Stonayova, Sara Stymne, Ashwini Vaidya, Veronika Vincze in Abigail Walsh. 2017. *PARSEME shared task 1.1 annotation guidelines* (last updated on November 30, 2017). http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/.

Marie Candito, Fabienne Cap, Silvio Cordeiro, Vassiliki Foufi, Polona Gantar, Voula Giouli, Carlos Herrero, Mihaela Ionescu, Verginica Mititelu, Johanna Monti, Joakim Nivre, Mihaela Onofrei, Carla Parra Escartín, Manfred Sailer, Carlos Ramisch, Monica-Mihaela Rizea, Agata Savary, Ivelina Stonayova, Sara Stymne in Veronika Vincze. 2016. *PARSEME shared task 1.0 annotation guidelines - version 1.6b* (last updated on November 26, 2016). http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/.

Kaja Dobrovoljc. 2017. Multi-word discourse markers and their corpus-driven identification: the case of MWDM extraction from the reference corpus of spoken Slovene. *International journal of corpus linguistics*, 22(4), 551–582.

Kaja Dobrovoljc, Simon Krek in Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. V *Zbornik Osme konference Jezikovne tehnologije,* str. 42–47, Ljubljana, Institut Jožef Stefan.

Polona Gantar, Simon Krek in Taja Kuzman. 2017. Verbal multiword expressions in Slovene. *Europhras 2017,* str. 247–259. Springer.

Lara Godec Soršak. 2013. Glagoli z oslabljenim pomenom v Slovarju slovenskega knjižnega jezika. *Slavistična revija*: 61(3): 507–522.

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek (ur.). 2015. *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana, Znanstvena založba Filozofske fakultete.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana, Institut Jožef Stefan.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar in Taja Kuzman. 2017. Training corpus ssj500k 2.0, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1165.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek in Anja Zajc. 2018. Training corpus ssj500k 2.1, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1181.

Erika Kržišnik. 1994. *Slovenski glagolski frazemi (ob primeru glagolov govorjenja). Doktorska disertacija*. Filozofska fakulteta, Univerza v Ljubljani.

Jože Toporišič. 2000. *Slovenska slovnica*. Maribor, Založba Obzorja.

Jože Toporišič. 1982. *Nova slovenska skladnja*. Ljubljana, Državna Založba Slovenije.

Jože Toporišič. 1976. *Slovenska slovnica*, Maribor, Obzorja.

Jože Toporišič. 1973/74. K izrazju in tipologiji slovenske frazeologije. *Jezik in slovstvo* (8): 273–279.

Ada Vidovič-Muha. 1998. Pomenski preplet glagolov imeti in biti – njuna jezikovnosistemska stilistika. *Slavistična revija* [na spletu], 46(4): 293–323.

Andreja Žele. 2002. Prostomorfemski glagoli kot slovarska gesla. *Jezikoslovni zapiski* 8(1), 95–108.

Andreja Žele. 2012. *Pomensko-skladenjske lastnosti slovenskega glagola*, (Zbirka Linguistica et philologica, 27). Ljubljana, Založba ZRC, ZRC SAZU.

Andreja Žele. 2011. Povedkovnik kot skladenjska in slovarska kategorija. *Jezikoslovni zapiski*, 17(1): 27–34.

[5] http://www.parseme.eu.

[6] http://www.elexicography.eu.

Towards Semantic Role Labeling in Slovene and Croatian

Polona Gantar,* Kristina Štrkalj Despot, Simon Krek,† Nikola Ljubešić‡**

* Department of translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@guest.arnes.si
** Institute for the Croatian Language and Linguistics
Republike Austrije 16, 10000 Zagreb
kdespot@ihjj.hr
† Artificial Intelligence Laboratory, Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana
simon.krek@ijs.si
‡ Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
nikola.ljubesic@ijs.si

Abstract

In the paper, the semantic role labeling framework is presented, which was developed within the project *Semantic Role Labeling in Slovene and Croatian.* The main goal of the project was the development of an annotated corpus to be used as training data for supervised machine learning systems. In building this framework we follow the path of previous SRL endeavours such as PDT, Vallex, FrameNET, Propbank etc. In compiling the list of semantic roles and their respective formal descriptions, we follow the approach developed by Prague Dependency Treebank, PDT. The paper describes both corpora used for semantic role annotation, as well as tools used in manual annotation tasks. Special attention is directed towards the description of the experimental automatic semantic role labeling based on supervised machine learning methods, and to its possible improvements. A preliminary quantitative analyses is performed for both languages (in terms of verbs range and frequencies, semantic roles, and typical syntactic-semantic patterns for the most frequent verbs).

Označevanje semantičnih vlog za slovenščino in hrvaščino

V prispevku opisujemo model semantičnega označevanja za slovenščino in hrvaščino, ki smo ga razvili v okviru mednarodnega bilateralnega projekta. Osnovni namen projekta je bil izdelati ročno označena korpusa, ki ju bo mogoče uporabiti kot učno množico v sistemih nadzorovanega strojnega učenja za oba jezika. Model sledi dobrim jezikovnim praksam ter široko uveljavljenim modelom na tem področju (PDT, Vallex, FrameNET, Propbank), hkrati pa upošteva značilnosti obeh jezikov kot tudi robustnost semantičnih oznak. V članku opišemo oba učna korpusa in nabor semantičnih oznak ter na kratko povzamemo rezultate poskusnega avtomatskega označevanja s pomočjo nadzorovanega strojnega učenja. V jedrnem delu prispevka opišemo prve rezultate kvantitativnih analiz za oba jezika, in sicer z vidika zastopanosti glagolov, semantičnih oznak in tipičnih pomensko-skladenjskih vzorcev za najfrekventnejše glagole.

1. Introduction

Semantic Role Labeling (SRL) within natural language processing refers to the process of detecting and assigning semantic roles to semantic arguments determined by the predicate or verb of a sentence. This means that in the sentence *My parents gave me a weird name*, the verb *to give* should be recognized as the predicate with three arguments: the one who deliberately performs the action or the agent (*parents*), the one who is the recipient or the experiencer of the event (*me*), and the one that undergoes the action or the patient/theme of the action (*name*). The analysis of semantic roles (both of the arguments and adjuncts) is important both within theoretical linguistics and within applied linguistics in compiling semantic lexicons and valency dictionaries. From the point of view of language technologies, the task of semantic role labeling is important within the development of the information extraction systems, question answering systems, improving syntactic parsing systems, in machine translation tasks etc. (Shen in Lapata, 2007; Christensen et al., 2011). In comparison with syntactic trees, semantic role labeling requires higher level of abstraction, and it is a very important step towards the understanding of the meaning of a sentence. This is why SRL plays a major role in natural language processing. For instance, in the sentence *A weird name was given to me by my parents*, the morphosyntactic representation of the sentence is different than in the sentence mentioned earlier. However, semantic roles are the same in both sentences.

A comprehensive comparative analysis performed within META-NET white book series (Krek et al., 2012) has shown that both Slovene and Croatian may be considered as under-resourced languages in terms of language technologies, especially in the area of machine readable semantic resources and advanced tools for the processing of those resources.

Therefore, SRL will improve the existing levels of linguistic annotation of both Slovene and Croatian training corpora. With close cognate[1] languages it is advisable and beneficial to use similar principles and annotation schemes in the same natural language processing tasks.

Therefore, a project *Semantic Role Labeling in Slovene and Croatian* was conducted. The aim of the project was to build a semantic role labeling system which will be added

[1] Both languages in question belong to South Slavic branch of Slavic language family.

to the existing syntactic dependencies in both Slovene and Croatian training corpora used hitherto for machine learning algorithms. The core project tasks included: 1) development of the common Slovene-Croatian semantic annotation scheme and the creation of the list of semantic role labels based on the existing resources for other languages; 2) compiling the instructions for annotation; 3) manual annotation of the sample parts of both learning corpora using compatible tags. This served as the basis for the automatic annotation experiments using supervised machine learning methods, performed later on both corpora.

In the paper, we will present the resulting semantic role labeling framework in detail. The framework follows the path of similar previous SRL endeavours such as PDT, Vallex, FrameNET, Propbank, Crovallex etc. (see Krek et al., 2016). The paper describes both corpora used for semantic role annotation, as well as tools used in manual annotation tasks. Special attention is directed towards the description of the data obtained from the experimental automatic semantic role labeling based on supervised machine learning methods, and to its possible improvements. A preliminary quantitative analysis is performed for both languages (in terms of verbs range and frequencies, semantic roles, and typical syntactic-semantic patterns for the most frequent verbs).

2. Semantic Role Labeling framework for Slovene and Croatian

In compiling the list of semantic roles and their respective formal descriptions, we follow the approach developed by Prague Dependency Treebank, PDT (Mikulová et al., 2005), in which verbs or predicates determine arguments and adjuncts (usually specifying circumstances: time, location etc.). In addition, multi- word predicate role can be specified. (Table 1).

In the framework which was developed for the annotation of the Slovene and Croatian corpus, in addition to PDT, we have consulted Valency Lexicon of Czech Verbs (Vallex), semantic role labeling within Croatian Dependency Treebank (SRL tagset compiled by Filko et al. 2012), and Crovallex (Croatian version of Czech Vallex) which contains 1740 verbs selected from the Croatian frequency dictionary (Mikelić Preradović et al., 2009).

Our final SRL tagset (Table 1) contains 25 semantic labels (5 of those are arguments, 17 adjuncts, and 3 labels for multi-word predicates). The concept of obligatoriness or "coreness" was not used in the framework as compatible semantic resources (e.g. valency lexicons or FrameNets with a defined concept of obligatoriness) for both languages are not available at the moment.

SLO/CRO		**PDT**	
agent	ACT	actor	ACT
patient	PAT	patient	PAT
recipient	REC	addressee	ADDR
		benefactor	BEN
origin	ORIG	origo	ORIG
		inheritence	HER
result	RESLT	effect	EFF
location	LOC	direction	DIR2
		locative	LOC
source (location)	SOURCE	direction	DIR1
goal (location)	GOAL	direction	DIR3
event	EVENT		
time	TIME	temporal	TWHEN
		temporal	TPAR
		temporal	TFRWH
		temporal	TOWH
duration	DUR	temporal	TFHL
		temporal	THL
		temporal	TSIN
		temporal	TTILL
frequency	FREQ	temporal	THO
aim	AIM	aim	AIM
		intent	INTT
cause	CAUSE	cause	CAUS
contradiction	CONTR	contradiction	CONTRD
		concession	CNCS
condition	COND	condition	COND
regard	REG	regard	REG
		criterion	CRIT
		comparison	CPR
accompaniment	ACMP	accompaniment	ACMP
restriction	RESTR	restriction	RESTR
manner	MANN	manner	MANN
		result	RESL
means	MEANS	means	MEANS
quantification	QUANT	difference	DIFF
		extent	EXT
multi-word predicate	MWPRED		
modal	MODAL		
phraseological unit	PHRAS	dependant part of phraseme	DPHR

Table 1: SRL Tagset in SLO/CRO in comparison with PDT system.

3. Corpora and Tools for Annotation

On the Slovene side, the SSJ500k 2.0 (Krek et al., 2015) corpus was used for manual annotation of semantic roles. The corpus contains 500,293 words (27,829 sentences) sampled from the FidaPLUS corpus (Arhar Holdt and Gorjanc, 2007). The whole corpus is manually annotated on morphosyntactic level (Grčar et al., 2012), and partly on syntactic level (Dobrovoljc et al., 2012). Named entities and multi-word expressions are also identified (Gantar et al., 2017). The total of 5,491 sentences were annotated with semantic roles, with the first 500 sentences used for test annotation by four annotators. The second phase included automatic annotation (see Chapt. 3.1) of the remaining 4,991 sentences and their manual check by 5 annotators. These represent the basis for the quantitative analysis of the Slovene training corpus.

For the Croatian language, we used the SETimes.HR part of the hr500k corpus (Ljubešić et al., 2018), which is based on a sample of the Croatian part of the SETimes parallel corpus. It contains 3,757 sentences manually lemmatized and morphosyntactically tagged (Agić et al., 2013), and annotated for syntactic dependencies using the Universal Dependencies formalism (Agić and Ljubešić, 2015). Within this project, these sentences were being manually semantically annotated by 2 annotators. This then served as the resource for automatic labeling and quantitative analysis.

3.1. Automatic Semantic Role Labeling

Both annotated corpora were split in training and test data in a 80:20 fashion. This data split is available for each of the languages at https://github.com/clarinsi/bilateral-srl/tree/master/data.

Publishing the specific data split publicly has the goal of fostering comparing various tools on both languages and identifying that or those that perform best, or with the minimum memory and time footprint.

Currently the well-known baseline mate-tools semantic role labeler (Björkelund et al. 2009) was benchmarked on the data with the per-label F1 metric reported in Table 2. The weighted F1 score for all classes for Croatian was 0.72, while for Slovene it was 0.75. The parser is available from https://storage.googleapis.com/google-code-archive-downloads/v2/code.google.com/mate-tools/srl-4.31.tgz, and it was used without any modifications, using the German feature set.

Label	Croatian	Slovene
PAT	0.81	0.88
ACT	0.91	0.94
RESLT	0.83	0.80
TIME	0.65	0.62
REC	0.78	0.74
MODAL	0.94	0.90
MANN	0.45	0.76
LOC	0.56	0.59
DUR	0.64	0.50
ORIG	0.65	0.24
CAUSE	0.14	0.35
REG	0.43	0.34
AIM	0.47	0.20
GOAL	0.38	0.53
QUANT	0.54	0.62
MWPRED	0.72	0.91
EVENT	0.68	0.29
ACMP	0.80	0.08
MEANS	0.44	0.64
FREQ	0.50	0.59
CONTR	0.21	0.14
COND	0.59	0.46
PHRAS	0.11	0.31
SOURCE	0.29	0.37
REST	0.0	0.0

Table 2: Results (F1) of the experiments on automatic labeling of Croatian and Slovene with mate-tools for each label.

The data on both languages are quite similar, with F1 metrics corresponding to the frequency of each phenomenon. More concretely, on the Croatian dataset, the Pearson correlation between frequency and F1 is 0.517 with a p-value of 0.008, while on the Slovene dataset the same correlation coefficient is 0.611 with a p-value of 0.001. We can conclude that both correlation coefficients are strong and statistically highly significant

4. Quantitative analyses

In the next chapters, the preliminary quantitative analysis of both corpora is presented from the point of view of verb frequencies, semantic roles, and syntactic-semantic patterns that are recognized in the corpus as being stable and typical for individual verbs (here only for the most frequent verbs).

4.1. Verbs representation in both corpora

The Slovene SRL-annotated corpus contains 15,988 verbal tokens with 1,953 lemmas. The percentage of verbal lemmas appearing only once in the corpus is 47,5.

The Croatian SRL-annotated corpus contains 12,605 verbal tokens with 1,094 lemmas. The percentage of verbal lemmas appearing only once in the corpus is 40.8.

As expected, most frequent in both corpora are verbs with broad meaning spectrum such as *biti* 'to be', *imeti/imati* 'to have', *dobiti* 'to get'; modal verbs: *morati* 'must', *moči/moći* 'can', *hoteti/htjeti* 'will', *želeti/željeti* 'want', and verbs of communication *reči/reći* 'to say', *povedati/kazati* 'to tell'. Significantly higher frequency of the verbs of communication in the Croatian corpus (*kazati, izjaviti, reći, priopćiti, navoditi* = 'to tell, say, state etc.') is the result of the fact that SETimes.HR corpus consists only of news texts.

The list of verb lemmas with the minimum frequency of 50 in Slovene and Croatian corpora are in Table 3.

SSJ500k		SETimes.HR	
biti	7203	**biti**	4969
imeti	333	**htjeti**	670
morati	178	**kazati**	276
iti	114	izjaviti	210
vedeti	95	**moći**	195
dobiti	83	**imati**	163
moči	83	**reći**	160
začeti	80	trebati	146
videti	75	**morati**	117
reči	74	**željeti**	65
priti	72	očekivati	62
povedati	72	**dobiti**	57
hoteti	69	**postati**	57
želeti	59	postojati	56
postati	54	priopćiti	54
govoriti	51	predstavljati	53
misliti	50	navoditi	50

Table 3: Verbs with frequency f>=50 in SSJ500k and SETimes.HR. Verbs that are present in both corpora are indicated in bold.

Further qualitative analysis included the most frequent verbs (Table 3) and arguments (Figure 1). In case of arguments, we considered their presence in various patterns and their frequency in patterns. Individual verbs were taken as the basis for pattern formulation, however, polysemy (in case of polysemous verbs) was not taken into account. The reason for this is non-existence of compatible valency lexicons in Slovene and Croatian, and the size of the annotated corpora which cover only a limited set of senses per verb. This will be addressed in future versions of SRL annotations when this type of resources will be available.

4.2. Semantic roles representation in both corpora

All 25 semantic labels proposed in our framework are found in both training corpora. As can be observed from the Figure 1, the most frequent semantic roles in both corpora are argument roles of PAT, ACT and RESLT. They are followed by adjunct roles of TIME, MANN, and LOC (the last two being significantly more frequent in the Slovene corpus). In addition to these, other notable differences include significantly higher frequency of patients (PAT) and recipients (REC) in the Slovene corpus. On the other hand, the frequency of agent roles is extremely balanced in both corpora.

A more detailed comparative analysis could explain weather these differences are the result of differences in the corpora design (the Slovene and the Croatian corpora differ in genre representation - the Croatian one containing primarily news texts while the Slovene one being balanced in terms of genre representation). However, different genre representation in corpora certainly has the effect on the higher frequency of communication semantic group of verbs in the Croatian data.

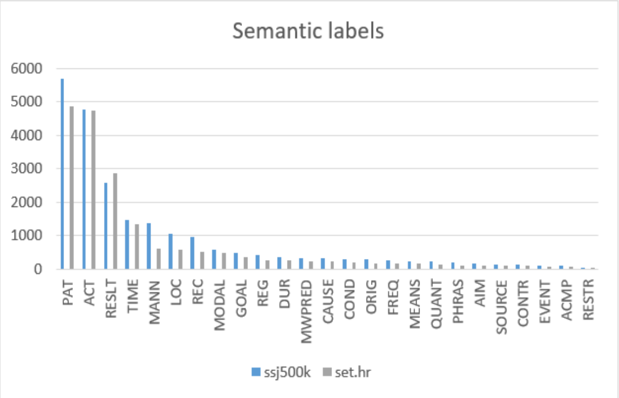


Figure 1: Semantic roles (labels) in Slovene (SSJ500k) and Croatian (SETimes.HR) training corpus.

Fquency of verbs in both corpora is relevant in relation to frequency of arguments in their patterns. Semantic roles with 50 or more hits in patterns are similar in both languages in case of verbs with similar basic meaning (in Table 4 and 5 indicated in bold).

biti	ACT	PAT	RESLT	TIME	MANN
imeti	ACT	PAT			
iti	ACT				
dobiti		PAT			
videti		PAT			
vedeti			RESLT		
postati			RESLT		

Table 4: Most frequent label (f>=50) per verb in Slovene SSJ500k

biti	ACT	PAT	RESLT	TIME
kazati	ACT		RESLT	
izjaviti	ACT		RESLT	TIME
reći	ACT		RESLT	
moći	ACT			
trebati	ACT			
imati	ACT	PAT		
uključivati		PAT		
predstavljati		PAT		
dobiti		PAT		
postati			RESLT	
priopćiti			RESLT	

Table 5: Most frequent label (f>=50) per verb in SETimes.HR

A verb *to be*, due to its broad semantics, is able to take on all of the semantic roles in both languages. Among the most frequent verbs, there are a few other such verbs with obligatory semantic roles (arguments): *imeti/imati* 'to have' (WHO has WHAT), *dobiti* 'to get' (WHO gets WHAT), and *postati* 'become' (WHO becomes WHO/WHAT).

4.3. Syntactic-semantic patterns

From both corpora, we have extracted stable syntactic-semantic patterns characteristic for each individual verb. Those patterns are similar in both languages despite the differences in the corpus design. Here, we will list those patterns (together with the example of their exact linguistic realization from the corpus) for the most frequent verbs in both corpora. To make the formalizations of these patterns more readable, we use "Who did What to Whom, and How, When and Where?" form (ACT = Who, PAT = What, RESLT=Who/What, LOC = Where etc.). Semantic tags are being put in the brackets next to their respective pronouns. The first part of the pattern represents its stable section which includes arguments that are typical for the given verb. In relation to (non-)obligatory nature of arguments, it has to be mentioned that patterns do not include arguments that are "obligatory" but are not explicitly present, e.g. agents (ACT) included in finite verbal forms, as exemplified in case of verbs *vedeti, začeti, videti, reči* etc. Since verb *biti* (to be) is found in combination with all arguments, this pattern was omitted in the analysis of both corpora.

As is the case with PropBank, our framework is also, at this stage, more focused on literal meaning and we did not clearly mark metaphorical usages.

4.3.1 SSJ500k

Slovene training corpus contains relatively stable patterns in case of verbs *imeti, morati, iti, vedeti, dobiti, moči* etc. (*potrebno/treba je* are also in this category) which appear in the corpus more than 70 times:

'to have' *imeti* (333)

- WHO (ACT) has WHAT (PAT 316) [for WHOM (REC), from whom (ORIG), where (LOC), when (TIME) ...]: *Na zadnji hrbtni bodici ima veliko črno piko.*

'must' *morati* (178)

- WHO (ACT) must INF (MODAL): *Država bi morala plačati stroške presoje vplivov na okolje.*

'to go' *iti* (114)

- WHO (ACT) goes WHERE (GOAL) [how

(MANN), when (TIME), under what conditions (COND) …]: *Šel sem prvič k vedeževalki.*

- to go (PHRAS 11): *Zgodba mi ni in ni šla iz glave.*
- to go SUPINE (MWPRED): *Verjetno bom šla smučat na Krvavec.*

'to know' ***vedeti*** (95)

- to know WHAT (RESLT) [how (MANN) …]: *Je pa treba nekaj jasno vedeti.*

'to get' ***dobiti*** (83)

- WHO (ACT) gets WHAT (PAT) [from whom (ORIG), in regard to (REG), with what (MEANS), when (TIME), under what conditions (COND) …]: *Mala je dobila ime po Prometeju.*

'can' ***moči*** (83)

- WHO (ACT) can INF (MODAL): *Ne moremo ga spregledati.*

'to start' ***začeti*** (80)

- WHO (ACT) starts WHAT (PAT) [how often (FREQ), when (TIME) …]: *Razpravo o tem je začel parlamentarni odbor.*
- to start INF (MWPRED): *Najprej začne pripravljati sladice.*

'to see' ***videti*** (75)

- WHO (ACT) sees WHAT (PAT) [when (TIME), in regard to what (REG), from where (SOURCE) …]: *Tukaj so jo zadnjič videli.*

'to say' ***reči*** (74)

- WHO (ACT) says TO WHOM (REC) WHAT (RESLT) [when (TIME) …]: *Neka psihologinja mi je rekla, da moram živeti le zase.*

'to come' ***priti*** (72)

- WHO (ACT) comes [to what (RESLT), when (TIME), where (GOAL), by what means (MEANS), why (CAUSE), under what conditions (COND) …]: *Na kongres v Ljubljano je prišlo več kot 500 gostov.*
- to come (PHRAS): *Vse to je prišlo na dan.*

Also, verb *iti* needs to be explained, with its pattern with the prepositional phrase *gre za* + WHO/WHAT (ENG: it is about). In this case the semantic role chosen for the argument expressed with WHO/WHAT was agent and not patient: *gre za vprašanje/preteklost/rešitev* etc. (ACT) (ENG: it's about the question/past/solution). If the verb in the same pattern is used for expressing motion, e. g. *iti v Evropo/samostan/desno* (GOAL) (ENG: to go to Europe/monastery/the right) the agent is not necessarily present. The verb *iti* and its counterpart *priti* are also somewhat special in the sense that they form phraseological units such as *ne iti v račun* ('not being able to comprehend'), *ne iti iz glave* ('not being able to forget'), *iti na živce* ('to make nervous'), *priti v poštev* ('to (be able to) be considered') with the label PHRAS.

4.3.2. SETimes.HR

From the Croatian training corpus, we have recognized and extracted fixed and stable syntactic-semantic patterns in case of verbs that appear in the corpus more than 50 times (*htjeti, kazati, moći, imati, trebati* etc.).

'to want' ***htjeti*** (670), ***željeti*** (65)

- WHO (ACT) wants WHAT (PAT) [for WHOM (REC), from WHOM (ORIG)...]: *Oni žele autonomiju sjevera, a za druge enklave žele takozvani Ahtisaari plus.*
- WHO (ACT) wants INF (MODAL): (WHAT) (PAT): *Mnoge žrtve ne žele podnijeti tužbu.*

'to tell, say' ***kazati*** (276), ***izjaviti*** (210), ***reći*** (160)

- WHO (ACT) says WHAT (RESLT) to WHOM (REC) about WHAT (PAT) [WHERE (LOC), WHEN (TIME)]: "*U suprotnom ćemo biti neozbiljni političari", rekao je Lagumdžija novinarima u Beogradu nakon sastanka s Jeremićem 14. ožujka.*

'can' ***moći*** (195)

- WHO (ACT) can INF (MODAL) WHAT (PAT): *Privatizacija je mogla donijeti bolje usluge.*

'to have' ***imati*** (163)

- WHO (ACT) has WHAT (PAT) [WHEN (TIME) for WHOM (REC), from WHOM/WHAT (ORIG)...]: *Moldavija sada ima novog predsjednika.*
- imati u vidu (PHRAS): *Imajući u vidu nesuradnju Beograda s Haaškim tribunalom ...*
- [(WHO) (ACT)] imati za cilj (PHRAS) WHAT (PAT): *Reforme za cilj imaju stavljanje oružja pod nadzor.*

'to need' ***trebati*** (146)

- WHO (ACT) needs INF (MODAL) WHAT (PAT) [to WHOM (REC)]: *Mi trebamo dati potporu Jeremiću.*

'must' ***morati*** (117)

- WHO (ACT) must INF (MODAL): *Čelnici moraju voditi.*

'to expect' ***očekivati*** (62)

- WHO (ACT) expects WHAT (PAT): *Katastarski dužnosnici očekuju registraciju oko 6,7 milijuna katastarskih čestica.*

'to get' ***dobiti*** (57)

- WHO (ACT) gets WHAT (PAT): *Manjinske zaklade dobit će naknadu za imovinu.*

5. Summary and Conclusions

In the paper, the data obtained from the experimental automatic semantic role labeling based on supervised machine learning methods, and the preliminary quantitative analyses of Slovene and Croatian training corpora (in terms of verbs range and frequencies, semantic roles, and typical syntactic-semantic patterns for the most frequent verbs) are presented.

The data for both languages are quite similar from all the above perspectives, despite the differences in corpora design.

From the preliminary analysis of the data, it seems that the SRL framework that was being developed within this bilateral project is suitable for semantic role labeling tasks in both languages. Moreover, the framework has been successfully implemented to serve as the solid base for the automatic SRL (using supervised machine learning methods).

Having a common framework for semantic annotation of cognate languages (Slovene and Croatian) was proved to be advantageous in terms of saving time and resources. Moreover, developing and applying a common framework was very beneficial from the perspective of mutual evaluation and corrections as well. This framework is also a solid base for future more detailed comparative semantic analyses.

Building a corpus with SRL annotations is an ongoing work and both corpora will be upgraded in the future. Upgrades will include the increase in size, calculation of inter-annotator agreement and segmentation of patterns according verb senses (when compatible semantic resources for both languages are available).

6. Acknowledgments

The paper was prepared as part of two projects, Nova slovnica sodobne standardne slovenščine: viri in metode (New grammar of contemporary standard Slovene: sources and methods, ARRS J6-8256) and by Slovene-Croatian bilateral government program. We would like to thank annotators of both corpora (in alphabetical order): Lucija Jezeršek, Taja Kuzman, Dafne Marko, Ivan Pandžić, Iza Škrjanec and Anja Zajc.

7. References

Špela Arhar Holdt and Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52(2), 95–110.

Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013. Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, pages 48–57. Sofia, Bulgaria, Association for Computational Linguistics.

Željko Agić and Nikola Ljubešić. 2015. Universal Dependencies for Croatian (that work for Serbian, too). In *Proceedings of the Workshop on Balto-Slavic Natural Language Processing*, pages 1–8.

Collin F. Backer, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*, pages 86–90. Montreal, Canada.

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)*, Boulder, June 4–5, pages 43–48.

Christensen, Janara, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An Analysis of Open Information Extraction based on Semantic Role Labeling. *International Conference on Knowledge Capture* (KCAP), pages 113–120. Banff, Alberta, Canada.

Kaja Dobrovoljc, Simon Krek, and Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. In *Zbornik Osme konference Jezikovne tehnologije*, pages 42–47. Ljubljana, Institut Jožef Stefan.

Matea Filko, Daša Farkaš, and Danijela Merkler. 2012. SRL Tagset for Croatian. Institute of Linguistics, Faculty of Humanities and Social Sciences, Zagreb. http://hobs.ffzg.hr/static/docs/SRL_tagset.pdf.

Polona Gantar, Simon Krek, and Taja Kuzman. 2017. Verbal multiword expressions in Slovene. *Europhras 2017*, pages 247–259. Springer.

Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. *Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik*. In *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana, Institut Jožef Stefan.

Simon Krek. 2012. *Slovenski jezik v digitalni dobi*. Berlin, Heilderberg, Springer Verlag.

Simon Krek, Polona Gantar, Kaja Dobrovoljc, and Iza Škrjanec. 2016. Označevanje udeleženskih vlog v učnem korpusu za slovenščino. In *Proceedings of the Conference on Language Technologies & Digital Humanities*, Faculty of Arts, pages 106–110. University of Ljubljana.

Krek, Simon et al. 2015. Training corpus ssj500k 1.4, *Slovenian language resource repository* CLARIN.SI, http://hdl.handle.net/11356/1052.

Nives Mikelić Preradović, Damir Boras, and Sanja Kišiček. 2009. CROVALLEX: Croatian Verb Valence Lexicon. In *Proceedings of the 31st International Conference on Information Technology Interfaces*, pages 533–538.

Marie Mikulová et al. 2006. Annotation on the tectogrammatical level in the Prague Dependency Treebank. *Annotation manual. Technical Report 30*, pages 5–11.

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

Dan Shen and Mirella Lapata. 2007. Using Semantic Roles to Improve Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning*, pages 12–21. Prague.

Zbirka primerov rabe vejice Vejica 1.3

Peter Holozan

Amebis, d. o. o., Kamnik
Bakovnik 3, 1241 Kamnik
peter.holozan@amebis.si

Povzetek

Pripravljena je bila nova verzija zbirke primerov rabe vejice Vejica, v kateri so bile popravljene najdene napake iz prejšnje verzije 1.0, dodan pa je bil del z nestandardnimi tviti iz korpusa Janes-Vejica 1.0. Nova verzija je izrazito izboljšala rezultat popravljanja vejic slovničnega pregledovalnika Besana pri delu iz korpusa Lektor. Slovnični pregledovalnik Besana je bil dopolnjen za pregledovanje nestandardnih besedil.

Corpus of comma usage Vejica 1.3

New version of corpus if comma usage Vejica was prepared, correcting the found problems from previous version 1.0. New part wa added containing samples of non-standard twits from corpus Janes-Vejica 1.0. The new version of Vejica improved results for Besana grammar checker comma correcting significantly in the part from corpus Lektor. Besana grammar checker was improved for better handling of non-standard texts.

1. Uvod

Jezikovnotehnološko raziskovanje kot osnovo potrebuje ustrezno označene korpuse oz. primere, kar nam omogoča potem bodisi uporabo strojnega učenja bodisi preizkušanje različnih metod.

Za področje postavljanje vejic je bilo že pripravljenih nekaj zbirk podatkov, vendar imajo obstoječe zbirke nekatere slabosti, zato je smiselno to področje še dopolniti z novo zbirko primerov rabe vejice, ki bo omogočala nadaljnje delo na tem področju.

2. Pregled dosedanjih zbirk primerov

Obstaja že kar nekaj prosto dostopnih korpusov oz. zbirk primerov rabe vejice v slovenščini, ki imajo označene napačno postavljene (torej manjkajoče in odvečne) vejice.

2.1. Korpus Šolar

V korpusu šolskih pisnih izdelkov Šolar so besedila, ki so jih učenci v slovenskih osnovnih in srednjih šolah samostojno tvorili pri pouku. Zajeta so besedila, pri katerih je slovenščina materni jezik avtorjev in ki niso bila napisana posebej za projekt, ampak kot šolska produkcija, jezikovni popravki so pa taki, kot so jih naredili učitelji. (Rozman et al., 2010)

Vključena so besedila od 6. do 9. razreda osnovnih šol, od 1. do 5. letnika srednjih šol in besedila, ki so bila napisana na maturitetnih tečajih. Del korpusa so tudi popravki, ki so jih naredili učitelji.

Prvič je korpus Šolar kot zbirko primerov rabe vejice uporabil Holozan (2012), vendar so bile uporabljene le povedi, ki so imele označene kakšno napako pri vejicah (torej bodisi manjkajočo bodisi odvečno vejico), kar pa ni primerno za ugotavljanje natančnosti. Nekoliko dopolnjena verzija (popravljene se bile nekatere opažene napake pri popravkih vejic) je bila uporabljena še pri preizkusu strojnega učenja za postavljanje vejic (Holozan, 2013).

Korpus je dostopen na naslovu https://www.korpus-solar.net/ in v repozitoriju CLARIN.SI s povezavo http://hdl.handle.net/11356/1036.

2.2. Korpus Lektor

Korpus lektorskih popravkov Lektor je nastal v okviru doktorske naloge (Popič, 2014), vsebuje približno milijon besed. Besedila v korpusu so napisali profesionalni pisci, dobra polovica so prevodi. Baza je v formatu XML in označeni so vsi lektorski popravki.

Korpus je dostopen na naslovu http://www.korpus-lektor.net.

Korpus je bil za analizo vejic uporabljen v Piškur (2015).

2.3. Vejica 1.0

Zbirka primerov rabe vejice Vejica 1.0 je bila pripravljena v okviru doktorske naloge (Holozan, 2016). Sestavljena je iz štirih delov, dva dela imata še poddele (kar je podrobneje opisano v nadaljevanju).

Zbirka primerov vsebuje 113.308 povedi, pri čemer je označenih 17.768 (11,36 %) manjkajočih vejic, 4.608 (3,22 %) vejic pa je označenih za odvečne. (Holozan, 2016)

Zbirka primerov je dostopna v repozitoriju CLARIN.SI s povezavo http://hdl.handle.net/11356/1055 pod licenco CC BY-NC-SA 4.0.

2.3.1. Korpus KUST

Korpus usvajanja slovenščine kot tujega jezika (KUST) je zbirka besedil, ki so jih napisali govorci drugih jezikov, ki se šele učijo slovensko. Korpus je bil predlagan v Stritar (2006) in bil narejen v okviru projekta ESS Uspešno vključevanje otrok, učencev in dijakov migrantov v vzgojo in izobraževanje. Projekt je izvajal Center za slovenščino kot drugi/tuji jezik Filozofske fakultete Univerze v Ljubljani (Rozman et al., 2010). Vključen je bil le manjši del korpusa KUST, in sicer je bilo izbranih 388 povedi, v katerih je bila vsaj ena odvečna ali manjkajoča vejica, pravilne povedi niso vključene (Holozan, 2016).

2.3.2. Korpus Šolar

V poddelu iz korpusa Šolar so bila uporabljena besedila iz korpusa Šolar (bolj podrobno opisanega v točki 2.1)., ki so vsebovala vsaj en učiteljski popravek (ker vsa besedila niso vsebovala popravkov). Ker pa se je pokazalo, da je veliko napak pri vejicah neoznačenih (za kar je možno več razlogov: če je besedilo vsebovalo preveč napak, je učitelj lahko obupal in nehal popravljati, učitelj je lahko popravil poved drugače in vejica potem ni bila več potrebna, kakšne vejice pa so učitelji najbrž tudi spregledali). Zato so bili vsi ti primeri ročno pregledani in popravljeni, skupaj skoraj 50.000 povedi. (Holozan, 2016)

Primere je popravljala ena oseba (s konzultacijami s strokovnjaki pri problematičnih primerih), kar pomeni, da so vejice čim bolj enotno postavljene, kar lahko pomeni prednost pri računalniški obdelavi, predvsem pri strojnem učenju, težave zaradi neenotne uporabe vejic v uporabljenem korpusu so npr. imeli pri strojnem učenju postavljanja vejic v baskovščini (Alegría et al., 2006).

Primeri iz tega dela imajo označen tudi poddel, ki pove, v katerem razredu oziroma letniku je bil napisan posamezen primer.

Ta poddel je uporabljen v Krajnc (2015) in Krajnc in Robnik-Šikonja (2015).

2.3.3. Korpus Lektor

Primeri iz korpusa Lektor (podrobneje opisanega v točki 2.2) so bili pretvorjeni tako, da so bili označeni le lektorski popravki pri vejicah, vsi drugi lektorski popravki pa so bili izpuščeni. Izpuščene so bile tudi povedi, ki so v celoti napisane v tujih jezikih, vendar to ni bilo narejeno čisto dosledno.

Celoten del je ena enota, posamična besedila iz korpusa Lektor niso bila razporejena kot morebitni poddeli.

2.3.4. Wikipedija

Zadnji del predstavljajo članki iz Wikipedije, pri čemer so bili izbrani članki, ki niso preveč lektorirani, za kar je bila uporabljena kategorija »Članki, ki so potrebni čiščenja«. Vse manjkajoče in odvečne vejice so bile ročno označene. Uporabljenih je bilo 9 člankov, ki skupaj vsebujejo 870 povedi. (Holozan, 2016)

Namen tega dela je bil dobiti domeno, ki bi bila med šolarji iz korpusa Šolar in profesionalnimi pisci iz korpusa Lektor.

2.4. Janes–Vejica 1.0

Ta zbirka primerov je bila narejena iz 500 tvitov (čivkov) v nestandardni slovenščini (izbrani so bili le tviti, ki jim je bila pripisana jezikovna nestandardnost), zbranih v okviru projekta Janes. V njih so bili ročno označeni vsi primeri nestandardne rabe vejic, vsa ta mesta in tudi vse obstoječe vejice pa so bile označene z razlogom za vejico. Primere sta označevala dva označevalca, na koncu pa je v primeru neskladja kuratorka določila končno oznako. (Popič et al., 2016).

Zaradi označenih razlogov za vejice je ta zbirka primerov zelo zanimiva za nadaljnje analize, žal pa ni velika. Zbirka primerov je dostopna v repozitoriju CLARIN.SI pod licenco CC BY-SA 4.0 na naslovu http://hdl.handle.net/11356/1088.

3. Dopolnjevanje zbirke rabe vejice

Po izdelavi je bila zbirka primerov Vejica 1.0 uporabljena za preizkušanje slovničnega pregledovalnika Besana. Pri tem se je pokazalo, da je v Vejica 1.0 kar nekaj težav, ki jih je bilo treba rešiti. Vmes sta bili narejeni zbirki Vejica 1.1 in Vejica 1.2, ki sicer nista bili objavljeni, je pa Vejica 1.2 bila omenjena pri zagovoru doktorata Holozan (2016), po tem pa je bilo narejeno še kar nekaj novih popravkov, zato sem se odločil, da bo nova izdana zbirka označena kot Vejica 1.3.

3.1. Izločanje tujih povedi

Kot prva težava se je pokazalo, da je v delu iz korpusa Lektor še nekaj deset povedi, ki so v celoti v tujem jeziku.

Harlem est né a Lausanne.
Omar est né a Evian.
Les gens nés a Lausanne sont généralement sujets helvétiques.

Slika 1: Primeri izločenih tujih povedi.

V prvih dveh primerih je analizator opozarjal na vejico pred »a«, ker ni zaznal, da gre za tuji jezik, ker so besede Harlem, ne, a, Omar in Evian tudi v slovenskem slovarju. Tu bi se sicer morda dalo še kaj nadgraditi s tem, da bi lastna imena izvzeli iz ugotavljanja jezika, vendar je pri kratkih povedih ugotavljanje jezika vseeno lahko problematično, še najbolj zanesljiva rešitev bi bila uporaba slovarjev in analizatorjev za vse te druge jezike, kar pa je v praksi sitno tako zaradi počasnosti kot potrebnega prostora za tuje slovarje (in npr. v Amebisu bi lahko tako dobro zaznavali le angleščino in nemščino, delno tudi francoščino in albanščino, za hrvaščino imamo le zelo majhen testni slovar, drugih jezikov pa niti nismo pokrivali).

V realni uporabi je pri povedih v tujem jeziku tudi sicer smiselno, da se pravilno nastavi jezik v urejevalniku besedil (že zaradi črkovalnika), kar pomeni, da slovenski slovnični pregledovalnik potem ne bo pregledoval teh povedi.

Ostali so še primeri, v katerih je v tujem jeziku le del, največkrat naslovi del ali pa citirani primeri. Ti primeri sicer delajo težave analizatorju, vendar so to čisto realna uporaba, ki jo bo moral slovenski slovnični pregledovalnik nekoč rešiti.

3.2. Pravilno upoštevanje lektorskih popravkov

Pregled problematičnih primerov je pokazal, da se je pri uvozu iz korpusa Lektor v nekaterih primerih (predvsem pri gnezdenih napakah, torej ko sta bili napaki ena v drugi) zgodilo, da so bili uvoženi tudi drugi lektorski popravki, in sicer tako, da je bilo napisano tako originalno besedilo kot popravek.

Kaj vi mislite menite in zakaj?
Razdelite izročite vsakemu po en flomaster/barvno pisalo.
kupujmo Kupujemo izdelke, narejene iz lokalnega lesa s FSC oznako

Slika 2: Nekaj primerov težav zaradi napačnega upoštevanja popravkov.

Še posebej v primerih, ko je bil popravljen glagol, je to analizatorju naredilo težave, ker je zaradi dvojnega glagola domneval, da nekje manjka vejica, kar je potem poslabševalo natančnost iskanja manjkajočih vejic.

3.3. Popravljene vejice

Pokazalo se je, da nekateri primeri manjkajočih oz. odvečnih vejic niso bili označeni v zbirki primerov. Največ takih primerov je bilo v delu iz korpusa Lektor, potem v delu iz Wikipedije, najmanj (po deležu) napak je bilo v delu iz korpusa Šolar, ki je bil ob izdelavi zbirke primerov Vejica 1.0 najbolj sistematično ročno preverjen (Holozan, 2016).

Izračuna se kot povprečje indeksiranih vrednosti BDP-ja na prebivalca, pričakovane življenske dobe§, in povprečnega pričakovanega obdobja obiskovanja izobraževalnih ustanov v državi.
Poleg zmanjšanja potrebne količine surovine za dano uporabo§, izboljšana učinkovitost zmanjšuje relativno ceno uporabe določene surovine, kar poveča povpraševanje po surovini, to pa lahko izniči prihranke, ki jih je prinesla povečana učinkovitost.
Določite§ kateri korporaciji pripadajo naslednje znamke:
Da bi si lažje zapomnili§ pišite na tablo: Različna gledišča: Kdo?

Slika 3: Primeri popravljenih vejic, označeni z znakom §.

Sistematično so bili za Vejico 1.3 preverjeni le primeri, pri katerih se z oznakami v zbirki primerov niso ujemali rezultati slovničnega pregledovalnika Besana. Zato se ob dopolnitvah Besane redno najdejo še novi taki primeri napak v zbirki primerov, kar kaže, da bi bilo za še natančnejše popravke treba še enkrat ročno preveriti celotno zbirko primerov. Po drugi strani pa glede na dosedanje število odkritih napak in delež napak, ki jih že popravi Besana, ocenjujem, da število še neodkritih najbrž ne presega enega odstotka sedanjega števila najdenih napak.

3.4. Izboljšanje zaradi popravkov zbirke primerov

Popravljanje zbirke primerov je zelo izboljšalo rezultate popravljanja vejic, dosežene s programom Besana, in sicer predvsem v delu iz korpusa Lektor, opazna izboljšava pa je bila tudi v delih iz korpusa KUST in Wikipedije, kjer so bile predvsem napake pri vejicah, ker sta bila ta dva dela označena ročno in nista bila potem prej še enkrat preverjena.

Za del iz korpusa Lektor se je priklic pri iskanju manjkajočih vejic izboljšal iz 40,5 % na 50,5 %, natančnost pa kar iz 18,8 % na 35,6 %. Pri iskanju odvečnih vejic se je pri delu iz korpusa Lektor priklic popravil iz 26,8 % na 40,5 %, natančnost pa iz 29,7 % na 58,4 %. Izboljšanje je torej res veliko. To izboljšanje kaže, da bo morda smiselno ponoviti poskus s strojnim učenjem postavljanja vejic, pri katerem je Holozan (2016) pokazal, da deluje veliko slabše na delu iz korpusa Lektor kot na delu iz korpusa Šolar. Po drugi strani pa je tam šlo za postavljanje vseh vejic, pri čemer vpliv popravkov najbrž ne bo tako velik, kot je za popravljanje vejic, kajti v delu iz korpusa Lektor je delež napak pri postavljanju vejic majhen – 1,2 % vejic manjka, odvečnih je pa 0,8 % (Holozan, 2016).

3.5. Priključitev zbirke Janes–Vejica 1.0

Odločil sem se, da v zbirko primerov vključim še primere iz korpusa Janes–Vejica 1.0, s čimer bo zbirka dopolnjena še s spletno slovenščino.

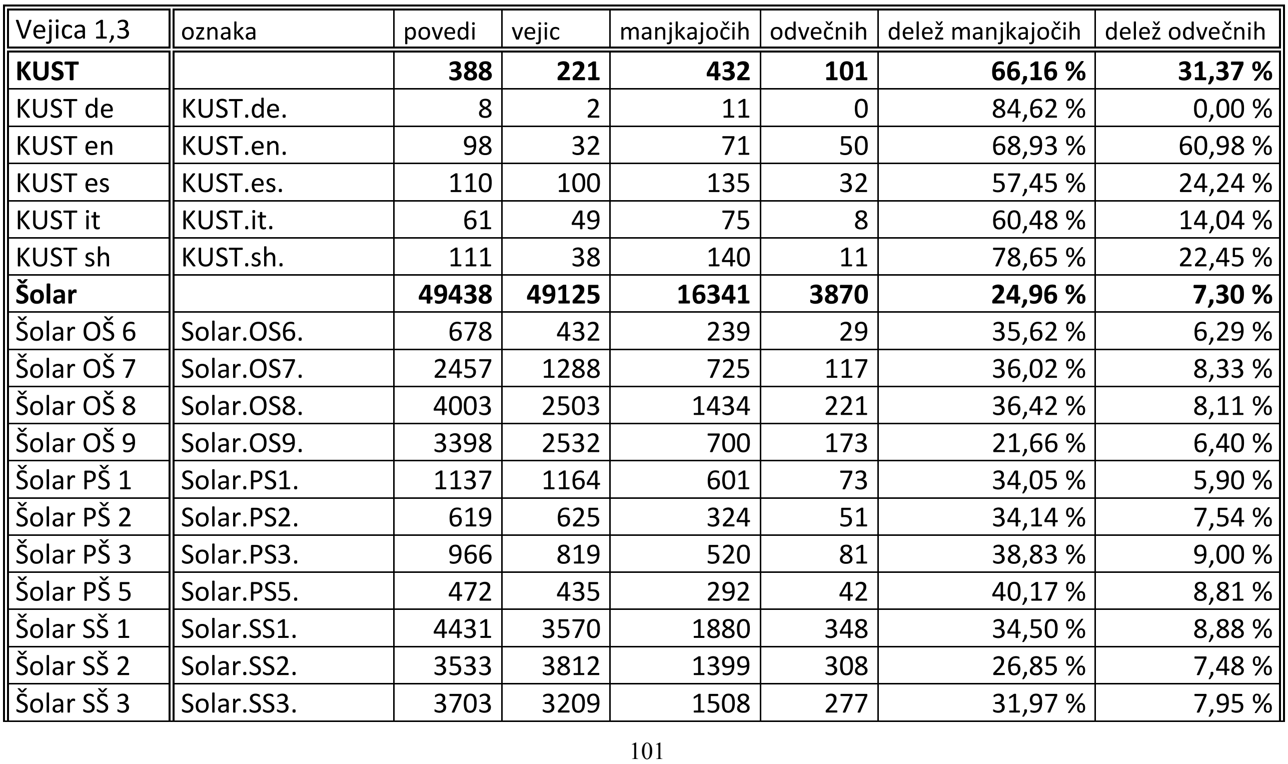
Raziskava Popič in Fišer (2018), ki je bila izvedena na tem korpusu, je pokazala, da so v večini primerov vejice postavljanje v skladu s standardom, nestandardne vejice pa so predvsem manjkajoče, odvečnih je malo.

Primeri so bili pretvorjeni tako, da so označeni enako kot drugi deli zbirke primerov (torej z oznakami za manjkajoče in odvečne vejice), odstranjeni so bili sicer zanimivi podatki o tipih vejic. Poenotenje teh primerov v format preostale zbirke primerov Vejica je koristno zato, da lahko nekdo na enak način uporabi zelo raznolike vire pri preizkušanju popravljanja vejic oz postavljanja vseh vejic besedilo.

Vsega skupaj je v tem delu 1369 povedi.

4. Sestava nove zbirke rabe vejice

Vejica 1.3 je sestavljena iz petih delov, KUST in Šolar pa sta še naprej razdeljena na poddele. Število primerov je prikazano v spodnji tabeli.



Vejica 1,3	oznaka	povedi	vejic	manjkajočih	odvečnih	delež manjkajočih	delež odvečnih
KUST		**388**	**221**	**432**	**101**	**66,16 %**	**31,37 %**
KUST de	KUST.de.	8	2	11	0	84,62 %	0,00 %
KUST en	KUST.en.	98	32	71	50	68,93 %	60,98 %
KUST es	KUST.es.	110	100	135	32	57,45 %	24,24 %
KUST it	KUST.it.	61	49	75	8	60,48 %	14,04 %
KUST sh	KUST.sh.	111	38	140	11	78,65 %	22,45 %
Šolar		**49438**	**49125**	**16341**	**3870**	**24,96 %**	**7,30 %**
Šolar OŠ 6	Solar.OS6.	678	432	239	29	35,62 %	6,29 %
Šolar OŠ 7	Solar.OS7.	2457	1288	725	117	36,02 %	8,33 %
Šolar OŠ 8	Solar.OS8.	4003	2503	1434	221	36,42 %	8,11 %
Šolar OŠ 9	Solar.OS9.	3398	2532	700	173	21,66 %	6,40 %
Šolar PŠ 1	Solar.PS1.	1137	1164	601	73	34,05 %	5,90 %
Šolar PŠ 2	Solar.PS2.	619	625	324	51	34,14 %	7,54 %
Šolar PŠ 3	Solar.PS3.	966	819	520	81	38,83 %	9,00 %
Šolar PŠ 5	Solar.PS5.	472	435	292	42	40,17 %	8,81 %
Šolar SŠ 1	Solar.SS1.	4431	3570	1880	348	34,50 %	8,88 %
Šolar SŠ 2	Solar.SS2.	3533	3812	1399	308	26,85 %	7,48 %
Šolar SŠ 3	Solar.SS3.	3703	3209	1508	277	31,97 %	7,95 %

Šolar SŠ 4	Solar.SS4.	2940	3174	1234	246	27,99 %	7,19 %
Šolar G 1	Solar.G1.	7240	8475	2150	775	20,24 %	8,38 %
Šolar G 2	Solar.G2.	3134	3918	816	221	17,24 %	5,34 %
Šolar G 3	Solar.G3.	3613	4391	841	304	16,07 %	6,47 %
Šolar G 4	Solar.G4.	6892	8593	1565	590	15,41 %	6,42 %
Šolar MT	Solar.MT.	222	185	113	14	37,92 %	7,04 %
Lektor	Lektor.Lektor.	**52121**	**71204**	**1133**	**717**	**1,57 %**	**1,00 %**
Wikipedija	Wiki.Wiki.	**869**	**929**	**124**	**71**	**11,78 %**	**7,10 %**
Janes	Janes.Janes.	**1368**	**646**	**387**	**20**	**37,46 %**	**3,00 %**
skupaj		104184	122125	18417	4779	13,10 %	3,77 %

Tabela 1: Sestava zbirke primerov Vejica 1.3.

Tabela 1 prikazuje zgradbo zbirke primerov Vejica 1.3 skupaj z dodatno razdelitvijo delov KUST in Šolar.

Enako kot v Holozan (2016) je delež manjkajočih vejic izračunan tako, da delimo število manjkajočih vejic z vsoto napisanih vejic in števila manjkajočih vejic (niso upoštevanje odvečne vejice), delež odvečnih vejic pa je izračunan kot kvocient števila odvečnih vejic z vsoto števila napisanih vejic in števila odvečnih vejic (niso upoštevanje manjkajoče vejice).

4.1. Format nove zbirke rabe vejice

Zbirka je zgrajena enako, kot je bil zgrajena zbirka primerov rabe vejice Vejica 1.0. Vsaka poved je v svoji vrstici, v vrstici je najprej oznaka povedi, ki je sestavljena iz oznake dela, oznake poddela (uporabljene oznake so naštete v Tabeli 1) in zaporedne številke v poddelu, deli oznake so med sabo ločeni s pikami. Potem sledi poved, ločena s tabulatorjem, mesta, na katerih vejice manjkajo, so označene z znakom »¤«, odvečne vejice pa so nadomeščene z znakom »÷«.

KUST.de.4 Ko sva prišli do skupine¤ so vsi vprašali¤ kaj smo kupili in midve tudi.
KUST.de.5 Danes¤ če gremo ven¤ bom oblekla mojo novo obleko, se že veselim.
Wiki.Wiki.487 Pogosto se v spastične mišice spodnjih okončin dajejo injekcije botulina÷ z namenom¤ da se zmanjša spastično povečan mišični tonus, ki je lahko zelo boleč.
Solar.G1.643 Na leto imamo približno 120 snežnih dni÷ ter 220 kondicijskih enot.
Solar.PS2.170 Odlomek govori o tem¤ kaku je David Goldstein tekel pred smogom¤ voznik avta¤ ki÷ je šel mimo¤ pa mu ni hotel ustaviti.
Lektor.Lektor.266 V drugih državah pa delovanje sindikatov režim zatira÷ ali pa dovoljuje zasebnim podjetjem¤ da ga zatirajo, tudi če v ta namen uporabljajo silo.
Janes.Janes.1182 je dost¤ da je topla, za čez šmorn glih kul.

Slika 4: Nekaj primerov iz zbirke.

Datoteka je zapisana v formatu UTF-8.

4.2. Dostopnost nove zbirke rabe vejice

Zbirka primerov rabe vejice Vejica 1.3 je objavljena v repozitoriju CLARIN.SI pod imenom Vejica 1.3 pod licenco CC BY-NC-SA 4.0 na naslovu http://hdl.handle.net/11356/1185.

5. Izboljševanje popravljanja vejic

Ker so podatki v Holozan (2016) že zastareli, je hkrati z novo zbirko primerov rabe vejice smiselno objaviti še kratko poročilo delu pri računalniškem popravljanju vejic s programom Besana in najnovejše rezultate, ki lahko služijo kot referenčne vrednosti za nadaljnje raziskave na tem področju.

5.1. Izboljšave slovničnega pregledovalnika Besana

Rezultati, objavljeni v Holozan (2016), so bili izboljšani že do zagovora konec leta 2016. Tako se je za problem iskanja manjkajočih vejic skupen priklic popravil iz 57,5 % na 63,5 %, natančnost pa iz 74,9 % na 79,8 %. Do tega trenutka se je Besana še popravila na priklic 75,3 % in natančnost 80,6 %. Vsi ti podatki so na zbirki primerov Vejica 1.0, kot pa kažejo ugotovitve v točki 3.4, bodo rezultati v novi verziji primerov še boljši.

Po eni strani je za izboljšanje Besane pomagalo povečanje slovarja (nove verzije prepoznajo veliko več lastnih imen), še bolj pa dopolnitve stavčnega analizatorja (oz. analizatorja povedi). Ena od pomembnejših dopolnitev za področje vejic je bila obravnava veznika »kot« v različnih skladenjskih vlogah. Vse izboljšave analizatorja so na primer označevanje jos100k izboljšale do te mere, da so zdaj leme pravilno označene v 99,31 %, oblikoskladenjske oznake pa v 96,87 %.

Izboljšano je bilo tudi opozarjanje na postavljanje vejic v primerih, pri kateri sta dve možnosti, prava pa je odvisna od pomena. Tak primer sta npr. »tako da« in »zato ker«, pri katerih je vejica lahko spredaj ali pa vmes, odvisno od poudarka. Prej je Besana vedno postavila vejico na začetku, kar je v tistih primerih, v katerih bi morala vejica biti vmes, pomenilo, da najprej ni ugotovila manjkajoče vejice, dodatno pa je še postavila vejico tja, kjer je ne bi smelo biti. Zato v teh primerih zdaj Besana na začetku opozori, da nekje v stavku manjka vejica, in prepusti uporabniku, da sam izbere, kje bi vejica morala biti.

5.2. Izboljšanje pri nestandardni slovenščini

Na prvi pogled se morda zdi, da dopolnjevanje stavčnega analizatorja za Besano z zelo nestandardno slovenščino ni preveč smiselno, saj takih besedil ne bo nihče popravljal z Besano. Vendar je kar nekaj razlogov, zaradi katerih je to smiselno.

Take nestandardne oblike se ljudem prikradejo tudi, kadar želijo pisati v knjižni slovenščini (ker marsikdo bere predvsem spletno slovenščino in vedno manj besedil v knjižni slovenščini), zato je smiselno, da jih Besana popravi.

Še pomembnejši razlog pa je, da je stavčni analizator uporabljen tudi drugje. Ena taka uporaba je na primer označevalnik korpusov in v korpusu Gigafida je kar precej nestandardnih besedil, ki so v tem trenutku nenatančno označena, ker je označevalnik naučen predvsem na besedila v knjižni slovenščini (je pa bil pri tem označevanju dosežen velik napredek v okviru projekta Janes, in sicer z vključitvijo nestandardnih učnih podatkov in z vključitvijo Brownovih gruč, pridobljenih iz velikih zbirk surovih nestandardnih podatkov (Ljubešić et al., 2018)). Tak primer je npr. beseda »jas« kot nestandardna oblika osebnega zaimka za prvo osebo ednine. Beseda »jas« ima v Gigafidi 1947 pojavitev, med prvimi 20 zadetki je 16 takih, kjer je mišljen osebni zaimek, vendar so vse označene kot oblika samostalnika »jasa«.

Druga uporaba je v sistemu za virtualne asistente SecondEgo, ki tako bolje prepoznavajo nestandardno slovenščino (pregled uporabe agentov v SecondEgu kaže, da je predvsem pogosta neuporaba strešic pri čšž in velikih začetnic).

Tretja pa je strojni prevajalnik Presis, ki pa je že lahko čisto realna uporaba tudi pri tvitih, še posebej zato, ker ima Google Prevajalnik precej težav z nestandardno slovenščino, kar je logično, ker paralelni korpusi, iz katerih se uči, vsebujejo večinoma le standardno slovenščino.

5.2.1. Nestandardno besedišče

Kako pogosto je nestandardno besedišče že v korpusu Gigafida, kaže npr. primer zapisa »nč« namesto »nič«, za kar je v Gigafidi kar 6410 konkordanc. V nekaj malega primerih gre sicer tudi za kratico »NČ« v pomenu »nedoločen čas«, še vedno pa je praktično 6000 stavkov, v katerih o analizator bolje deloval, ker bo to besedo pravilno prepoznal.

Primer neoznačene manjkajoče vejice v korpusu Šolar, ki ga je našla Besana po tem, ko je bila dodana beseda »kokr« kot nestandardni zapis besede »kakor«, je naslednji:

Pozno zvečer so uzeli najnujnejše stvari¤ pobrali§ kokr so hitro lahko¤ saj se jim je počas že istekal na barko.

Slika 5: Naslov slike.

Besana je dodatno opozorila, da manjka vejica na mestu z oznako §.

Primer nestandardno zapisane besede, ki dela težave pri postavljanju vejic, če je analizator ne prepozna, je npr. »poj« v pomenu »potem«, ki se prekriva z velelnikom glagola »peti«, zaradi česar je Besana pri »Poj bom pa vsakmu pokazal napis na seb, pa naj se mal zamisljo.« prej pred »bom« pričakovala vejico.

Primer nestandardne oblike, ki pa ni bila dodana, ker bi verjetno povzročila več težav, kot pa bi jih rešila, pa je »ja« kot zapis osebnega zaimka »jaz«, ker se prekriva z zelo pogostim medmetom (oz. členkom) »ja«, ki je zelo pogost ravno v nestandardnih besedilih. Če bi bil dodan kot osebni zaimek, Besana v veliko primerih ne bi dodala manjkajoče vejice za »ja« na začetku povedi, ker bi menila, da gre za osebni zaimek, ki je del nadaljevanja. Po drugi strani pa se je kot koristno izkazalo to, da je bil dodan nestandardni osebni zaimek »jas«, ki se prekriva z obliko samostalnika »jasa«.

5.2.2. Tipične redukcije pri pisanju

V nestandardni slovenščini obstajajo nekatere tipične redukcije pri pisanju, ki jih je smiselno upoštevati, kadar besede ni v slovarju.

mogoče bo kdo clo zastopu kdaj kk je brez šihta
haha zlobirite interrail pa pridte :)
Men prej pr stran grejo, k pa zadi.

Slika 6: Primeri redukcij pri pisanju.

V analizator je bilo dodano pravilo, da »u« na koncu besede tipično nadomešča »il«, »al« ali »el«, če gre za deležnik na -l. Podobno je pri »i« treba preveriti, ali je mogoče namesto »aj«.

Razlika med temi tipičnimi redukcijami in nestandardnim besediščem je v tem, da je nestandardno besedišče dodano v slovar, redukcije pa se upoštevajo sistemsko (je pa treba v slovar dodati oblike, ki se prekrivajo z drugimi standardnimi oblikami).

5.2.3. Neuporaba čšž

V tvitih (in tudi nasploh pri nestandardnem pisanju) je pogosto, da so izpuščene strešice na čšž in so torej namesto teh uporabljene črke csz. Amebisov analizator ima nastavitev, ki pri vseh csz upošteva tudi možnost, da gre za čšž. Težava pa je, ker pri tem lahko nastanejo tudi dvoumnosti, ki lahko otežijo analizo povedi in postavljanje vejic.

@SanjaModric kaj to pomeni za nase ce zvecer zmagajo ...

Slika 7: Primer tvita brez strešic.

Pri tipičnih veznikih (npr. »če«) je bilo treba dodati tudi obliko »ce« pri preverjanju mest, na katerih pogosto manjka vejica. Dodatno se je pri »ce« zapletlo še v kombinaciji z neuporabo velikih začetnic, ker se je pokazala možnost, da je »CE« še oznaka kraja Celje, kar je dodatno zapletlo delo analizatorja povedi.

5.2.4. Neuporaba velikih začetnic

Podobno kot strešice so v nestandardnih besedilih pogosto izpuščene tudi velike začetnice. Zato ima Amebisov analizator tudi nastavitev, da se ne zanaša, da so pravilno uporabljene velike začetnice (ta nastavitev je na primer vključena v virtualnih asistentih v sistemu SecondEgo). Seveda pa ta nastavitev zaplete razdvoumljanje.

ja želim ti čim več dobrih ocen..

Slika 8: Primer neuporabe velikih začetnic.

Besana običajno pri vejicah za medmeti na začetki povedi preveri, da mora biti medmet napisan z veliko začetnico, ta omejitev je bila ob nastavitvi, da začetnice niso pomembne, odstranjena, tako da Besana postavi vejico za »ja« v primeru zgoraj.

5.2.5. Vejice pri začetnih medmetih

Na tvite lahko gledamo kot na premi govor. Ena od značilnosti je veliko število uporabljenih medmetov, ki morajo na začetku povedi biti z vejico ločeni od nadaljevanja.

Bravo Risi!
Mah ne grem se vec!!
Aja hvala za sveče, grem dans na britof zastojn
ajde falil so do 13h ampak pol se je pa zacel #FB http://t.co/pnGkF3N46T

Slika 9: Primeri, pri katerih manjka vejica za začetnim medmetom.

Po eni strani je bilo treba dodati nekatere medmete v slovar (npr. »mah«), po drugi pa je bila prej tudi omejitev, da je pogoj za vejico velika začetnica.

5.2.6. Začetni pozivi pri tvitih

Tviti se zelo pogosto začnejo s seznamom naslovnikov.

@leaathenatabako Evo poznavalke.
@jureflux A si vedu, da je to na Islandiji?
@JJansaSDS @vladaRS dej ze enkrat tih bod no, kaj si pa ti naredu?

Slika 10: Primeri z začetnimi pozivi.

Po eni strani bi se na te primere dalo gledati, kot da gre za začetne zvalnike in bi jim potem sledila vejica. Vendar so ti pozivi običajno dodani samodejno in zato pisci tvitov nanje ne gledajo kot del besedila (kar kažejo tudi velike začetnice za njimi). Analizator je bil dopolnjen tako, da v primeru, da je nastavljen na obravnavo tvitov, preskoči vsa imena na začetku.

5.3. Trenutni rezultati za slovnični pregledovalnik Besana

Besana trenutno pri popravljanju vejic v zbirki primerov rabe vejice Vejica 1.3 doseže rezultate, prikazane v tabeli 2.

del	priklic	natančnost
KUST	87,04 %	94,35 %
Šolar	77,62 %	90,76 %
Lektor	50,45 %	38,06 %
Wikipedija	79,84 %	86,73 %
Janes	49,48 %	79,26 %
SKUPAJ	75,59 %	85,78 %

Tabela 2: Popravljanje manjkajočih vejic.

Pri popravljanju manjkajočih vejic so rezultati zelo raznoliki, najboljši priklic je pri delih KUST in Wikipedija, najslabši pa pri delih Janes in Lektor. Slabši rezultat pri Janesu kaže na to, da bo treba še dopolniti analizo nestandardne slovenščine, po drugi strani pa je slabši rezultat v Lektorju posledica tega, da tam v precejšnji meri manjkajo vejice, ki jih je težko avtomatsko postaviti (torej zanje ne zadošča, da pogledamo le besede v neposredni okolici), kar se npr. vidi iz tega, da je beseda, pred katero v Lektorju manjka največ vejic, »in«, na katerega odpade 18 % vseh manjkajočih vejic (Holozan, 2015).

Pri natančnosti je najboljši rezultat pri delih KUST in Šolar, najslabši pa pri delih Lektor in Janes. Pri delu Lektor je natančnost izrazito slabša (več kot pol slabša od vseh drugih delov), razlog za to najbrž to, da je delež manjkajočih vejic v tem delu (1,57 %) veliko manjši kot v drugih delih (na drugem mesti je del Wikipedija, ki vsebuje 11,78 % manjkajočih vejic, torej več kot sedemkrat več, pri drugih delih pa je razlika še večja).

del	priklic	natančnost
KUST	27,72 %	96,55 %
Šolar	37,42 %	93,72 %
Lektor	40,45 %	58,47 %
Wikipedija	45,07 %	96,97 %
Janes	30,00 %	40,00 %
SKUPAJ	37,75 %	85,17 %

Tabela 3: Popravljanje odvečnih vejic.

Priklic pri popravljanju odvečnih vejic je slabši kot pri popravljanju manjkajočih, natančnost pa je primerljiva.

Zanimivo je, da je najslabši priklic pri delu KUST, pri katerem je pri manjkajočih vejicah priklic najboljši, drugi najslabši pa je pri Janesu, vendar je pri Janesu le malo primerov odvečnih vejic in tudi sicer delež odvečnih vejic ni velik. Delež odvečnih vejic pri delu KUST je tako več kot desetkrat večji od Janesa in najbrž so vse te odvečne vejice preveč nepredvidljive za Besano (bi se pa bilo temu smiselno bolj posvetiti, če bi želeli narediti verzijo Besane, ki bi bila bolj uporabna za osebe, ki se učijo slovenščino kot drugi jezik).

Natančnost je spet najslabša pri Lektorju, tudi tukaj pa velja, da je delež manjkajočih vejic v Lektorju daleč najmanjši od vseh delov (1 %).

6. Možnosti za nadaljnje delo

Ena smer je še nadaljnje preverjanje pravilnosti označenih napak pri vejicah, vendar takih napak najbrž ni več veliko.

Bolj zanimivo bi bilo razširiti zbirko primerov še z dodatnimi deli. Ena možnost bi bila uporaba SSJ500k, kot je predlagal že Holozan (2016). Ta del sicer najbrž ni tako zanimiv za napake pri vejicah, ker teh verjetno ni tako veliko, je pa zanimiv, ker je ročno označen (približno polovica je tudi skladenjsko razčlenjena), kar omogoča, da se preizkusijo različne metode postavljanja vseh vejic in se ugotovi, kakšen rezultat bi bilo mogoče doseči, če bi imeli res dobro oblikoskladenjsko označevanje in stavčno razčlenjevanje. Bi pa bilo za ta del treba ročno pregledati vejice, kar zahteva precejšen časoven vložek (po drugi strani pa je res, da če bi nas zanimalo postavljanje vseh vejic, in ne popravljanje napačnih vejic, bi lahko uporabili SSJ500k tudi brez ročnega označevanja napačnih vejic, ker delež napačnih vejic najbrž ni prevelik, če pa bi dodatno za iskanje napačnih vejic uporabili kar Besano, bi lahko polovico napak odkrili z dovolj malo dela, kar bi za postavljanje vseh vejic bilo že čisto uporabno).

Glede na to, da imamo že vključena besedila šolarjev, dijakov in profesionalnih piscev, bi bilo zanimivo izdelati še zbirko primerov študentskih besedil. Morda bi se dalo za pripravi take zbirke primerov izkoristiti dejstvo, da morajo na veliko fakultetah biti diplomska in magistrska dela obvezno lektorirana. Ta besedila so zbrana v Korpusu akademske slovenščine (KAS) (Erjavec at al., 2016). Zanimiva bi bila primerjava napak pri postavljanju vejic med besedili s fakultet, ki zahtevajo lekturo, v primerjavi s tistimi, ki lekture ne zahtevajo. Bi pa bilo treba vejice spet ročno preveriti, kar zahteva kar nekaj dela.

Glede izboljševanja slovničnega pregledovalnika Besana pri popravljanju vejic je še veliko rezerve pri popravljanju nestandardnih besedil, po drugi strani pa to ni ravno tipična uporaba. Zato bo bolj smiselno izboljšati rezultat pri delu iz korpusa Lektor, še posebej natančnost (Besana ima kar nekaj težav pri strokovnih besedilih, in sicer analizator še ne zna pravilno upoštevati citiranj, težave pa mu delajo tudi seznami literature, pri katerih pa je težava, da obstajajo različni formati in da vsebujejo veliko tujih besed).

7. Sklep

Izboljšana verzija zbirka primerov rabe vejice Vejica 1.3 omogoča še boljše preizkušanje programov za popravljanje vejic. Dodatek dela s tviti v spletni slovenščini omogoča še preizkušanje tovrstne nestandardne slovenščine.

Analizator povedi in s tem Besana sta bila dopolnjena, da bolj učinkovito obdelujeta nestandardna besedila, rezultati pa kažejo, da se to da še izboljšati, saj so ti rezultati v delu Janes slabši kot pri drugih delih.

Rezultati popravljanja vejic so sicer občutno boljši od tistih, ki so bili objavljeni v Holozan (2016), po eni strani zaradi izboljševanja Besane, po drugi strani pa zaradi čiščenja primerov. V delu Šolar takoj Besana najde 77,62 % manjkajočih vejic (pri natančnosti 90,76 %), medtem ko je leta 2016 našla 64,95 % manjkajočih vejic (z natančnostjo 89,48 %), še večja pa je razlika pri delu Lektor (priklic na 50,45 % iz 33,84 %, natančnost pa na 38,06 % iz 18,19 %).

Nasploh rezultati kažejo, kako pomembna je izbira primeru pri preizkušanju in je zato za primerjavo nujno, da se uporabijo isti primeri, kar je omogočeno z licenco CC in objavo zbirke primerov rabe vejice v repozitoriju CLARIN.SI.

8. Literatura

Iñaki Alegria, Bertol Arrieta, Arantza Diaz de Ilarraza, Eli Izagirre, Montse Maritxalar. 2006. Using Machine Learning Techniques to Build a Comma Checker for Basque. V *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, str. 1–8. Association for Computer Linguistics. 1–8.

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar, Milan Ojsteršek. 2016. Slovenska akademska besedila: prototipni korpus in načrt analiz. V: *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2016*, str. 58–64. Znanstvena založba Filozofske fakultete v Ljubljani.

Peter Holozan. 2012. Kako dobro programi popravljajo vejice v slovenščini. V: Jezikovne tehnologije: *Zbornik C 15. mednarodne multikonference Informacijska družba IS 2012, 8. do 12. oktober 2012*, str. 101–106. Institut Jožef Stefan. 101–106.

Peter Holozan. 2013. Uporaba strojnega učenja za postavljanje vejic v slovenščini. *Uporabna informatika*, XXI/3: 196–209.

Peter Holozan. 2015. Možnosti uporabe jezikovnih tehnologij za določanje težav pri rabi vejice. V: Helena Dobrovoljc in Tina Lengar Verovnik, ur., *Pravopisna razpotja*, str. 77–92. Založba ZRC, Ljubljana.

Peter Holozan. 2016. *Računalniško postavljanje vejic v slovenščini*. Doktorsko delo, Filozofska fakulteta, Univerza v Ljubljani.

Anja Krajnc. 2015. *Postavljanje vejic v slovenščini s pomočjo strojnega učenja.* Diplomsko delo, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

Anja Krajnc, Marko Robnik-Šikonja. 2015. Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšanega korpusa Šolar. V *Zbornik konference Slovenščina na spletu in v novih medijih*, str. 38–43. Znanstvena založba Filozofske fakultete v Ljubljani.

Nikola Ljubešić, Tomaž Erjavec, Darja Fišer. 2018. Orodja za procesiranje nestandardne slovenščine. V: Darja Fišer, ur., *Viri, orodja in metoda za analizo spletne slovenščine*, str. 74–99. Znanstvena založba Filozofske fakultete Univerze v Ljubljani, Ljubljana.

Karin Piškur. 2015. *Postavljanje vejic v slovenskih besedilih z orodjem LanguageTool.* Diplomsko delo, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani.

Damjan Popič. 2014. *Korpusnojezikovna analiza vplivov na slovenska prevodna besedila.* Doktorsko delo, Filozofska fakulteta, Univerza v Ljubljani.

Damjan Popič, Darja Fišer, Katja Zupan, Polona Logar. 2016. Raba vejice v uporabniških spletnih vsebinah. V: *Zbornik konference Jezikovne tehnologije in digitalna humanistika 2016*, str. 149–153. Znanstvena založba Filozofske fakultete v Ljubljani.

Damjan Popič in Darja Fišer. 2018. (Ne)normativnost računalniško posredovane komunikacije v slovenščini: merilo vejice. V: Darja Fišer, ur., *Viri, orodja in metoda za analizo spletne slovenščine*, str. 140–159. Znanstvena založba Filozofske fakultete Univerze v Ljubljani, Ljubljana.

Tadeja Rozman, Mojca Stritar, Irena Krapš Vodopivec, Iztok Kosem, Simon Krek. 2010. *Nova didaktika poučevanja slovenskega jezika : sporazumevanje v slovenskem jeziku*. Ministrstvo za šolstvo in šport: Amebis. http://www.slovenscina.eu/Media/Kazalniki/Kazalnik15/Nova_didaktika_Sporazumevanje.pdf.

Mojca Stritar. 2006. Oblikovanje korpusa usvajanja slovenščine kot tujega jezika. V: *Zbornik 5. slovenske in 1. mednarodne konference Jezikovne tehnologije*. Institut "Jožef Stefan".

Peter Holozan. 2016. *Corpus of comma placement Vejica 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1055.

Damjan Popič; et al. 2017. *Tweet comma corpus Janes-Vejica 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1088.

Tadeja Rozman; et al. 2013. *Learners' corpus Šolar 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1036.

Croatian Web Dictionary Mrežnik: One year later - What is different?

Lana Hudeček,[*] Milica Mihaljević†

[*]Institute of Croatian Language and Linguistics
Republike Austrije 16, Zagreb, Croatia
lhudecek@ihjj.hr

†Institute of Croatian Language and Linguistics
Republike Austrije 16, Zagreb, Croatia
mmihalj@ihjj.hr

Abstract

The authors compare the lexicographic experience of compiling a paper desk dictionary *Školski rječnik hrvatskoga jezika* (2012, *School Dictionary of the Croatian Language*, *ŠR*) with the compilation of *Hrvatski mrežni rječnik – Mrežnik* (*Croatian Web Dictionary – Mrežnik, M*) which is now in progress. They focus on the new insights brought to the lexicographic work by the use of Sketch Engine giving examples from both dictionaries.

1. Introduction

In the Institute of Croatian Language and Linguistics, the *Croatian Web Dictionary – Mrežnik* is being compiled within the research project IP-2016-06-2141 financed by the Croatian Science Foundation. It is a four-year project and the work on the project started on 1st March 2017. The dictionary consists of three modules: the module for adult native speakers of Croatian with 10,000 entries, the module for elementary school children with 3,000 entries, and the module for foreigners with 1,000 entries.[1] The dictionary is based on two Croatian corpora: *the Croatian Web Corpus hrWaC* (http://nlp.ffzg.hr/resources/corpora/hrwac/)[2] and *the Croatian Language Repository* (CLR; riznica.ihjj.hr)[3]. The lexicographers select freely data from the corpus (the dictionary is corpus-based and not corpus-driven) as well as from other Croatian dictionaries, websites, and other resources. The dictionary is corpus-based due to the normative aspect of the dictionary and the unrepresentativeness of the two available Croatian corpora. The corpora differ in size and the source of the texts: *hrWaC* is much bigger and consists of texts mostly from newspapers, forums, blogs, etc. written in the journalistic and/or colloquial style. *CLR* is smaller and consists mostly of older texts often written in the literary style. The dictionary work is supported by Sketch Engine[4], a corpus query system used to support the analysis of the lemmas. The compilation of the dictionary is based on Word Sketches[5] specially adapted to the needs of the project, which are based on a developed Sketch Grammar[6] and the application of the GDEx[7] module for finding appropriate examples in the corpus. Some categories were added to Word Sketches while sometimes regular expressions were used, e.g. with interjections and conjunctions where Sketch Grammar didn't give adequate results, (see Table 1).

To support the preparation of the dictionary text the TLex software package, a professional software application for compiling dictionaries is used. TLex is adapted to the needs of the project as entry fields in TLex have been designed according to the dictionary entry model developed by the editors of the dictionary, i.e. the authors of the paper. An important characteristic of *Mrežnik* is a system of links within a module (synonyms, antonyms, masculine/feminine pairs, derivatives) and to other databases outside *Mrežnik,* i.e. links to repositories which will be created as a part of this project and compiled simultaneously with the dictionary (*Conjunction Repository*, *Repository of Idioms*, *Repository of Ethnics and Ktetics*, *Male/Female Portal*) as well as with repositories which have already been compiled within other projects conducted at the Institute of Croatian Language and Linguistics. The result of the *Mrežnik* project will be a free, monolingual, hypertext, searchable, online dictionary of standard Croatian. The focus of this paper will be on the module for adult native speakers.[8]

[1] More about the project and the structure of the three modules see in Hudeček and Mihaljević, (2017a, 2017b, 2017c).

[2] *hrWaC* is a Croatian web corpus made up of texts collected from the Internet, i.e. from the .hr top-level domain. The corpus was created in January 2014 with the total size over 1.2 billion words. The current version of the corpus (v2.0) contains 1.9 billion tokens and is annotated with the lemma, morphosyntax, and dependency syntax layers.

[3] *Croatian Language Repository* is a project which started in 2005, and the corpus consists of Croatian literature, non-fiction, scientific publications and university textbooks, school books, literature translated by outstanding Croatian translators, journals and newspapers, books from the pre-standardization period of Croatian language that are adapted to standard Croatian.
The *Croatian Language Repository* corpus was processed using ReLDI tagger with Word Sketches version 1.4 by Nikola Ljubešić. It has 101,782,863 tokens and 85,273,724 words.

[4] All terms used in this paper are defined in the *Glossary* on the *Mrežnik* website ihjj.hr/mreznik/page/pojmovnik. More on Sketch Engine see in Kilgarriff et al (2014: 7-36).

[5] Croatian Word Sketches give Croatian collocations categorized by grammatical relations. Croatian Word Sketches do not have a reference yet but as they are adapted from the Slovenian model. For Slovenian Word Sketches see Krek (2006).

[6] See Kilgarriff et al (2010: 372–379).

[7] See Kilgarriff et al (2008).

[8] More about the module for children see in Hudeček and Mihaljević (in print) and more about the module for foreigners see in Hudeček et al (2018).

Categories added to Word Sketches					Regular expressions used to search the corpus	
Verb (V)	Preposition (Pr)	Pronoun (P)	Adjective (Adj)	Adverb (Adv)	Interjection (I)	Conjunction (C)
V + *se*, *se* + V	Pr + N, Pr + Adv, V + Pr, N + Pr, Adj + Pr	P + Par, Par + P	Par + Adj	Par + Adv, Adv + Par	Adv + I, V + I, I + Pr, I + N, N + I, I + *i*, I + I, I + P, Par + I	Adv + C, Par + C, C + Adv, C + C

(N = noun, Par = particle)

Table 1: Addition to Word Sketches by categories

2. The Road from *ŠR* to *Mrežnik*

The project team of *Mrežnik* consists of experienced lexicographers most of whom have already compiled the *School Dictionary of the Croatian Language* (*ŠR*), published in 2012, a normative dictionary consisting of 30,000 entries. It is based on a corpus of elementary and high school textbooks from which the lexicographers manually composed the alphabetical list of entries. It was written consulting *CLR*, i.e. all entry words were checked in the corpus for examples and collocations but examples and collocations were not taken from the corpus. It was written in the Softlex dictionary compilation program. Our expectations were that the work on *Mrežnik* would be similar to the work on *ŠR* and that its modified methodology and maybe even some definitions could be used in *Mrežnik*. However, starting work with *Word Sketches* and taking over examples from the corpus gave us new insights into lexicographic work.

2.1. Comparing *ŠR* and *Mrežnik*

The difference between *ŠR* and *Mrežnik* can easily be shown by comparing a prototype entry in both dictionaries. This is the structure of the entry *nastavnik* (teacher) in *ŠR*:

ACCENTUATED HEADWORD nástāvnīk
WORD CLASS *im. m.*
SELECTED ACCENTUATED FORMS OF THE HEADWORD <G nástāvnīka, V nástāvnīče; *mn.* N nástāvnīci, G nástāvnīkā>
DEFINITION osoba koja vodi nastavu
COLLOCATIONS COMPOSED BY THE LEXICOGRAPHER [~ *hrvatskog jezika*; *sveučilišni* ~]
This is the structure of the entry *nastavnik* in *Mrežnik*:

HEADWORD nastavnik

ACCENTUATED HEADWORD nástāvnīk

WORD CLASS *im. m.*

ALL ACCENTUATED FORMS OF THE HEADWORD

(GA nástāvnīka, DL nástāvnīku, V nástāvnīče, I nástāvnīkom; *mn.* NV nástāvnīci, G nástāvnīkā, DLI nástāvnīcima, A nástāvnīke)

1. MEANING – DEFINITION Nastavnik je odrasla osoba bez obzira na spol ili muškarac koji vodi nastavu u srednjoj školi ili na fakultetu.[9]

EXAMPLES FROM THE CORPUS *U tijeku ponavljanja godine studija studenti su se dužni uključiti u izvođenje nastavnog procesa za one predmete za koje im uredno pohađanje nastave nije potvrđeno potpisom predmetnog nastavnika, te trebaju uredno izvršavati nastavne obveze.*
Kako bi se solidarizirali s kolegama u Sindikatu učitelja i srednjoškolskih nastavnika, Ribićev sindikat spreman je odreći se isplate prosvjetnog dodatka od 2,3 posto čim odluka o ukidanju koeficijenata 3, 5, 7 i 9 stupi na snagu.
Profesionalni put Anđelko Klobučar započeo je kao srednjoškolski nastavnik u Zagrebu, a od 1958. do 1963. godine bio je glazbeni suradnik Dubrava-filma, skladajući mnogobrojne glazbe za igrane, kratke igrane, crtane i dokumentarne filmove, od kojih su mnoga od njegovih ostvarenja postala antologijskim primjerima žanra.

COLLOCATIONS FROM WORD SKETCHES INTRODUCED BY QUESTIONS
Kakav je nastavnik? (What is the teacher like?) dežurni, gostujući, honorarni, predmetni, srednjoškolski, strukovni *razg.*, sveučilišni, visokoškolski
Što nastavnik može? (What can the teacher do?) osmišljavati (program, radionicu), predavati, pripremati (nastavu, predavanja), sudjelovati (u projektu, u radu), voditi (aktivnosti, program, radionice)
Što se s nastavnikom može? (What can we do with the teacher?) javiti mu (da je učenik bolestan, da učenik ima problem), olakšati mu rad, omogućiti mu (napredovanje, praćenje, rad), preporučiti ga (za napredovanje, za zaposlenje), zaposliti ga
Čega je tko nastavnik? (What does the teacher teach?) engleskoga, fizike, informatike, hrvatskoga, matematike
Koordinacija (Coordination)*:* nastavnik i mentor, pedagog i nastavnik, profesor i nastavnik, roditelj i nastavnik; nastavnici i nenastavno osoblje, nastavnici i ravnatelj; odnosi se samo na muškarca: nastavnica i nastavnik

[9] Using male forms for both male and female and only for male is a common phenomenon in Croatian. This is stressed in all definitions in the module for adult users (odrasla osoba bez obzira na spol ili muškarac – an adult regardless of his sex or a male). As such a definition would be too complicated for school children in the module for schoolchildren this is explained in a special note. More about this see in Mihaljević (in print).

Povezuje se s (Is often connected with): dežurstvom, edukacijom, izobrazbom, kompetencijom, mobilnošću, obrazovanjem, osposobljavanjem, plaćama, usavršavanjem

SYNONYM FOR THE 1ST MEANING (LINK TO THE ENTRY) SINONIM: professor 4

STYLISTIC LABEL (COLLOQUIAL STYLE) [2] *razg.*

2. MEANING – DEFINITION Nastavnik je odrasla osoba bez obzira na spol ili muškarac koji vodi nastavu u višim razredima osnovne škole.[10]

EXAMPLES FROM THE CORPUS: *S učenicima će raditi mladi voditelji iz Saveza, daroviti informatičari, studenti i srednjoškolci koji su na natjecanjima postizali najbolje rezultate, kao i nastavnici informatike u osnovnim školama.*

Osnovnoškolski nastavnik (49) iz Slavonskoga Broda, osumnjičen za spolnu zloporabu djeteta mlađeg od 15 godina i iskorištavanje djece za pornografiju, pušten je nakon dva tjedna iz pritvora.

COLLOCATIONS FROM WORD SKETCHES INTRODUCED BY QUESTIONS

Kakav je nastavnik? dežurni, osnovnoškolski

Što nastavnik može? osmišljavati (program, radionicu), predavati, pripremati (nastavu, predavanja), sudjelovati (u projektu, u radu), voditi (aktivnosti, program, radionice)

Što se s nastavnikom može? javiti mu (da će učenik izostati s nastave), olakšati mu (posao, rad), omogućiti mu (napredovanje, praćenje, rad), preporučiti ga za što, zaposliti ga

Čega je tko nastavnik? engleskoga, fizike, informatike, hrvatskoga, matematike

Koordinacija: odgajatelj i nastavnik, profesor i nastavnik, roditelj i nastavnik, pedagog i nastavnik, učitelj i nastavnik; odnosi se samo na muškarca: nastavnica i nastavnik

SYNONYMS FOR THE 2ND MEANING (LINKS TO THE ENTRIES): SINONIMI: učitelj (predmetni učitelj 1), profesor 4

3. MEANING – DEFINITION [3] Nastavnik je odrasla osoba bez obzira na spol ili muškarac koji komu prenosi kakva znanja ili ga poučava kakvim vještinama.

EXAMPLES FROM THE CORPUS *Rukovoditelj letenja mora biti punoljetan, mora imati letačko iskustvo od najmanje 30 sati letenja i 30 uzlijetanja i slijetanja kao zapovjednik jedrilice nakon izdavanja dozvole pilota jedrilice, a kada lete učenici piloti mora imati važeće ovlaštenje nastavnika letenja.*

Rukovoditelj tehničkih poslova i nastavnik padobranstva u Aeroklubu Borovo Branislav Mišić izjavio je da se mladim padobrancima u prvih 10 - tak skokova padobran otvara automatski te da je takav slučaj trebao biti i s poginulim mladićem.

COLLOCATIONS FROM WORD SKETCHES INTRODUCED BY A QUESTION *Čega je tko nastavnik?* crtanja, gitare, kuhanja, letenja, padobranstva

SYNONYMS FOR THE 3RD MEANING (LINKS TO THE ENTRIES): SINONIM: učitelj 3

SUBENTRY strukovni nastavnik

DEFINITION Strukovni nastavnik organizator je i voditelj strukovno-teorijske nastave te praktične nastave i vježba u strukovnim školama; svoje je temeljno obrazovanje stekao na nekome od nepedagoških fakulteta, a pedagoške kompetencije stekao je dopunskim pedagoško-psihološkim i didaktičko-metodičkim obrazovanjem.

EXAMPLES FROM THE CORPUS *Agencija za strukovno obrazovanje i obrazovanje odraslih, od 26. do 28. ožujka 2018. godine u Opatiji, organizira „Dane strukovnih nastavnika".*

Može se zaključiti da je neophodno osigurati odgovarajuću izobrazbu i kontinuirano usavršavanje strukovnih nastavnika u pedagoško-didaktičkom, ali i strukovnom području.

FEMALE PAIR OF SUBENTRY (LINK TO THE ENTRY) ŽENSKO nastavnica (strukovna nastavnica), učiteljica (strukovna učiteljica 2)

SYNONYM FOR THE MEANING OF SUBENTRY (LINK TO THE ENTRY) SINONIM učitelj (strukovni učitelj)

PRAGMATIC NOTE[11] Riječi *učitelj, nastavnik* i *profesor* drukčije su određene u zakonu danas (*Zakon o odgoju i obrazovanju u osnovnoj i srednjoj školi* 2008.) nego što je to bilo prije, pa to dovodi do nedosljedne uporabe tih riječi. Danas prema Zakonu u osnovnoj školi rade učitelji, a u srednjoj i na fakultetu nastavnici. Značenje se tih riječi usklađeno sa Zakonom dosljedno nalazi npr. u natječajima za posao te na mrežnim stranicama škola. Međutim, značenje je tih riječi često drukčije upotrebljava u publicističkome i razgovornome stilu, pa se često govori i o osnovnoškolskim nastavnicima/profesorima i srednjoškolskim profesorima. U srednjim strukovnim školama rade strukovni učitelji, koji se često zovu i strukovni nastavnici. Zbrku povećava i to što učitelji mogu napredovati u zvanje učitelja mentora i učitelja savjetnika, a nastavnici u zvanje profesora mentora i profesora savjetnika. Na fakultetu rade sveučilišni nastavnici, koji mogu imati znanstveno-nastavno i umjetničko-nastavno zvanje docenta, izvanrednoga profesora i redovitoga profesora ili nastavno zvanje predavača, višega predavača ili profesora visoke škole. Dakle, profesori su redoviti profesori, izvanredni profesori i profesori visoke škole. Od radnih se mjesta i znanstveno-nastavnih i umjetničko-nastavnih zvanja razlikuju titule koje se dobivaju završetkom određenoga

[10] We decided to separate meanings 1 and 2 although the definitions differ only in emphasizing different levels of education for pragmatic reasons and the difference in synonyms. The first meaning is the official term used in documents (laws, certificates, diplomas, etc.), and the second is a colloquial and historic term which doesn't occur in contemporary official documents. It is synonymous to the official term *učitelj* (or *razredni učitelj*).

[11] In the pragmatic note the difference between the usage of the words *učitelj, nastavnik,* and *profesor* are explained as these words have similar meanings which vary according to the formality of style and are used differently e.g. in legal documents and newspapers. Also, they are used differently now than they were ten years ago so the meaning of these words also varies according to the time when the text was written.

fakulteta. Onaj tko završi učiteljski fakultet, dobiva titulu diplomirani učitelj/diplomirana učiteljica. Onaj tko završi nastavnički fakultet po bolonjskome procesu, dobiva titulu mag. edu., odnosno magistar edukacije, ali onaj tko je diplomirao prije uvođenja bolonjskoga procesa, dobivao je titulu profesora, pa danas u školama radi još mnogo profesora. Učenici osnovne škole danas se najčešće svojim učiteljima obraćaju riječju *učitelj*, a učenici srednje škole i studenti svojim se nastavnicima obraćaju riječju *profesor*. Odnos među riječima *učitelj*, *nastavnik* i *profesor* odražava se i na odnos među riječima *učiteljica*, *nastavnica* i *profesorica*.

By comparing these two entries we can conclude:

1. the entries in *Mrežnik* are much longer
2. while the entry for the headword *nastavnik* in *ŠR* has only these fields: word class, selected grammatical forms, definition, two collocations, the structure of the entry for the same headword in *Mrežnik* is as follows: word class, all grammatical forms, three different definition with separate examples and collocations (introduced by questions or phrases *What is xx like?*, *What can xx do?*, *What can we do with xx?*, *What does xx teach?*, *Coordination:*, *Often connected with:*), some collocations having a stylistic label (*razg.* – colloquial), subentry with examples, synonyms connected with particular meanings (links), male/female pairs connected with particular meanings (links), pragmatic note, word formation information and derivatives. If the derivatives are separate entries in the dictionary, they are connected via hyperlink to their entry and if not, they are only listed.

There are two answers to the question why the structure of the same entry in these two dictionaries differs so much:

1. As *Mrežnik* is a web dictionary more data than in the printed dictionary could be given: all accentuated noun forms, information on the male/female pair (as we have experience that this is the data that users require very often[12]), word-formation information and derivatives (as this is the data very often required by students), questions for collocations based on the model used in elexiko[13] and a pragmatic note.
2. Using Word Sketches we found new meanings not recorded in Croatian dictionaries. Every new meaning has examples from the corpus, which led us to the awareness that often there are useful pragmatic comments which could be inserted into the dictionary. In the example above while using Word Sketches and corpus examples we became aware of the complex net of relations between the two sets of words: *učitelj*, *nastavnik*, *professor* (male teacher) and *učiteljica*, *nastavnica*, *profesorica* (female teacher) as these terms are used differently (and sometimes inconsistently even in the same document) in legal documents, school practice, newspapers, and everyday speech. We interviewed a few schoolteachers to get a clearer view of how these terms are used in schools. For this reason, a short pragmatic note was introduced into all six of the above-mentioned entries.

The relation between male and female pairs is complicated by the fact that the male form can mean *male* (in the dictionary definitions this is expressed by *muškarac* if it applies only to adults or *muška osoba* if it applies to male children as well) but also male and female (especially in the plural) regardless of gender. This is expressed by the formula *osoba bez obzira na spol* or *muška osoba bez obzira na spol* (person regardless of gender or adult person regardless of gender).

2.2. Collocations and examples

As collocations and examples are extracted from the corpus this brings us to another difference between *ŠR* and *Mrežnik*. This will be shown on the example of the feminative of the entry *nastavnik* (teacher) *nastavnica* (female teacher).

Comparing these two entries we can see that in *ŠR* the entry for the word *nastavnik* closely corresponds to the entry for the word *nastavnica* and once having compiled the entry *nastavnik* we could compile *nastavnica* in a matter of minutes. On the other hand, in *Mrežnik* the entry *nastavnica* corresponds to the entry *nastavnik* in definitions and questions asked for collocations but differs considerably in most of the examples and collocations (see Table 2).

This, however, brings us to the new problem of how closely our collocations should correspond to that what we get from Word Sketches. So far we have noticed two major problems:

1. Collocations for male and female agent nouns differ considerably and in a way that shouldn't (because the dictionary has an educational purpose and not only a scientific one) be reflected in a dictionary of the standard language. This will be illustrated by comparing the collocations for *konobar* (waiter) and *konobarica* (waitress) and *pekar* and *pekarica* (male and female baker). If we look for the lemma *konobarica* in Sketch Engine the first collocation is *brkata* (with a moustache) followed by *sisata*, *prsata* (having big breasts), and the first verb having *konobarica* as an object is an impolite verb *zajebavati*. On the other hand, if we check the corresponding male word *konobar* the collocations are quite different: *neljubazan* (impolite), *ljubazan* (polite), *pomoćni* (assistant), *priučeni* (trained on the job, inadequately trained), and the verbs having the highest score are *zamoliti*/*moliti* (ask), *dozvati*/*zovnuti* (call). With the female noun *pekarica* (baker) among the top collocations are *fatalna* (fatal) and *pohotna* (lusty). Even with the word *čistačica* (cleaning lady) *seksi* (sexy) has a very high score. No such adjectives occur with the masculine words *pekar* and *čistač*.

[12] We give language advice on a daily basis by telephone and e-mail and questions about female/male pairs occur very often. That is the reason we have launched a new project *Male and female in Croatian*. More on the project see on ihjj.hr/projekt/musko-i-zensko-u-hrvatskome-jeziku.

[13] See Klosa (ed.) (2011) and Möhrs (2014).

ŠR	**nástāvnica** *im. ž.* <G nástāvnicē; *mn.* N nástāvnice, G nástāvnīcā> žena koja vodi nastavu [~ *hrvatskog jezika*; *sveučilišna* ~]
M	nástāvnica **nastavnica** im. ž. (G nástāvnicē, DL nástāvnici, A nástāvnicu, V nástāvnice, I nástāvnicōm; *mn.* NAV nástāvnice, G nástāvnīcā, DLI nástāvnicama) **1** Nastavnica je žena koja vodi nastavu u srednjoj školi ili na fakultetu . *Nastavnica biologije u istoj srednjoj školi Dolores Dobrinkić na posao putuje triput tjedno iz Splita, za što je u siječnju potrošila dvije trećine plaće. Članica ste pokreta »Opus Dei«, što nekako sugerira, s obzirom i na Vaš posao sveučilišne nastavnice, angažman u javnosti.* *Kakva je nastavnica?* mlada, omražena, omiljena, predmetna, srednjoškolska, stroga, sveučilišna, umirovljena, x-godišnja; *Što nastavnica može?* organizirati (nastavu, predavanja, radionice), osmišljavati (program, radionicu), predavati, pripremati (nastavu, predavanja), sudjelovati (u projektu, u radu), voditi (aktivnosti, program, radionice); *Što se s nastavnicom može?* napadati je se, pitati je se što, pozdravljati je; *Čega je tko nastavnica?* biologije, engleskoga, fizike, informatike, hrvatskoga, matematike , povijesti, vjeronauka; *Koordinacija:* mentorica i nastavnica, nastavnica i nastavnik, profesorica i nastavnica, razrednica i nastavnica (čega), ravnateljica i nastavnica, učenice i nastavnice, učiteljice i nastavnice **MUŠKO: nastavnik :1** **SINONIM: profesorica** 2 razg. Nastavnica je žena koja vodi nastavu u višim razredima osnovne škole . *Osnovnoškolska nastavnica koja je bez ikakvog jasnog razloga izbacivala s nastave dijete koje je štićenik doma za nezbrinutu djecu, odbijajući priznati da je učenik škole, dobila je zabranu rada od Prosvjetne inspekcije. Pretvorili smo se u tvornice 'biflanja', a mi nastavnici pretvoreni smo u birokrate - istaknula je osnovnoškolska nastavnica Ivana Kovač.* *Kakva je nastavnica?* dežurna, osnovnoškolska; *Što nastavnica može?* organizira (nastavu, predavanja, radionice), osmišljava (program, radionicu), predaje, priprema (nastavu, predavanja), sudjeluje (u projektu, u radu), vodi (aktivnosti, program, radionice); *Što se s nastavnicom može?* napadati je se, pitati je se, pozdravljati je se; *Čega je tko nastavnica?* biologije, engleskoga, fizike, informatike, hrvatskoga, matematike povijesti, vjeronauka; *Koordinacija:* nastavnice i nastavnici, odgajateljice i nastavnice, učiteljice i nastavnice **MUŠKO: nastavnik 2** **SINONIM: učiteljica predmetna učiteljica :1** 3 Nastavnica je žena koja komu prenosi kakva znanja ili ga poučava kakvim vještinama. *Pa zbog velikog broja polaznika tečajeva gitare angažirali smo još jednu kolegicu kao drugu nastavnicu gitare.* *Čega je tko nastavnica?* gitare, klavira, pjevanja **MUŠKO: nastavnik 3** **SINONIM: učiteljica 2** **strukovna nastavnica** Strukovna nastavnica organizatorica je i voditeljica strukovno-teorijske nastave te praktične nastave i vježba u strukovnim školama; svoje je temeljno obrazovanje stekao na nekom od nepedagoških fakulteta, a pedagoške kompetencije stekao je dopunskim pedagoško-psihološko i didaktičko metodičkim obrazovanjem. **MUŠKO: nastavnik strukovni nastavnik :1** **SINONIM: učiteljica (strukovna učiteljica 1)** Riječi *učitelj*, *nastavnik* i *profesor* drukčije su određene u zakonu danas…

Table 2: Entries *nastavnica* (female teacher) in *ŠR* and *Mrežnik* (M)

2. As both *hrWaC* and *Riznica* have many newspaper examples, very often collocations reflect news reporting on murder, rape, drugs, etc. not really characteristic for a certain word. This is especially common for female agent nouns, e.g. with many female agent nouns verbs *maltretirati*, *ubiti*, and *silovati* occur (mistreat, kill, rape), e.g. with the word *nastavnica* collocations *pretući*, *gađati*, *napasti* (beat, hit, attack) are the first three results in the row *koga – što*, i.e. the row in which *nastavnica* is in the accusative case. If we compare the same row for the word *djevojka* girl and the word *nastavnik* we get these results: The collocations for the word *djevojka* (girl) with the

highest score are *zaprositi*, *silovati*, *oženiti*, *upoznati*, *ubiti* (ask to marry, rape, marry, meet, kill) and the collocations for *nastavnik* are *educirati*, *obrazovati*, *osposobljavati*, *pozivati* (educate, specialize, invite). While analyzing these results we have to bear in mind the above-mentioned fact that *nastavnik* often refers to a male as well as a female person, especially in legal texts. As the dictionary is not corpus-driven but only corpus-based we avoided collocations which are not characteristic for a certain word or which can be offensive to a general dictionary user or which we feel are only the result of the unrepresentativeness of the corpus in which there are too many journalistic texts. However, such collocations will occur with the verbs *kill*, *rape*, *mistreat*, etc. as they illustrate the basic meaning of these words. The corpus gives also many uninformative collocations as *postati konobarica/konobaricom*, *raditi kao konobarica* (become a waitress, work as a waitress). They are uninformative as they can occur with any professional noun, e.g. *medicinska sestra* / *nastavnica* / *pekarica* / *profesorica* / *učiteljica*, etc.

2.3. Synonyms and antonyms

Working with the corpus and Sketch Engine our approach to synonyms and antonyms has also changed. In *ŠR* synonymous entries resembled each other completely, i.e. they had the same example differing only in the synonymous word and they had same collocations. If a word is used only in the colloquial style it was marked with the label *razg*. and directed to the entry belonging to the neutral standard language (*v.* – see). In *Mrežnik* such words have complete entries, consisting of a definition (or definitions), which is the same as the definition of their synonym belonging to the neutral standard language, examples and collocations (from the corpus, i.e. different from that of their synonyms). The words *stomatolog* and *zubar* were selected to illustrate this point as, although the word *zubar* is often used in Croatian, it is not the official professional term. This can be seen on the page of Croatian Terminological Database *Struna*. Entries *zubar* and *stomatolog* in *Mrežnik* will be linked to the entries in *Struna*.

ŠK	**stomatòlog** *im. m.* <G stomatòloga, V stomatòlože; *mn.* N stomatòlozi, G stomatòlōgā> liječnik koji se bavi stomatologijom; ***sin.***: zubar *razg.*
	zùbār *im. m.* <G zubára, V zȕbāru/zȕbāre; *mn.* N zubári, G zubárā> *razg. v.* stomatolog
M	stomatòlog **stomatolog** im. m. (GA stomatòloga, DL stomatòlogu, V stomatòlože, I stomatòlogom; *mn.* NAV stomatòlozi, G stomatòlōgā, DLI stomatòlozima, A stomatòloge) Stomatolog je liječnik koji se bavi stomatologijom. *Najnižu cijenu usluga određuje Stomatološka komora a potom je svaki privatni stomatolog usklađuje sa svojim mogućnostima naplate, zarade itd., što ovisi o mnogo čimbenika. U ordinaciji stomatologa dr. Živka Dijana jučer je u Zadru održana edukacijska radionica na kojoj su prezentirane najsuvremenije metode pomlađivanja lica, usana i zubnog mesa filerima Esthelis.* *Kakav je stomatolog?* budući, dežuran, dječji, estetski, izabrani, kvalitetni, nezaposlen, obiteljski, odabrani, poznati, privatni, ugovorni, vrhunski; *Što stomatolog može?* izvaditi zub/živac, preporučiti (terapiju, vađenje zuba, zubni konac); *Što se sa stomatologom može?* izabrati ga, obavijestiti ga (o lijekovima koje tko uzima, o terapiji, o trudnoći), posjetiti/posjećivati ga, preporučiti ga komu, zamoliti ga (da što preporuči/objasni); *Koordinacija:* ginekolozi i stomatolozi, liječnici i stomatolozi, pacijenti i stomatolozi, pedijatri i stomatolozi, stomatolozi i zubni tehničari; stomatolog ili specijalist (dentalne patologije, endodoncije, oralne kirurgije, paradontolog, stomatološke protetike, za plastičnu kirurgiju); *Povezuje se s*: dežurstvom, intervencijom, kongresom, kontrolom, nadzorom, pomoći, posjetom, pregledom, preporukom, savjetom, udrugom, udruženjom, uputom **ŽENSKO: stomatologica :1, stomatologinja :1, zubarica :1** **SINONIM: zubar :1**
	zùbār **zubar** im. m. (GA zubára, DL zubáru, V zȕbāru/zȕbāre, I zubárom/zubárem; *mn.* NV zubári, G zubárā, DLI zubárima, A zubáre) *razg*. Zubar je liječnik koji se bavi stomatologijom. *Tko želi zdrave zube i čvrste desni, mora zube prati 2 do 3 puta dnevno, svilenim koncem pročistiti prostor između zuba i redovito ići zubaru na kontrolu. U nekoliko trenutaka možete*

	postići osmijeh iz snova, a da pritom ne idete zubaru na zahvate koji koštaju i oštećuju zubnu caklinu. *Kakav je zubar?* besplatan, dežurni, izvrstan, poznati, privatni, socijalni, ugledan, uspješan, vrhunski ; *Što zubar može?* brusiti zub, liječiti desni/zub, otvoriti zub, praviti/raditi navlaku/plombu/protezu/zub, vaditi zub/živac; *Što se sa zubarom može?* bojati ga se, ići k njemu, dolaziti k njemu, imati ga, javljati mu se, nazivati ga se, plaćati mu, posjećivati ga; *Koordinacija:* kirurzi i zubari, zubar i ortodont, zubari i doktori, zubari i frizeri, zubari i ginekolozi, zubari i liječnici, zubari i okulisti; odnosi se samo na muškarce: zubar ili zubarica; *Povezuje se s:* preporuka, strah, usluga **ŽENSKO: stomatologica :1, stomatologinja :1, zubarica :1** **SINONIM: stomatolog :1**

Table 3: Synonymous entries *stomatolog* and *zubar* (dentist) in *ŠR* and *Mrežnik*

All this holds for antonyms as well. Moreover, as there is no need to save space in a web dictionary in polysemic entries synonyms and antonyms are connected to each meaning, i.e. not connected to the lemma even in cases of full synonyms/antonyms.

2.4. Corpus versus system

Working with the corpus often made us choose between the language system and data derived from the corpus. This was a common problem with lemmas which have a low frequency in the corpus. In Croatian the professional noun of masculine gender can in certain contexts (especially used in the plural) refer to persons of both sexes (or of unknown sex) and in other only to males. However, with certain entries and subentries such contexts were difficult to find (e.g. subentry *strukovni nastavnik* professional teacher, teacher in a vocational school).

On the other hand, for some entries, a corresponding female (e.g. *strukovna nastavnica*) or male noun (e.g. *pomoćnik porodničara* male midwife) could not be found in the corpus or had a very low frequency. Nevertheless, we decided to include such entries not only because they reflect the language system but also as many users of our language advice services often ask for information on the formation and usage of masculine/feminine pairs. Often native speakers of Croatian have problems in forming feminine/masculine pairs and due to the changing sociolinguistic context the need to use them occurs. As there are many problems connected with the relations between male and female nouns in Croatian a separate research project *Muško i žensko u hrvatskome jeziku* closely connected with the *Mrežnik* project is also conducted at the Institute.

Some examples in which Croatian native speakers have problems in forming a feminine or a masculine noun of a more frequently used feminine/masculine noun denoting professions are shown in the table below:

Masculine	**Feminine**	**English**
diskdžokej	diskdžokejica	disk-jockey
knjigoveža	knjigoveškinja	bookbinder
krupje	krupjeica	croupier
mornar	mornarica	sailor
ronilac	roniteljica	diver
tekstopisac	tekstopiskinja	songwriter

Table 4: Masculine/feminine pairs native speakers of Croatian find difficult

Some feminine/masculine forms are explained in a special note (advice), e.g. *čitalac* > *čitatelj*, *psihologica* > *psihologinja*. In *ŠR* only some of these examples were marked by the labels *v.* (see) and → (replace with) while examples which do not occur as often were not included in the dictionary.

2.5. Pleonasms and paronyms

The corpus made us aware of the fact that pleonasms occur very frequently. For example, the expression *no međutim* has 1611 occurrences in hrWaC, so the users of *Mrežnik* will be made aware of this very common mistake in a note. Another very common pleonasm has the structure *žena* (woman) + feminine form of the agent noun, e.g. *žena vozačica* (woman + driver in the feminine form). The corpus makes us aware of the fact that this is mostly used in the negative context connected with the stereotype 'women are bad drivers'. *Mrežnik* offers the advice that instead of the pleonastic construction *žena vozačica* in a stylistically unmarked context only *vozačica* should be used.

The corpus also makes us aware of mistakes connected with the use of paronyms, e.g. words *psihički*, *psihološki*, and *psihologijski* (mental or psychic, psychological); *genski*, *genetički*, and *genetski* (referring to gens, genetics or genesis); *religijski* and *religiozan* (religious or poius, religious) are often confused. The meanings of terms *smrtni list* and *smrtovnica* (both terms can be translated as *death certificate* into English) are often confused or these terms are considered as synonymous although they refer to two different documents as *smrtovnica* is a document issued by a registrar on the basis of *smrtni list* issued by the coroner. The relations between these words are then explained in a pragmatic note.

3. Conclusion

The possibilities of linking different resources and giving different data are enormous and one should not fall into the trap of giving some information only because it is available (as one of the reviewers of the project warned us) and not because the users need it. If we receive certain questions repeatedly from the users of our language advice service or notice it is the topic of discussion on different blogs (e.g. the above-mentioned difference between *učitelj*, *nastavnik*, and *profesor*) we consider that an explanation should be given in the dictionary. A note on language usage will be edited within the dictionary while additional data (verb valency, the etymology of idioms, metaphoric extensions, terminological data, etc.) will be given as additional information on a link.

The work with collocations from Sketch Engine has after only one year broadened our lexicographic views:

1. Often after looking at Word Sketches, we have decided we need to have more than one meaning for a word that had only one meaning in *ŠR*.

2. We have concluded that although a semantic relation (synonyms, antonyms, male/female relation) exists between particular words their collocations can differ considerably. Thus, collocations are given for each meaning of the lemma separately, i.e. synonyms have the same definition but can have different collocations.

3. We couldn't rely on Word Sketches completely and had to make a selection between offered data as many examples were uninformative, not characteristic for the lemma but for the corpus from which they were taken, not polite or biased towards a particular sex, social or ethnical group, etc.

4. We have concluded that data that we got from the corpus and Word Sketches could often be useful and should be included in a pragmatic or a language advice note.

4. Acknowledgments

This paper is written within the research project *Croatian Web Dictionary – Mrežnik.* (IP-2016-06-2141), financed by the Croatian Science Foundation.

5. References

Baza hrvatskih glagolskih valencija – GLAVA. Accessed at http://ihjj.hr/projekt/baza-hrvatskih-glagolskih-valencija/27/ 18 January 2018.

Matea Birtić et al. 2012. *Školski rječnik hrvatskoga jezika*. Zagreb: Školska knjiga – Institut za hrvatski jezik i jezikoslovlje.

Bolje je hrvatski. Accessed at http://bolje.hr/ 18 January 2018.

Hrvatsko strukovno nazivlje – STRUNA Accesed at http://struna.ihjj.hr/ 18 January 2018).

Lana Hudeček and Milica Mihaljević. 2017a. Hrvatski mrežni rječnik – Mrežnik. *Hrvatski jezik*, 4(4):1–7.

Lana Hudeček and Milica Mihaljević. 2017b. A new project – Croatian web dictionary MREŽNIK. In *The Future of Information Sciences. INFuture 2017, Integrating ICT in Society*, pages 205-2013, Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences. Zagreb.

Lana Hudeček and Milica Mihaljević. 2017c. The Croatian Web Dictionary Project – Mrežnik. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pages 172–192. Lexical Computing CZ s.r.o., Brno – Leiden.

Lana Hudeček et al. 2017. Radionica na Croaticumu – provjera rječničke koncepcije modula za strance na terenu. *Hrvatski jezik*, 4/4: 9–12.

Lana Hudeček and Milica Mihaljević.2018. Normiranje hrvatskoga jezikoslovnog nazivlja. In Hrvatski prilozi 16. međunarodnom slavističkom kongresu, pages 49 – 62. Hrvatsko filološko društvo, Zagreb.

Jezični savjetnik Accessed at http://jezicni-savjetnik.hr/ 20 February 2018.

Adam Kilgarriff et al. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Adam Kilgarriff et al. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII Euralex International Congress*, pages 425–432. Universitat Pompeu Fabra. Barcelona. Institut Universitari de Linguistica Aplicada.

Adam Kilgarriff et al. 2010. A quantitative evaluation of word sketches. In *Proceedings of the XIV Euralex International Congress*, pages 372–379. Fryske Akademy. Leeuwarden.

Annette Klosa. (ed.). 2011. *elexiko. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs*. Narr. Verlag. Tübingen.

Simon Krek and Adam Kilgarriff. 2006. Slovene word sketches. In *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference*. Institut Jožef Štefan. Ljubljana. http://www.kilgarriff.co.uk/Publications/2006-KrekKilg-Ljub-SloveneWS.pdf.

Milica Mihaljević. Hrvatski mrežni izvori za djecu i strance. In Zbornik *20 godina kroatistike u Lavovu*. Lavov (in print)

Christine Möhrs. 2014. Landeskundliche Wortschatzübungen auf der Basis von Kollokationen. Zur Nutzung von elexiko für Deutschlehrende. In: *Themenheft »Dateninterpretation und -präsentation in Onlinewörterbüchern am Beispiel von elexiko«*. Deutsche Sprache 4/2014, pp 309-324.

Repozitorij metafora. Accessed at http://ihjj.hr/metafore/ 20 February 2018.

Portuguese Corpora of the 18th century: old Medicine texts for teaching and research activities

Maria José Bocorny Finatto*, Paulo Quaresma†, Maria Filomena Gonçalves ‡

* Federal University of Rio Grande do Sul - UFRGS, Linguistics Department,
Instituto de Letras, Av. Bento Gonçalves, 9500 - Campus do Vale - Prédio 43221, sala 217
Caixa Postal 15002 - 91501-970 - Porto Alegre –RS - Brasil
maria.finatto@gmail.com

†Universidade de Évora, Department of Computer Science, Laboratory of Informatics, Systems and Parallelism, R. Romão Ramalho, 59 - 7000 Évora, Portugal.
pq@di.uevora.pt

‡ Universidade de Évora, ECS/Department of Linguistics and Literatures,
CIDEHUS-UÉ/FCT (UID/HIS/00057/2013)
Largo dos Colegiais
7002-554 Évora, Portugal.
mfg@uevora.pt

Abstract

The aim of this paper is to demonstrate the application of the methodologies of Corpus Linguistics and of the Natural Language Processing (NLP) tools to an 18th century Portuguese medicine book. The general objective of this work is to apply the digital humanities tools to a text that has not yet received this kind of approach, in view of teaching and research activities.

1. Introduction

The aim of this paper is to demonstrate the application of the methodologies of corpus linguistics and of Natural Language Processing (NLP) tools to an 18th century Portuguese medicine book. Therefore, the purpose of this work is to present a preliminary essay with a view to a major project on a historical study of the medical terminology in the Portuguese language. It should be noted that, until now, the Portuguese old terminologies had not been studied with computing tools.

First of all, it is important to draw up the theoretical and methodological framework of the analysis, starting with the concept of Corpus Linguistics. Therefore, the general objective of this work is to apply digital humanities (Berry and Fagerjord, 2017; Marquilhas and Hendrickx, 2016) tools to a text from the 18th century that had not yet received this kind of approach, in view of teaching and research activities.

A historical corpus is a set of documents "intentionally created to represent and investigate past stages of a language and/or to study language change" (Claridge, 2008: 242). Nowadays, as mentioned by Kytö (2011), empirical research in Linguistics has increasingly relied on material drawn from a wide range of electronic corpora. In this regard, the history of various languages has (re)emerged as a research area where electronic resources and various kinds of search tools can represent a new stage in the way research has been carried out to investigate mechanisms involved in language change, as well as the features possibly accounting for different phenomena. This kind of corpora have proved particularly useful in some areas of linguistic research, such as: historical lexicology, terminology and lexicography. These areas involve problems and procedures that nowadays can be recognized as a new "digital philology" (Driscoll and Pierazzo, 2016; Paixão de Sousa, 2013a, 2013b).

As mentioned by Froehlich (2015), if we have a collection of documents organized as a corpus, it is possible to find patterns of grammatical use, or frequently recurring phrases in it. A researcher may also want to find statistically likely and/or unlikely phrases for a particular author or kind of text, particular kinds of grammatical structures or a lot of examples of a particular concept across a large number of documents in context. Corpus analysis, conduced with the help of different kinds of computational tools, "is especially useful for testing intuitions about texts and/or triangulating results from other digital methods" (Froehlich, 2015).

However, in spite of the progress made by these new digital collections of data, with the support of Natural Language Processing (NLP) and Corpus Linguistics tools, there are many difficulties to overcome when handling old documents in digital format. One of the greatest difficulties remains at the computational processing of written language in ancient texts, whether handwritten or printed. Identifying spelling and even updating them are important challenges for the linguists as well as for the NLP researchers.

Taking this challenge into account, this article presents a set of initial procedures for the design of a corpus consisting of samples of ancient medical texts printed in Portuguese of the 18th century on the subject "diseases and their treatments". Our starting point was the book ***Observaçoens medicas doutrinaes de cem casos gravissimos*** (Semedo, 1707). It was printed in Lisbon, Portugal, in 1707, with 635 pages, published by João

Curvo Semedo (1635-1719), a Portuguese physician from Monforte, Alentejo, a region within Portugal.

It is important to emphasize that Semedo produced several medical treatises and handbooks of which the following are examples: *Polyanthea medicinal* (1697), our selected book ***Observaçoens medicas doutrinaes de cem casos gravissimos*** (1707) [Medical and doctrinal Observations of a hundred serious cases as Figure 1 shows] and *Atalaya da vida contra as emboscadas da morte* (1720) [freely translated as: An Observatory of life against the traps of death]. Thereby, the choice of Curvo Semedo is justified by being one of the "most popular doctors throughout the Portuguese empire in the eighteenth century" (Furtado, 2008: 147) and because the majority of treatments he prescribed ("Curvian secrets") were made with ingredients from Brazil, Africa and Asia. The works of Semedo confirm the opening of European medicine to products from other regions of the world.

In addition, his work represents, in linguistic terms, the period of the "classical Portuguese" (Castro, 2006: 73, 183-198; Banza and Gonçalves, 2018: 39-47), while illustrating the medical terminology of this period. It should be noted that, although the emergence of Portuguese language terminologies (Verdelho, 1998) represents a true technological metamorphosis of the language, its historical analyses still lacks a systematic study, a situation that also applies to the medical terminology.

In the scenario of the ancient Portuguese lexicography, the terms of Medicine received a specific mark ("medicine term") as we can see in the *Vocabulario Portuguez e Latino* (Portuguese and Latin Vocabulary) of Rafael Bluteau (1712-1728). This is a dictionary which is an indispensable work for the study of the different technical and scientific terminologies.

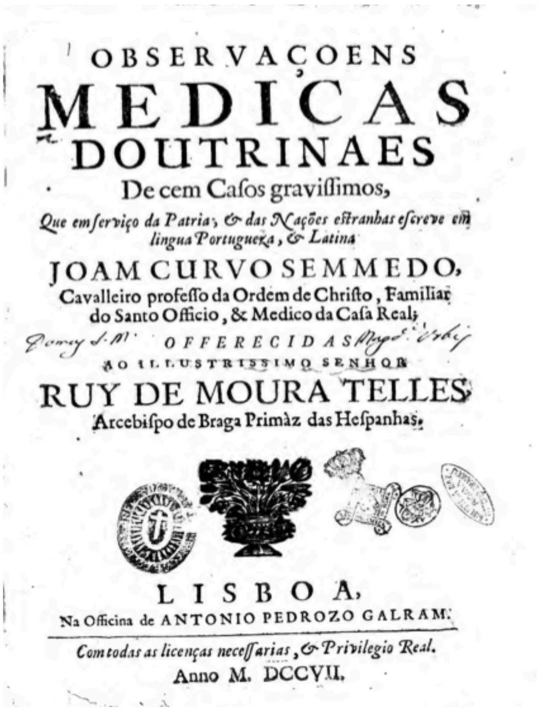
On the other hand, the works of Semedo inspired other treatises, namely works published by Portuguese doctors who practiced Medicine in Brazil. Thus, his book ***Observaçoens medicas doutrinaes de cem casos gravissimos*** (hereinafter ***Observaçoens***) and others are relevant to the history of Medicine in that territory and even of the so-called "popular pharmacopoeia", that is, curative methods based on the empirical knowledge of the properties of nature elements. Semedo himself added to the Medicine jargon some words of these pharmacopoeia, which are not actually terms, but popular names for plants, infusions and other "household remedies", which could even include blood from different animals, stones, seeds and roots.

At last, it is also important to emphasize that Semedo's proposal intended to present these texts, vocabularies and terminologies in a way to make it accessible to their readers, with special attention for the lower literate "young doctors" of his time, who did not know enough Latin but who could read a text in Portuguese.

For all these reasons, the ***Observaçoens*** of Curvo Semedo are a rich source of terminological information to which Digital Humanities research methods need to be applied.

Semedo's ***Observaçoens*** deal with 101 cases of a wide range of profiles, offering a historical overview of the most common diseases and intercurrences of the time, affecting different population segments: adults, men, women, pregnant women, newborns, young people, the elderly, children, noblemen, peasants or city people. Semedos' work was also chosen because it was not registered in any of the great historical corpora, not even by Mark Davies'Corpus[1], which has 45 million keywords covering a period between 1200 and 1900.

In file format, this scanned book is available for free at Google Books. In addition to this source, for our work on reading, familiarizing with and transcribing the text, it was important to have another complete digital version made from an original. It was available in the Reservation Sector of the Évora Public Library (BPE) in Portugal. Figure 1 below shows this book frontpage from BPE.



OBSERVAÇOENS
MEDICAS
DOUTRINAES
De cem Caſos graviſſimos,
Que em ſerviço da Patria, & das Nações eſtranhas eſcreve em lingua Portugueza, & Latina
JOAM CURVO SEMMEDO,
Cavalleiro profeſſo da Ordem de Chriſto, Familiar do Santo Officio, & Medico da Caſa Real;
OFFERECIDAS
AO ILLUSTRISSIMO SENHOR
RUY DE MOURA TELLES
Arcebiſpo de Braga Primàz das Heſpanhas.

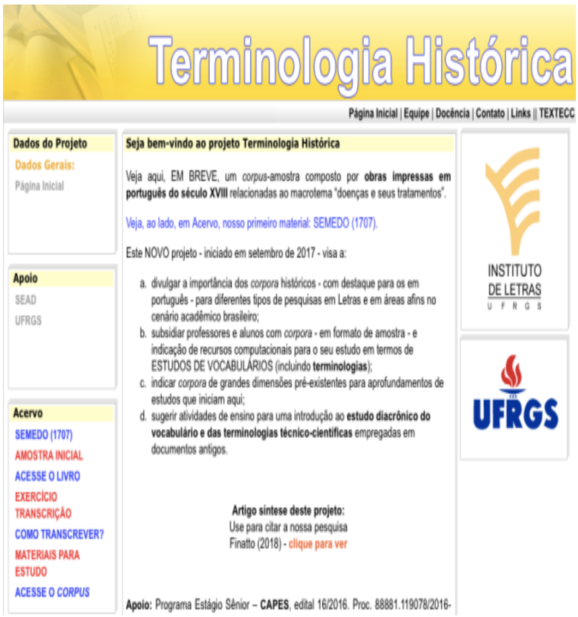
LISBOA,
Na Officina de ANTONIO PEDROZO GALRAM.
Com todas as licenças neceſſarias, & Privilegio Real.
Anno M. DCCVII.

Figure 1: The frontpage of Semedo (1707).
Scanned version by BPE.

The text of this book, as a corpus-sample, will be part of a website specially dedicated to the study of historical lexicology and terminology topics. It is a corpus with printed texts of the 18th century. These materials are integrated to the didactic initiative "Terminologia Histórica", within the scope of the TEXTECC Project www.ufrgs.br/textecc at Universidade Federal do Rio do Sul (UFRGS), Brazil. Texts and other data build an e-learning environment, where simple sets of texts and online tools will be offered for exploration to help studies on the historical terminology and, in particular, on the history of medical terminology in Portuguese. The tools planned for this website are: a word list generator, a word-context generator to search expressions in a given corpus and/or text, and a generator of lists of word groups to show blocks of repeated words (clusters) along a given text or several texts. Figure 2 below shows the front page of the didactic environment and some preliminary activities with Semedo's book. Starting from the left menu, the user has an initial sample of the corpus and some guided transcription exercises. It is also possible for

[1] Website of the Mark Davies Corpus: http://www.corpusdoportugues.org/interface2016.asp.

the user to access Semedo's book according to the scanned version freely offered by Google Books.



Seja bem-vindo ao projeto Terminologia Histórica

Veja aqui, EM BREVE, um *corpus*-amostra composto por **obras impressas em português do século XVIII** relacionadas ao macrotema "doenças e seus tratamentos".

Veja, ao lado, em Acervo, nosso primeiro material: SEMEDO (1707).

Este NOVO projeto - iniciado em setembro de 2017 - visa a:

a. divulgar a importância dos *corpora* históricos - com destaque para os em português - para diferentes tipos de pesquisas em Letras e em áreas afins no cenário acadêmico brasileiro;
b. subsidiar professores e alunos com *corpora* - em formato de amostra - e indicação de recursos computacionais para o seu estudo em termos de ESTUDOS DE VOCABULÁRIOS (incluindo **terminologias**);
c. indicar *corpora* de grandes dimensões pré-existentes para aprofundamentos de estudos que iniciam aqui;
d. sugerir atividades de ensino para uma introdução ao **estudo diacrônico do vocabulário e das terminologias técnico-científicas** empregadas em documentos antigos.

Artigo síntese deste projeto:
Use para citar a nossa pesquisa
Finatto (2018) - clique para ver

Apoio: Programa Estágio Sênior – **CAPES**, edital 16/2016. Proc. 88881.119078/2016-

Figure 2: The draft version of the website already available at http://www.ufrgs.br/textecc/terminologia/

In order to feed this website and its tools, a pilot study was conducted with Semedos' books content. The objective was to verify the advantages and disadvantages of the treatment of a set of texts with the original spelling and with the updated spelling. For this purpose, two free access computational tools for corpora processing were tested, AntConc (Antony, 2014) and TermoStat (Drouin, 2003). It is important to emphasize that both tools, developed by Corpus Linguistics researchers, are not built to deal with ancient texts orthography and old print characters. This means that the above-mentioned tools raise a few problems of philological nature, since, in order to comply with the text features, it is necessary to transcribe them and to prepare digital editions (Crane et al., 2008; Paixão de Sousa 2013a, 2013b). The tools will be very briefly described in the next section.

From Semedos' book only a complete section with 1,317 graphic words considering its spelling was examined. This excerpt, named ***Observaçam XCII*** (pages 528–532), is just one of the 101 that make up the whole book.

In addition, this sample was contrasted with the collection called ***Gazetas Manuscritas*** of the Évora Library (see a part of this in Menezes, 1673), a corpus of ancient journalistic texts (Quaresma, 2016). This is a large set of journalistic texts from the 18th century handwritten in Portuguese. Thus, the ***Gazetas Manuscritas*** [freely translated as "The handwritten News"] was considered as a contrastive reference corpus. In a document with 480,366 characters and 14,832 types (different items), a sample with 85,517 words/tokens was chosen for this contrast. This material is partially available – in a transcribed version – at the Tycho Brahe Corpus (Sousa, 2014): http://www.tycho.iel.unicamp.br/corpus/.

2. The tools for text processing tests

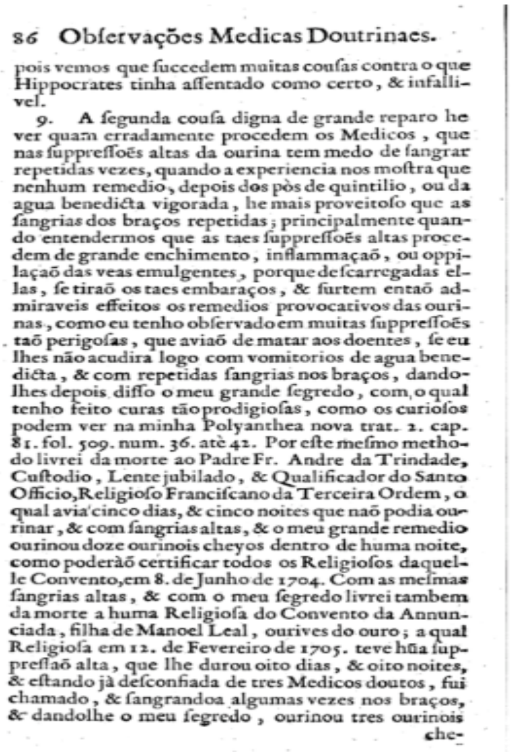
TermoStat receives an input text and returns as a main result a list of candidate terms (CT) derived from the text. A term – or a specific word item – can be either simple (a word) or complex (a sequence of words). Each term receives a score based on the frequency of the term in the analyzed corpus, the corpus of analysis (CA), and its frequency in another pre-processed corpus, a corpus of reference (CR). The Portuguese reference corpus has about 10,000,000 occurrences, which corresponds to approximately 542,000 different forms. It is a non-technical corpus. In our study, the input text can be made by an ancient orthography or an adapted one, but it will be compared with the same modern Portuguese corpus, the CR. The CR is a "resident" part of the TermoStat system for its Portuguese module.

On the other hand, AntConc is a freeware corpus analysis toolkit. This tool is useful for searching words in context and helps us to do different kinds of text analysis. AntConc, for example, allowed us to observe the usage of repeated stock phrases throughout much of the text. With AntConc, we can also make a wordlist of a whole text or texts and compare their frequencies. As TermoStat, AntConc receives, as input, a text file that will be processed. This software identifies each set of text characters which is separated by a blank as a "word" (token). Numbers and punctuation marks used in the text are disregarded. Thus, if we have in the ancient corpus three different forms of a Portuguese ancient word (today: PURGAÇÃO [PURGING, using laxatives]), as PURGAÇAÕ and PURGAÇÃO or PURGAÇAM, the AntConc system will identify them as three different "words". The same will happen with any flexional forms/variants, as plural and singular for Portuguese nouns, as the word MULHER [WOMAN] or MULHERES [WOMEN].

3. Steps of the pilot study

Some initial results of an experiment, only with the above-mentioned Semedos' sample processed by AntConc and TermoStat tools, indicate the advantages of dealing with the old orthographic forms (Gonçalves, 2003). More details are described by Finatto (2018). For an initial test, the performance of these tools was compared in processing the old spelling and the updated spelling. Figure 3 shows a complete page of Semedos' book and illustrates some special examples of problems in handling the orthographic system of this kind of ancient printed material.

As the Figure 3 exemplifies, there is a lot of orthographic challenges to face with our OCR systems and even with the typographical conventions. To support a future large scale better optical character recognition, it will be to use necessary different resources. One option to help us with the tasks of the corpus development with our students is the *eDictor* system, a tool for philological edition and automatic linguistic annotations (Sousa, Kepler and Faria, 2013). We intend to explore this system in the frame of the above cited e-learning environment "Terminologia Histórica". The Version beta 1.0 of the *eDictor* was developed in 2007 (https://www.ime.usp.br/~tycho/participants/psousa/edicto r/presentation/edictor_2007.html), and this first version already contained the core functions of the tool: an XML annotation module, the possibility of XSLT transformation exportation, and a morphosyntactic (Part of Speech) tagging function.



86 Obſervações Medicas Doutrinaes.

pois vemos que ſuccedem muitas couſas contra o que Hippocrates tinha aſſentado como certo, & infallivel.

9. A ſegunda couſa digna de grande reparo he ver quam erradamente procedem os Medicos, que nas ſuppreſſoẽs altas da ourina tem medo de ſangrar repetidas vezes, quando a experiencia nos moſtra que nenhum remedio, depois dos pòs de quintilio, ou da agua benedicta vigorada, he mais proveitoſo que as ſangrias dos braços repetidas; principalmente quando entendermos que as taes ſuppreſſoẽs altas procedem de grande enchimento, inflammaçaõ, ou oppilaçaõ das veas emulgentes, porque deſcarregadas ellas, ſe tiraõ os taes embaraços, & ſurtem entaõ admiraveis effeitos os remedios provocativos das ourinas, como eu tenho obſervado em muitas ſuppreſſoẽs taõ perigoſas, que aviaõ de matar aos doentes, ſe eu lhes não acudira logo com vomitorios de agua benedicta, & com repetidas ſangrias nos braços, dando-lhes depois diſſo o meu grande ſegredo, com o qual tenho feito curas tão prodigioſas, como os curioſos podem ver na minha Polyanthea nova trat. 2. cap. 81. fol. 509. num. 36. atè 42. Por eſte meſmo methodo livrei da morte ao Padre Fr. Andre da Trindade, Cuſtodio, Lente jubilado, & Qualificador do Santo Officio, Religioſo Franciſcano da Terceira Ordem, o qual avia cinco dias, & cinco noites que naõ podia ourinar, & com ſangrias altas, & o meu grande remedio ourinou doze ourinois cheyos dentro de huma noite, como poderàõ certificar todos os Religioſos daquelle Convento, em 8. de Junho de 1704. Com as meſmas ſangrias altas, & com o meu ſegredo livrei tambem da morte a huma Religioſa do Convento da Annunciada, filha de Manoel Leal, ourives do ouro; a qual Religioſa em 12. de Fevereiro de 1705. teve hũa ſuppreſſaõ alta, que lhe durou oito dias, & oito noites, & eſtando jà deſconfiada de tres Medicos doutos, fui chamado, & ſangrandoa algumas vezes nos braços, & dandolhe o meu ſegredo, ourinou tres ourinois che-

Figure 3: The page 86 of the Semedo's book ***Observaçoens***

A second round of testing involved the comparison between the ***Gazetas Manuscritas*** sample and ***Observaçam XCII***. These two steps, dealed only with the TermoStat and AntConc systems, are summarized below.

3.1. The first step

With the AntConc tool, a list of all the words from the text of ***Observaçam XCII*** according to the old original spelling was produced. It was a list with 1,317 words (tokens), where 536 were different word forms (types). In the proportion between types-tokens, with which the variety of the vocabulary of the text is estimated, the segment showed 40% of vocabulary variety and a set of 355 words of single occurrence (called *Hapax legomena*).

Below, we have an example of Semedos' book – with the ancient orthography – with entries of the words CAMARAS [today: EPISODES OF DIARRHEA], FEBRE [FEVER] and SANGRIAS/SANGRASSEM [related to BLEEDINGS]. The emphasis in bold does not exist in the original text:

> Em 14 de Outubro de 1702. fuy chamado para visitar a senhora Dona Violante Casimira Saldanha a quem Deos tinha feito merce de dar hum filho desejado com ancja & conseguido com grande alegria; mas como as felicidades temporaes sejaó mui pensionadas, & cheyas de sobresaltos, ao gosto do feliz nascimento se seguio o temor, & tristeza, com humas **camaras**, **febre**, & falta da descarga devida ao puerperio: perturbàraõ muito estes symptomas naó só aos pays da recem nascida criança, mas aos patentes, &familiares da casa, porque tinhaõ ouvido dizer, que **camaras** sobre parto eraõ muito para temidas: para se desatar este no Gordonio, naó obstante que na visita da tarde tinha dado ordem a que pela manháa **sangrassem** a dita senhora, o naó quizeraó fazer sem que eu a visitasse primeiro, porque entendèraó que os cursos era hum grande impedimento para a **sangria**;

Then, with the TermoStat, tool described above, we have contrasted the frequencies and word distributions used in the old text with the word frequencies of its collection of texts with current Portuguese spelling. With TermoStat, we would argue, in thesis, the major peculiarities of ***Observaçam XCII*** regarding the statistical distribution of a specific vocabulary of the past in relation to a current and broader vocabulary.

The test with AntConc was productive. That is, it has met the challenge of recognizing the words in their original (not modernised spelling of our 18th century medical text, even though it was not developed for this purpose. It is worth mentioning that it handled well the diversity and frequency of graphic forms, especially with the measure of the proportional variety of vocabulary (measure known as 'Type-Token Ratio') and indication of the proportion of words of single occurrence.

On the other hand, TermoStat worked by identifying and categorizing "words" by morphological classes, then contrasting the vocabulary of the segment from Semedos' with a large collection of current texts. The results with this tool require further studies on its modes of functioning and performance with ancient texts. It is necessary to consider what this system does, "its statistical guidelines", with the classification of invalid spellings and how it assesses "errors" – the unknown words – that are not recognized by their morphosyntactic parser. Although the contrast allowed by TermoStat is between the words of the unique old text versus a large number of modern texts, we believe that it could be used for some purposes, even if the old-modern comparison can be considered unequal and problematic. As TermoStat pointed out, the words SANGRIA [BLEEDING], MEDICO [DOCTOR] and PURGAÇAÕ [PURGING, using laxatives] are the most typical items with the ancient text. For the modern one, it showed the items PURGAÇÃO, SANGRIA and PURGAÇÃO LOQUIAL [CHILDBIRTH'S PURGING].

3.2. The second step

For a second set of tests we dealt only with texts in the old orthography version and only with the TermoStat tools. As the tool system showed, the main words of Semedos' ***Observaçam XCII*** are SANGRIA, FEBRE [FEVER], PURGA and MEDICO. These words also appear in the ***Gazetas Manuscritas*** text, but not with the highest frequency, as would be expected of a non-specific corpus of Medicine.

On the other hand, if we consider Semedo's entire book (1707), as a medical handbook, there is only 01 occurrence of the item BEXIGAS (plural) [WOUNDS CAUSED BY SMALLPOX or SMALLPOX, the disease itself] − along 635 pages, but there are only 10 occurrences for BEXIGA [URINARY BLADDER], word in the singular.

In another contextual frame, designed by the corpus of ***Gazetas Manuscritas***, considered as an ancient journalistic text, we can count 29 occurrences of the word BEXIGAS [in SMALLPOX sense]. Below, we have an example – with the ancient orthography – of an entry of the word BEXIGAS and SANGRIAS/SANGRALO [related to BLEEDINGS]. The emphasis in bold does not exist in the original text:

> Com grande susto esteve a corte em hũa grande febre do Prinçipe e na contenda dos medicos duvidando hũns, e querendo outros **sangralo** prevaleço a opinião de que não fizeçe este remedio, e secou a febre de todos os sintomas sahindo hũa espeçie de **bexigas**, tão benigna que senão fosse preçizo á fineza da Prinçeza bem podião chamar-se com outro epiteto, houve preçes, e assistencia dos reys, e de toda a corte foi, como mereçia couza tão justa. A Prinçeza ja se levanta, o Sr. Jnfante D. Carllos melhorou com as **sangrias.**

Table 1 below shows a comparison of the top-10 nominal expressions in examined Semedo's book segment ***Observaçam XCII*** and in ***Gazetas Manuscritas***.

Noun ***Observaçam XCII***	Frequency	Noun ***Gazetas Manuscritas***	Frequency
Parto	15	rey	1004
Sangria	14	sra	344
Febre	11	conde	309
Natureza	9	antonio	188
Purga	5	duque	183
Humor	5	infante	142
Medico	4	caza	137
Galeno	4	Sr	136
Purgaçaõ	4	annos	134
Puerperio	3	diario	101

Table 1: The comparison of the top-10 nominal expressions in Semedo's book ***Observaçam XCII*** and in ***Gazetas Manuscritas***

As we can see the top-10 nominal expressions are totally distinct, and they reflect the "textual genres" of both texts. In Semedos'book segment the most frequent item is PARTO [**childbirth**] while in ***Gazetas*** the top lexical item is REY [**the king**]. Indeed, the textual genre not only determines certain terminology characteristics, but the textual genre is also determined by certain factors. As Santos and Costa (2015: 160) point out "texts are the result of social and discursive activities" and "when considered from this perspective, texts are not only linguistics artefacts, but also the product of social, cultural and ideological factors".

It is also interesting to compare the way nominal expressions are created in both textual genres: "noun" is the most frequent word class, but in the ***Gazetas Manuscritas*** there is a high frequency of 'noun + noun' (36.0%) and in Semedo's this represents only 5.0%. Moreover, in Semedo's book the use of multi-word complex nominal expressions including adjectives has a higher frequency than in ***Gazetas Manuscritas*** (25.0% versus 2.0%). This fact suggests the need for more complex nominal structures to describe medical situations in comparison with a general domain text.

Table 2 and Table 3 show examples of the most frequent nominal expressions in Semedos' ***Observaçam XCII*** and in ***Gazetas Manuscritas***, respectively.

Semedos' *Observaçam XCII*	%	Examples
Noun (N)	62	parto, sangria, febre, natureza, mulher, falta, caso, perigo, humor, pé
N+prep+N	20	purgação de parto, falta de purgação, sangria de pé, via de purga, sangria de pès, inchação de pé, vizinho de parto, sinaes de crueza, enchimento de sangue, natureza de humor
N+adj	10	caso semelhante, purgação loquial, purgaçã principiante, humor cacochymicos, sentença definitivo, perigo urgente, varão douto, filho desejado, caminho errado
N+N	5	felicidade temporaes, reynavão soro, valerio martins
N+prep+N+adj	5	falta de purgação mensal, embaraço a purgaçã principiante, falta de purgação loquial

Table 2: Distribution of nominal expressions in Semedo's segment book ***Observaçam XCII***

Gazetas Manuscritas	%	Examples
N	44	rey, conde, filho, dia, sñra, cruzado, antonio, duque
N+N	36	el rey, d. maria, d. antonio, d. anna, s. francisco, d. manoel, d. joão, d. lourenço, campo grande, del rey

N+prep+N	15	filho de conde, duque de cadaval, secretario de estado, conde de assumar, joão de saldanha, rey de frança, cardeal de cunha, duque de aveyro, marques de alegrete, marques de abrantes
N+adj	2	monteiro mor, filho unico, diamante brilhante, sargento mor, camareira mor
N+N+N	2	jnfante d. francisco, jnfante d. carllos, jnfante d. antonio, jnfante d. carlos, el rey catholico, d. anna joaquina, assumar d. pedro, el rey stanislao, jnfante d. manoel, jnfanta d. francisca

Table 3: Distribution of nominal expressions in ***Gazetas Manuscritas***

4. Initial results: some considerations

As a result of our initial tests with the selected tools, we want to emphasize the importance to have historical corpora - especially in Portuguese - for different kinds of researches in Lexicology, Terminology and related areas as well as indicate the importance of diachronic studies of vocabulary and medical terminologies in ancient documents. However, besides the computational dimension highlighted here, an explicative philological-historical component should be included. This component, of course, is something that needs to be included in the online learning environment in which the corpus and computational tools to explore it will be offered.

Words identified as frequent and as "terminologies" by the computational tools or by a human reader have a source and a history. These ancient terminologies appear in Semedos' medical handbook as a particular conception of the functions of the human body. Thus, the vocabulary profile of the text manifests an epistemology of the late 17th and early 18th century. It is also concerned to the Semedos' scientific points of view before the Linnaean taxonomy and this scientific revolution to mankind. This prism related to these documental corpora is relevant to understand the language and terminology of the time, besides the automatic and comparative data. This shows a frame of elements that should be considered beyond quantitative evidences.

In addition, Semedo's proposal that intended to present these type of Medicine language, vocabularies and terminologies in a way to make it accessible to their readers serves as a good inspiration for today's researchers on the topic "plain language" for lower literate audiences.

5. Acknowledgments

We register here our special acknowledgments to the National Council for Scientific and Technological Development (CNPq), trough the financial lines *PQ* and *Edital Universal*, and to The Coordination for the Improvement of University Personnel (CAPES) in the scope of the *Estágio Sênior* CAPES, Proc. 88881.119078/2016-01, these are the foremost agencies to promote and support academic research in Brazil. We also thank for special support to the Interdisciplinary Center for History, Culture and Societies, University of Évora (CIDEHUS-UE/FCT), Portugal.

6. References

Laurence Anthony. 2014. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan, Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

Ana Paula Banza and Maria Filomena Gonçalves. 2018. *Roteiro de História da Língua Portuguesa*. Évora, UNESCO Chair in Intangible Heritage and Traditional Know-How: Linking Heritage. http://www.catedra.uevora.pt/unesco/index.php/unesco/Investigacao/Publications-et-al/Books/Roteiro-de-Historia-da-Lingua-Portuguesa.

David Berry and Anders Fagerjord. 2017. *Digital Humanities: knowledge and critique in digital Age*. Cambridge, UK/Malden, Ms, Polity Press.

Rafael Bluteau. 1712-1728. *Vocabulario portuguez e latino* (...). Vol. 1-4 (1712-1713), Coimbra, Colegio das Artes; Vol. 5-8 (171-1721), Lisboa, Pascoal da Sylva; *Supplemento ao Vocabulario Portuguez e Latino*, Vol. 1, Lisboa, Joseph Antonio da Silva; Vol. 2 (1728), Lisboa, Patriarchal Officina da Musica.

Ivo Castro. 2006. *Introdução à história do Português*. 2ª ed. revista e ampliada. Lisboa, Edições Colibri.

Claudia Claridge. 2008. Historical corpora. In: A. Lüdeling and M. Kytö ed., *Corpus linguistics*: *an international handbook*. Berlin/New York: Walter de Gruyter, Handbooks of Linguistics and Communication Science/Handbücher zur Sprach und Kommunikationswissenschaft 29.1-2.

Gregory Crane, David Bamman and Alison Jones. 2008. ePhilology: when the books talk to their readers. In: S. Schreibman and R. Siemens eds., *A Companion to Digital Literary Studies*. Oxford, Blackwell.

Matthew James Driscoll and Elena Pierazzo eds. 2016. *Digital scholarly editing: theories and practices*. Digital Humanities Series, Vol. 4. Open Book Publishers.

Patrick Drouin. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology*. 9(1):99–117. doi:10.1075/term.9.1.06dro.

Maria José Bocorny Finatto. 2018. Corpus-amostra português do século XVIII: textos antigos de Medicina em atividades de ensino e pesquisa. *Domínios de Linguagem,*12(1):434−464.doi:http://dx.doi.org/10.14393/DL33-v12n1a2018-15.

Heather Froehlich. 2015. *Tutorial. Corpus Analysis with Antconc.*https://programminghistorian.org/lessons/corpus-analysis-with-antconc#introduction.

Júnia Ferreira Furtado. 2008. Tropical empiricism: making medical knowledge in colonial Brazil. In: James Delbourgo and Nicholas Dew ed., *Science and empire in the Atlantic world*, pages 127–152. New York/London, Routledge.

Maria Filomena Gonçalves. 2003. *As ideias ortográficas em Portugal: de Madureira Feijó a Gonçalves Viana (1734-1911)*. Lisboa, Fundação Calouste Gulbenkian/Fundação para a Ciência e Tecnologia.

Hendrick J. Kockaert and Friede Steurs eds. 2015. *Handbook of Terminology*, Vol. 1. Amsterdam/Philadelphia, John Benjamins.

Merja Kytö. 2011. Corpora and historical linguistics. *Revista Brasileira de Linguística Aplicada*, 11(2): 417-457. https://dx.doi.org/10.1590/S1984-63982011000200007.

Rita Marquilhas and Iris Hendrickx. 2016. Avanços nas humanidades digitais. In: A. Maria Martins and Ernestina Carrilho eds., *Manual de Linguística Portuguesa*. MRL 16, pages 252–277. Berlin/Boston, De Gruyter,

Francisco Xavier de Menezes. 1673. *Gazetas manuscritas da Biblioteca de Évora*. Vol. I (1729-1731). http://www.tycho.iel.unicamp.br/corpus/texts/xml/m_008.

Maria Clara Paixão de Sousa. 2013a. A Filologia Digital em Língua Portuguesa: Alguns caminhos. In: Maria Filomena Gonçalves and Ana Paula Banza coord., *Património textual e Humanidades Digitais. Da antiga à nova Filologia*, pages113-138. Évora, Centro Interdisciplinar de História, Culturas e Sociedades da Universidade de Évora (CIDEHUS)/ Fundação para a Ciência e Tecnologia (FCT). http://books.openedition.org/cidehus/1089.

Maria Clara Paixão de Sousa. 2013b. Texto digital: uma perspectiva material. *Revista da ANPOLL*, 35: 17–60.

Maria Clara Paixão de Sousa. 2014. Tycho Brahe: contribuições para as humanidades digitais no Brasil. *Filologia e Linguística Portuguesa*, 16(nº esp. dez.): 53–93.

Maria Clara Paixão de Sousa; Fábio Natanael Kepler, Pablo Picasso Feliciano de Faria. 2013. *e-Dictor* (Version 1.0 Beta 10). Retrieved from https://edictor.net/download.

Paulo Quaresma. 2013. Análise linguística de documentos da Biblioteca Pública de Évora Uma abordagem informática. In: Maria Filomena Gonçalves and Ana Paula Banza coord., *Património Textual e Humanidades Digitais. Da antiga à nova Filologia,* pages 139-155. Évora, CIDEHUS. https://books.openedition.org/cidehus/1091.

Cláudia Santos and Rute Costa. 2015. Domain specificity: semasiological and onomasiological knowledge representation. In: H. J. Kockaert and F. Steurs eds., *Handbook of Terminology*, Vol. 1, pages 153–179. Amsterdam/Philadelphia, John Benjamins.

João Curvo Semedo. 1707. *Observaçoens medicas doutrinaes de cem casos graviissimos, que em serviço da pátria, & das nações estranhas escreve em língua portugueza, & latina*. Lisboa, Officina de Antonio Pedrozo Galram.

Telmo Verdelho. 1998. Terminologias na língua portuguesa (perspectiva histórica). In: Jenny Brumme ed., *La història dels llenguatges iberoromànics d'especialitat (segles XVII-XIX),* pages 98–131. Barcelona, Universitat Pompeu Fabra/Institut Universitari de Lingüistica Aplicada. http://clp.dlc.ua.pt/Publicacoes/Terminologias_lingua_portuguesa.pdf.

Interaktivna karta slovenskih narečnih besedil

Alenka Kavčič,* Ivan Lovrić,* Vera Smole**

* Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
alenka.kavcic@fri.uni-lj.si, ivan@lovric.si
** Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
vera.smole@ff.uni-lj.si

Povzetek

Sedem narečnih skupin, ki združujejo sedemintrideset narečij, uvršča slovenski jezik med narečno najbolj razčlenjene evropske jezike. Da to bogastvo jezika ne bi ostalo skrito in dostopno le jezikoslovcem in narečjeslovcem, smo izdelali spletno interaktivno karto za slovenska narečna besedila. Aplikacija omogoča prikaz posameznih krajev s posnetim narečnim govorom na interaktivnem zemljevidu, z izbiro kraja pa lahko poslušamo posnetek narečnega govora, si ogledamo njegovo fonetično transkripcijo in poknjižitev, dodana pa je tudi diahrona analiza govora. Aplikacija vključuje tudi skrbniški del, ki prijavljenim uporabnikom omogoča urejanje obstoječih vsebin in dodajanje novih vnosov. Z željo po uporabniku prijaznem in enostavnem uporabniškem vmesniku smo pri izdelavi aplikacije uporabili uveljavljena spletna orodja in rešitve.

Interactive map of Slovenian dialectal texts

Seven dialect groups combining 37 dialects, rank Slovenian language among the European languages with the greatest dialectical diversity. To offer this richness of the language, now available mostly to linguists and dialectologists, to wider public, an interactive web-based map of Slovenian dialectal texts has been developed. The application shows individual places with the recorded dialectal texts on the interactive map, while selecting a specific place enables listening to the sound recording of the dialectal speech, displays its phonetic transcription and translation to literary language as well as the diachronic analysis of the speech. The application includes an administrative part that enables logged-in users editing of the existing contents as well as adding new entries. Aiming to achieve user friendly and easy to use user interface, the application was developed using conventional web tools and solutions.

1. Uvod

Slovenščina je narečno zelo bogat jezik. S 37 narečji, ki jih združujemo v 7 narečnih skupin, se slovenski jezik uvršča med najbolj razčlenjene evropske jezike.

Čeprav se narečjeslovje kot samostojna veda razvija že od druge polovice 19. stoletja (Toporišič, 1987), so raziskovalni rezultati še vedno dostopni pretežno v tiskanih izdajah monografij, (zbirk) člankov, narečnih slovarjev in drugih tematskih publikacij, pri čemer so izjema Slovenski lingvistični atlas (1 in 2) in nekateri narečni slovarji, vključeni v spletni portal Fran.[1] Za večjo prepoznavnost in približanje tematike tudi mlajšim, digitalnim generacijam je zato nujno potrebno izkoristiti možnosti sodobnih informacijsko-komunikacijskih tehnologij, predvsem interneta in svetovnega spleta kot glavnega medija za dostop do informacij.

V tem prispevku bomo opisali spletno aplikacijo za prikaz interaktivne karte slovenskih narečnih besedil, ki je nastala v okviru diplomskega dela na Fakulteti za računalništvo in informatiko Univerze v Ljubljani (Lovrić, 2018). Aplikacija za podane kraje, označene na karti, omogoča predvajanje zvočnih posnetkov narečnih besedil ter prikaz njihove fonetične transkripcije, poknjižitve (zapis z grafemi in fonetiko knjižnega jezika ter oblikoslovnimi in leksičnimi knjižnimi ustreznicami v pomenskih oklepajih) in analize večine besedil. Gradivo in analize so rezultat raziskovalnega dela pri dveh predmetih na Oddelku za slovenistiko Filozofske fakultete Univerze v Ljubljani (Smole in Horvat, 2016; Smole, 2016), vsa besedila pa so na temo stara kmečka hiša (prostori in oprema v njih). Aplikacija vključuje tudi administrativni del, ki omogoča dodajanje novih vsebin in urejanje obstoječih.

2. Sorodni pristopi

Za slovenski jezik nismo našli nobene podobne aplikacije, tematsko še najbližja je elektronska oblika Slovenskega lingvističnega atlasa (e-SLA[2]), ki pa je namenjen bolj narečjeslovcem kot širši publiki.

Nekoliko bližji naši ideji spletne aplikacije je Zemljevid narečij bolgarskega jezika,[3] ki ga je izdelal Oddelek za dialektologijo in jezikovno geografijo Inštituta za bolgarski jezik in je prvi elektronski interaktivni zemljevid bolgarskih narečij. Zemljevid prikazuje glavne narečne skupine celotnega jezikovnega ozemlja in predstavlja vsa glavna narečja in njihove meje. Po zemljevidu so razporejene ikone, ki ob kliku prikažejo krajšo analizo govora (ikona knjige) ali omogočijo predvajanje zvočnega zapisa narečnega govora (ikona zvočnika). Aplikacija omogoča tudi dopolnjevanje karte z novimi zvočnimi zapisi.

Ena bolj zanimivih obstoječih rešitev je tudi Zvočna karta naglasov in narečij Velike Britanije,[4] ki je dostopna na spletnih straneh The British Library. Za prikaz podatkov uporablja zemljevide Google Maps, a na karti so dostopni

[1] Fran, Slovarji Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, dostopno na http://www.fran.si/.

[2] Slovenski lingvistični atlas v obliki html, dostopno na http://sla.zrc-sazu.si/eSLA/Zavihki_na_eSLA_JS.html.

[3] Zemljevid narečij bolgarskega jezika, dostopno na http://ibl.bas.bg/bulgarian_dialects/.

[4] Zvočna karta naglasov in narečij Velike Britanije, dostopno na https://sounds.bl.uk/Sound-Maps/Accents-and-Dialects.

le zvočni zapisi brez besedil ali analiz govora. Uporabniška izkušnja ni najboljša, saj se predvajalnik zvoka ne odpre v modalnem oknu, ampak uporabnika ob kliku na zvočni zapis preusmeri na drugo spletno stran.

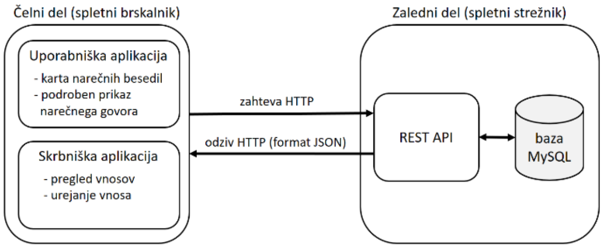
Podobno temelji na Google Maps tudi Zvočni atlas hrvaških govorov,[5] a se njegova ideja nekoliko razlikuje od naše aplikacije, saj gre za glasoslovni atlas, katerega namen je predstaviti določene foneme v kratkem besedilnem okolju. Na zemljevidu prikazane ikone za zvočne posnetke primerov narečnih govorov tako omogočajo poslušanje vedno le vnaprej izbranega stavka z določeno fonološko strukturo.

3. Interaktivna karta slovenskih narečnih besedil

V tem poglavju bomo opisali izdelano interaktivno aplikacijo, začenši z zasnovo in zgradbo aplikacije, navedbo uporabljenih spletnih tehnologij ter opisom priprave karte narečnih besedil v ustreznem formatu.

3.1. Zasnova in zgradba aplikacije

Spletno aplikacijo sestavljata čelni in zaledni del, kot je to shematično prikazano na sliki 1.



Slika 1: Zgradba aplikacije.

Čelni del se izvaja v spletnem brskalniku in predstavlja tisti del aplikacije, ki ga vidi uporabnik (neprijavljen uporabnik pri ogledu narečne karte in govora posameznega kraja ali prijavljen skrbnik pri pregledu in urejanju vnosov). Zajema tako uporabniško kot tudi skrbniško aplikacijo (obe sta opisani v nadaljevanju). Izdelava tega dela aplikacije je temeljila na sodobnih spletnih tehnologijah in ogrodjih, kot so HTML5, CSS, JavaScript, AngularJS, Bootstrap, za prikaz karte pa smo uporabili knjižnico Leaflet, ki temelji na prosto dostopnih zemljevidih OpenStreetMap.

Zaledni del aplikacije se izvaja na spletnem strežniku in zajema vmesnik za dostop do podatkovne baze MySql, v kateri so shranjeni vsi podatki. Vmesnik temelji na arhitekturi REST in je v celoti napisan v jeziku PHP. Vmesnik tako sprejema zahteve HTTP preko metod GET, POST, PUT in DELETE ter vrača odzive operacij v formatu JSON. Dostop do podatkov v bazi je možen le preko vmesnika s pomočjo zahtev HTTP, kar velja tako za uporabniško kot tudi za skrbniško aplikacijo.

3.2. Priprava interaktivne karte narečij

Pri izdelavi aplikacije smo posebno pozornost namenili enostavni in intuitivni uporabi karte, saj je aplikacija namenjena širšemu krogu končnih uporabnikov, ki niso vedno vešči uporabe računalniških programov. Sprva smo nameravali našo karto graditi na dobro poznanem in pogosto uporabljanem zemljevidu Google Maps (z uporabo vmesnika Google Maps API), preko katerega bi izrisali slovenska narečna območja. Vendar smo se zaradi zaprtosti in posledično manjše fleksibilnosti vmesnika Google Maps API na koncu odločili za uporabo odprtokodne Javascriptove knjižnice Leaflet in prosto dostopnih zemljevidov OpenStreetMap.

Osnova za izdelavo zemljevida z vrisanimi narečnimi območji je bila Karta slovenskih narečij Tineta Logarja in Jakoba Riglerja (1983), ki je nastala na osnovi Ramovševe (1931) *Dialektološke karte slovenskega jezika* in bila še večkrat dopolnjevana. Najprej smo na osnovi Karte slovenskih narečij izdelali vektorski izris zemljevida slovenskih narečij ter ga shranili v formatu svg. Nato smo vektorsko grafiko še geokodirali, in sicer s pretvorbo v format GeoJSON, ki vsebuje geografske podatke za vsak grafični element karte. Na koncu smo geokodirano karto uvozili v aplikacijo s pomočjo knjižnice za izdelavo interaktivnih spletnih zemljevidov Leaflet.

Tako pripravljena karta nam omogoča več interaktivnosti v aplikaciji, saj lahko dinamično spreminjamo oblikovne stile narečnih območij in mej med njimi, odziva pa se tudi na uporabniške vnose in premike z miško, kot je npr. lebdenje z miškinim kazalcem nad posameznim narečjem in podobno.

3.3. Uporabniška aplikacija

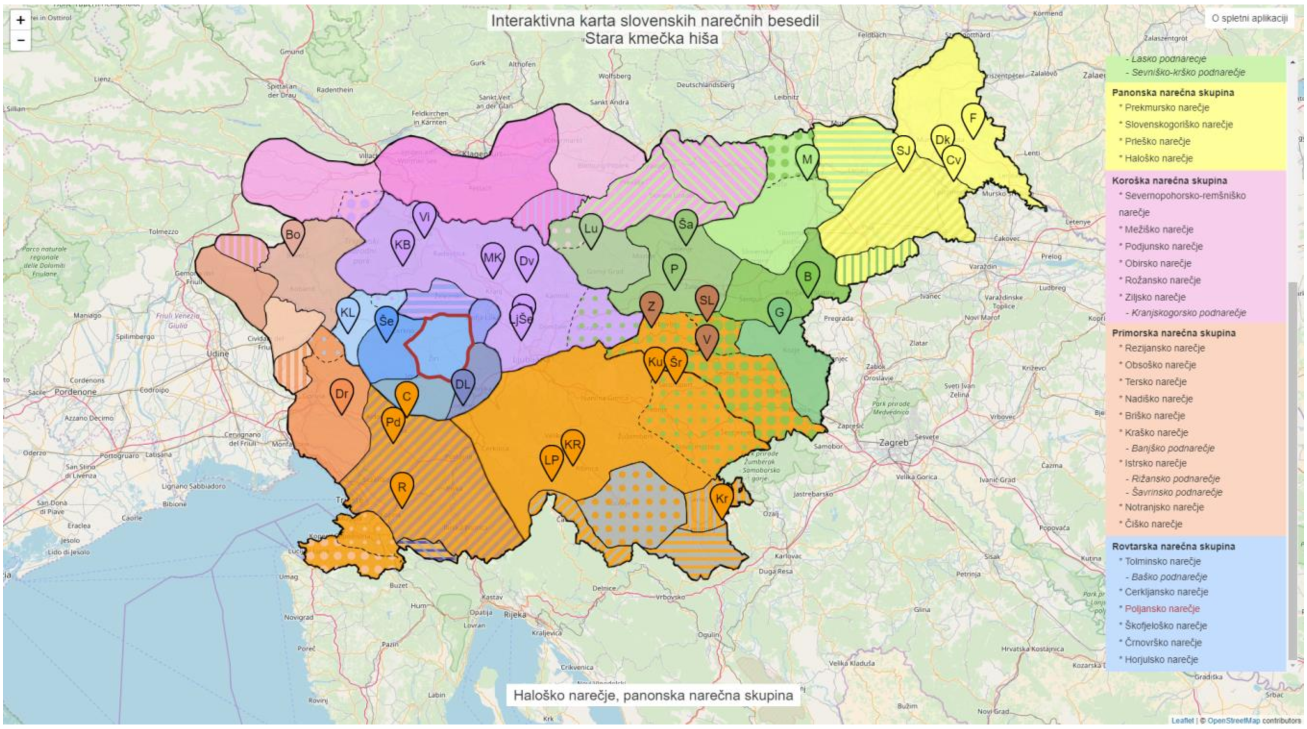
Uporabniški del aplikacije prikazuje spletni zemljevid,[6] na katerem so barvno označena vsa slovenska narečja, podnarečja in tudi narečne skupine. Ob zemljevidu je tudi legenda z navedenimi narečji in podnarečji za vsako od sedmih narečnih skupin.

Uporabniški pogled na karto je prikazan na sliki 2. Na zemljevidu so označena narečna območja, kjer je vsaka narečna skupina označena z določeno barvo, posamezna narečja in podnarečja pa še z dodatnimi grafičnimi simboli, pikami ali poševnimi črtami, ki ponazarjajo vplive drugih (pod)narečij. Karta namreč z izbranimi barvami, šrafurami in vzorci vsebuje tudi vizualne informacije o tem, kako se na nekaterih območjih prepletajo narečja in podnarečja.

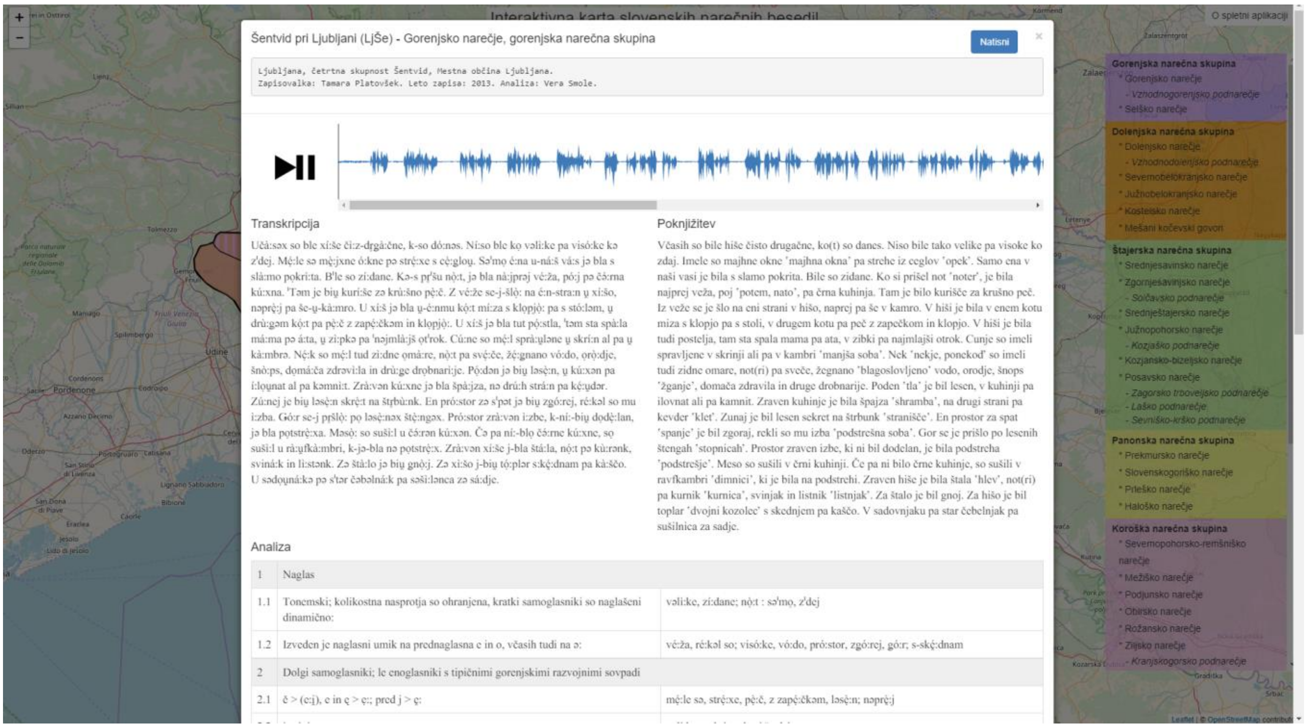
Na desni strani je prikazana tudi legenda z izpisanimi vsemi narečji in podnarečji (poševni tisk), ki so združena v narečne skupine (krepki tisk). S prehodom z miško čez legendo se na karti označi ustrezno narečno območje (primer na sliki 2 prikazuje označeno poljansko narečje, ki je v legendi izpisano rdeče, na karti pa označeno z debelejšo rdečo obrobo). Aplikacija omogoča tudi poljubno približevanje in premikanje po zemljevidu. To je posebej praktično v primeru, ko imamo na manjšem področju označenih več krajev z narečnimi posnetki in s približevanjem lažje ločimo med posameznimi posnetki.

[5] Zvočni atlas hrvaških govorov, dostopno na http://hrvatski-zvucni-atlas.com/.

[6] Interaktivna karta slovenskih narečnih besedil je dostopna na http://narecja.si.



Slika 2: Uporabniški del interaktivne karte.


Interaktivna karta slovenskih narečnih besedil
Natisni
×
O spletni aplikaciji
Gorenjska narečna skupina
* Gorenjsko narečje
- Vzhodnogorenjsko podnarečje
* Selško narečje
Dolenjska narečna skupina
* Dolenjsko narečje
- Vzhodnodolenjsko podnarečje
* Severnobelokranjsko narečje
* Južnobelokranjsko narečje
* Kostelsko narečje
* Mešani kočevski govori
Štajerska narečna skupina
* Srednjesavinjsko narečje
* Zgornjesavinjsko narečje
- Solčavsko podnarečje
* Srednještajersko narečje
* Južnopohorsko narečje
- Kozjaško podnarečje
* Kozjansko-bizeljsko narečje
* Posavsko narečje
- Zagorsko-trboveljsko podnarečje
- Laško podnarečje
- Sevniško-krško podnarečje
Panonska narečna skupina
* Prekmursko narečje
* Slovenskogoriško narečje
* Prleško narečje
* Haloško narečje
Koroška narečna skupina
* Severnopohorsko-remšniško narečje
* Mežiško narečje
* Podjunsko narečje
* Obirsko narečje
* Rožansko narečje
* Ziljsko narečje
- Kranjskogorsko podnarečje
Leaflet | © OpenStreetMap contributors

Šentvid pri Ljubljani (LjŠe) - Gorenjsko narečje, gorenjska narečna skupina

```
Ljubljana, četrtna skupnost Šentvid, Mestna občina Ljubljana.
Zapisovalka: Tamara Platovšek. Leto zapisa: 2013. Analiza: Vera Smole.
```

Transkripcija

Učá:səx so ble xí:še či:z-dr̥gà:čne, k-so dó:nəs. Ní:so ble kọ vəli:ke pa visó:ke kə z'dej. Mẹ́:le sə mẹ̀:jxne ó:kne pə strẹ́:xe s cẹ̀:glou̯. Sə'mọ é:na u-ná:š vá:s jə bla s slà:mo pọkri:ta. B'le so zí:dane. Kə-s pr̥'šu nọ̀:t, jə bla nà:jprəj vé:ža, pó:j pə čə̀:rna kú:xna. 'Təm je biu̯ kurí:še zə krù:šno pẹ̀:č. Z vé:že se-j-šlọ̀: na é:n-stra:n u̯ xí:šo, nəprẹ̀:j pa še-u̯-kà:mro. U xí:š jə bla u̯-é:nmu kọ̀:t mí:za s klọpjọ̀: pa s stó:ləm, u̯ drù:gəm kọ̀:t pa pẹ̀:č z zapẹ́:čkəm in klọpjọ̀:. U xí:š jə bla tut pọ́:stla, 'təm sta spà:la má:ma pə á:ta, u̯ zì:pkə pa 'nəjmlà:jš ọt'rok. Cú:ne so mẹ́:l sprà:u̯ləne u̯ skrí:n al pa u̯ kà:mbrə. Nẹ́:k so mẹ́:l tud zì:dne ọmá:re, nọ̀:t pa svẹ́:če, žẹ́:gnano vó:do, ọrọ̀:dje, šnò:ps, dọmá:ča zdrəvì:la in drù:ge drọbnari:je. Pọ́:dən jə biu̯ ləsẹ̀:n, u̯ kú:xən pa í:lọu̯nat al pa kəmni:t. Zrà:vən kú:xne jə bla špà:jza, nə drú:h strá:n pa kẹ́:u̯dər. Zú:nej je biu̯ ləsẹ̀:n skrẹ̀:t na štr̥bù:nk. En pró:stor zə s'pat jə biu̯ zgó:rej, ré:kəl so mu i:zba. Gó:r se-j pr̥šlọ̀: pọ ləsẹ̀:nəx štẹ̀:ngəx. Pró:stor zrà:vən i:zbe, k-ní:-biu̯ dọdẹ̀:lan, jə bla pọtstrẹ̀:xa. Məsọ̀: so suši:l u čə́:rən kú:xən. Čə pa ní:-blọ čə́:rne kú:xne, sọ suši:l u rà:u̯fkà:mbri, k-jə-bla nə pọtstrẹ̀:x. Zrà:vən xí:še j-bla štá:la, nọ́:t pə kù:rənk, sviná:k in li:stənk. Zə štà:lo jə biu̯ gnọ̀:j. Zə xi:šo j-biu̯ tó:plər s:kẹ́:dnam pa kà:ščo. U sədọu̯ná:kə pə s'tər čəbəlná:k pa səši:lənca zə sá:dje.

Poknjižitev

Včasih so bile hiše čisto drugačne, ko(t) so danes. Niso bile tako velike pa visoke ko zdaj. Imele so majhne okne 'majhna okna' pa strehe iz ceglov 'opek'. Samo ena v naši vasi je bila s slamo pokrita. Bile so zidane. Ko si prišel not 'noter', je bila najprej veža, poj 'potem, nato', pa črna kuhinja. Tam je bilo kurišče za krušno peč. Iz veže se je šlo na eni strani v hišo, naprej pa še v kamro. V hiši je bila v enem kotu miza s klopjo pa s stoli, v drugem kotu pa peč z zapečkom in klopjo. V hiši je bila tudi postelja, tam sta spala mama pa ata, v zibki pa najmlajši otrok. Cunje so imeli spravljene v skrinji ali pa v kambri 'manjša soba'. Nek 'nekje, ponekod' so imeli tudi zidne omare, not(ri) pa sveče, žegnano 'blagoslovljeno' vodo, orodje, šnops 'žganje', domača zdravila in druge drobnarije. Poden 'tla' je bil lesen, v kuhinji pa ilovnat ali pa kamnit. Zraven kuhinje je bila špajza 'shramba', na drugi strani pa kevder 'klet'. Zunaj je bil lesen sekret na štrbunk 'stranišče'. En prostor za spat 'spanje' je bil zgoraj, rekli so mu izba 'podstrešna soba'. Gor se je prišlo po lesenih štengah 'stopnicah'. Prostor zraven izbe, ki ni bil dodelan, je bila podstreha 'podstrešje'. Meso so sušili v črni kuhinji. Če pa ni bilo črne kuhinje, so sušili v ravfkambri 'dimnici', ki je bila na podstrehi. Zraven hiše je bila štala 'hlev', not(ri) pa kurnik 'kurnica', svinjak in listnik 'listnjak'. Za štalo je bil gnoj. Za hišo je bil toplar 'dvojni kozolec' s skednjem pa kaščo. V sadovnjaku pa star čebelnjak pa sušilnica za sadje.

Analiza

1	Naglas	
1.1	Tonemski; kolikostna nasprotja so ohranjena, kratki samoglasniki so naglašeni dinamično:	vəli:ke, zí:dane; nọ̀:t : sə'mọ, z'dej
1.2	Izveden je naglasni umik na prednaglasna e in o, včasih tudi na ə:	vé:ža, ré:kəl so; visó:ke, vó:do, pró:stor, zgó:rej, gó:r; s-skẹ́:dnam
2	Dolgi samoglasniki; le enoglasniki s tipičnimi gorenjskimi razvojnimi sovpadi	
2.1	ě > (e:i̯), e in ę > ẹ:; pred j > ẹ:	mẹ́:le sə, strẹ́:xe, pẹ̀:č, z zapẹ́:čkəm, ləsẹ̀:n; nəprẹ̀:j

Slika 3: Pojavno okno z narečnim besedilom za izbrano lokacijo.

Na zemljevidu so dodane še oznake v krajih, kjer so bili posneti zvočni zapisi. Posamezni kraj oz. postavitev oznake je določena z geografskimi koordinatami (geografsko dolžino in širino) kraja. S klikom na to oznako se odpre pojavno okno, prikazano na sliki 3, v katerem lahko uporabnik predvaja zvočni zapis narečnega govora, prebere fonetično transkripcijo tega zapisa, njegovo poknjižitev na glasoslovni ravnini ter velikokrat tudi narečjeslovno analizo govora z vidika značilnosti na sedmih jezikovnih ravninah narečnega govora (naglas, dolgi samoglasniki, kratki naglašeni samoglasniki, kratki nenaglašeni samoglasniki, soglasniki, oblikoslovni pojavi in leksika). Zaradi boljše uporabniške izkušnje se pri predvajanju zvočnega posnetka nismo omejili le na funkciji *predvajaj* in *ustavi*, ampak smo zvok tudi vizualizirali z valovnimi oblikami in omogočili premikanje naprej in nazaj po posnetku.

V pojavnem oknu je dodana tudi možnost za tiskanje vsebine okna, to je transkripcije, poknjižitve in analize, skupaj s podatki o kraju in narečju (privzeto se vsi ti podatki shranijo v datoteko pdf na lokalnem računalniku). Le zvočnega posnetka narečnega govora ne moremo izvoziti iz aplikacije in lokalno shraniti.

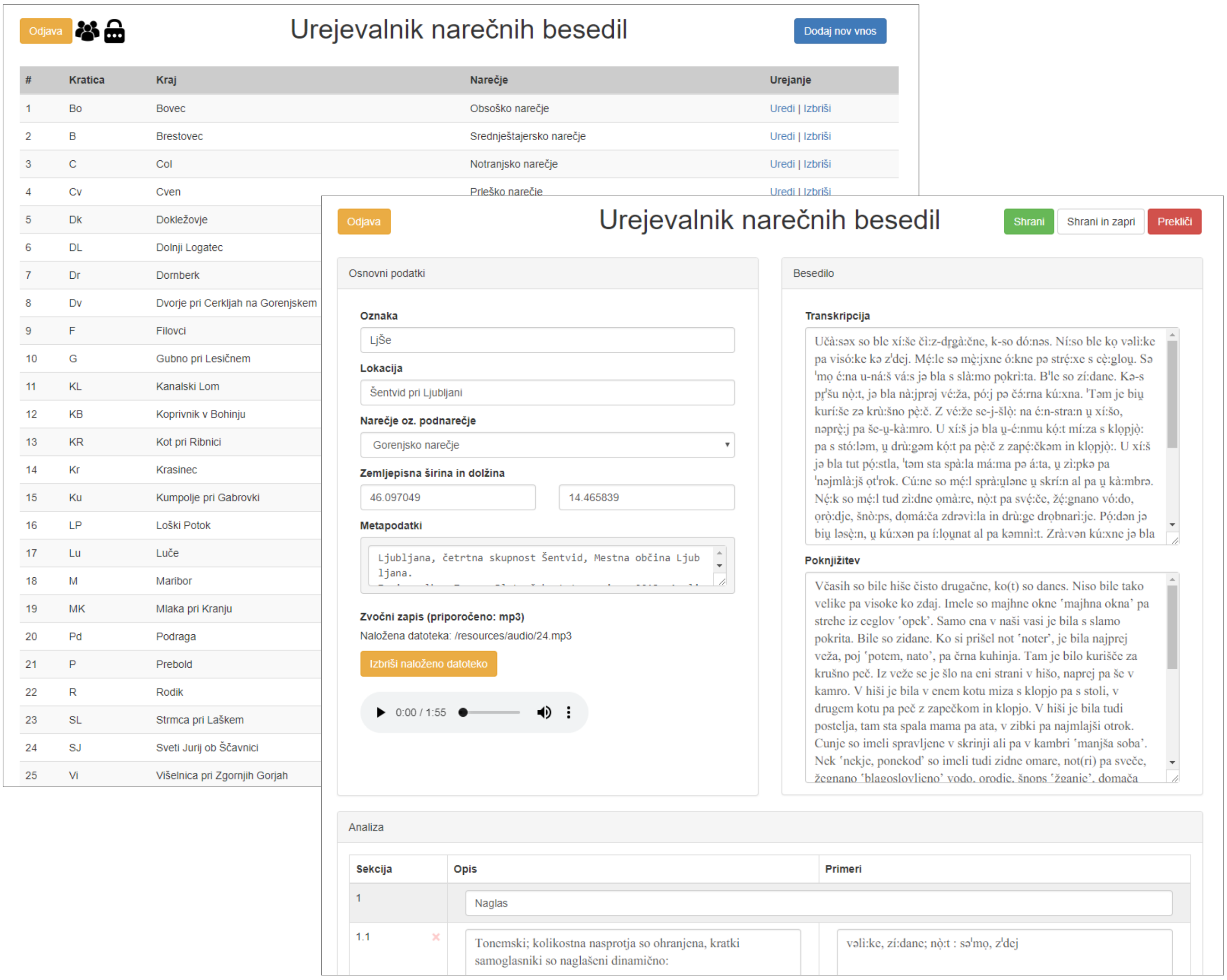
3.4. Skrbniška aplikacija

Skrbniška aplikacija zahteva najprej avtentikacijo uporabnika, saj je dostop do urejevalnika vsebine omejen. Ob uspešni prijavi se prikaže seznam vseh krajev, urejen abecedno (slika 4, levo okno). Vsak kraj lahko uredimo (spremenimo podatke) ali izbrišemo, lahko pa vnesemo tudi povsem nov kraj.

Pogled za urejanje vnosa prikazuje slika 4 (desno okno). Sestavljajo ga trije deli: Osnovni podatki, Besedilo in Analiza. Prvi del zajema podatke o lokaciji, kjer je bil posnet zvočni zapis narečnega govora: ime kraja, njegove zemljepisne koordinate (širino in dolžino), v katerih se na karti prikaže oznaka tega kraja, in oznako, to je kratico, ki se prikaže v oznaki kraja na karti. S spustnega seznama izberemo narečje oziroma podnarečje, kamor govor umeščamo, ter naložimo zvočni zapis govora v formatu mp3, ki se prenese in shrani na strežniku. Dodamo lahko tudi poljubne metapodatke, ki navadno vključujejo dodatne opise kraja, podatke o zapisovalcu in informatorju, leto zapisa ali katerekoli druge pomembne podatke.

V drugem delu sta dve polji za vnos besedila: transkripcija in poknjižitev. Transkribirano besedilo je vedno zapisano v pisavi ZRCola (Weiss, 2004) in ga lahko v aplikacijo kopiramo iz urejevalnika besedila, ki podpira vnašalni sistem ZRCola. Zaradi grafične doslednosti smo isto pisavo uporabili tudi v polju za poknjižitev, čeprav ta ne uporablja posebnih fonetičnih znakov. Poknjiženo besedilo lahko vnesemo neposredno v tekstovno polje v aplikaciji ali pa ga kopiramo iz urejevalnika besedila, v katerem smo ga predhodno pripravili.

Tretji del pa je namenjen vnašanju narečjeslovne analize narečnega govora. Analiza je sestavljena iz sedmih sekcij; vsaka opisuje značilnosti narečnih govorov na eni od jezikovnih ravnin in vsaka vsebuje poljubno število vnosov s primeri iz besedila, ki jih lahko dodajamo sproti.



#	Kratica	Kraj	Narečje	Urejanje
1	Bo	Bovec	Obsoško narečje	Uredi | Izbriši
2	B	Brestovec	Srednještajersko narečje	Uredi | Izbriši
3	C	Col	Notranjsko narečje	Uredi | Izbriši
4	Cv	Cven	Prleško narečje	Uredi | Izbriši
5	Dk	Dokležovje		
6	DL	Dolnji Logatec		
7	Dr	Dornberk		
8	Dv	Dvorje pri Cerkljah na Gorenjskem		
9	F	Filovci		
10	G	Gubno pri Lesičnem		
11	KL	Kanalski Lom		
12	KB	Koprivnik v Bohinju		
13	KR	Kot pri Ribnici		
14	Kr	Krasinec		
15	Ku	Kumpolje pri Gabrovki		
16	LP	Loški Potok		
17	Lu	Luče		
18	M	Maribor		
19	MK	Mlaka pri Kranju		
20	Pd	Podraga		
21	P	Prebold		
22	R	Rodik		
23	SL	Strmca pri Laškem		
24	SJ	Sveti Jurij ob Ščavnici		
25	Vi	Višelnica pri Zgornjih Gorjah		

Oznaka: LjŠe

Lokacija: Šentvid pri Ljubljani

Narečje oz. podnarečje: Gorenjsko narečje

Zemljepisna širina in dolžina: 46.097049 14.465839

Metapodatki:

```
Ljubljana, četrtna skupnost Šentvid, Mestna občina Ljub
ljana.
```

Zvočni zapis (priporočeno: mp3)

Naložena datoteka: /resources/audio/24.mp3

Transkripcija

Učà:səx so ble xí:šc čì:z-dr̥gà:čne, k-so dó:nəs. Ní:so ble kọ vəlì:ke pa visó:ke kə zˈdej. Mẹ́:le sə mẹ̀:jxne ó:kne pə strẹ́:xe s cẹ̀:glou̯. Səˈmọ é:na u-ná:š vá:s jə bla s slà:mo pọkrì:ta. Bˈle so zí:dane. Kə-s pr̥ˈšu nọ̀:t, jə bla nà:jprəj vé:ža, pó:j pə čə́:rna kú:xna. ˈTəm je biu̯ kurí:šc zə krù:šno pẹ̀:č. Z vé:že se-j-šlọ̀: na é:n-stra:n u̯ xí:šo, nəprẹ̀:j pa šc-u̯-kà:mro. U xí:š jə bla u̯-é:nmu kọ́:t mí:za s klọpjọ̀: pa s stó:ləm, u̯ drù:gəm kọ́:t pa pẹ̀:č z zapẹ́:čkəm in klọpjọ̀:. U xí:š jə bla tut pọ́:stla, ˈtəm sta spà:la má:ma pə á:ta, u̯ zì:pkə pa ˈnəjmlà:jš ọtˈrok. Cú:ne so mẹ́:l sprà:u̯ləne u̯ skrí:n al pa u̯ kà:mbrə. Nẹ́:k so mẹ́:l tud zì:dne ọmà:re, nọ̀:t pa svẹ́:če, žẹ́:gnano vó:do, ọrọ̀:dje, šnò:ps, dọmá:ča zdrəvì:la in drù:ge drọbnarì:je. Pọ́:dən jə biu̯ ləsẹ̀:n, u̯ kú:xən pa í:lọu̯nat al pa kəmnì:t. Zrà:vən kú:xne jə bla

Poknjižitev

Včasih so bile hiše čisto drugačne, ko(t) so danes. Niso bile tako velike pa visoke ko zdaj. Imele so majhne okne ʻmajhna oknaʼ pa strehe iz ceglov ʻopekʼ. Samo ena v naši vasi je bila s slamo pokrita. Bile so zidane. Ko si prišel not ʻnoterʼ, je bila najprej veža, poj ʻpotem, natoʼ, pa črna kuhinja. Tam je bilo kurišče za krušno peč. Iz veže se je šlo na eni strani v hišo, naprej pa še v kamro. V hiši je bila v enem kotu miza s klopjo pa s stoli, v drugem kotu pa peč z zapečkom in klopjo. V hiši je bila tudi postelja, tam sta spala mama pa ata, v zibki pa najmlajši otrok. Cunje so imeli spravljene v skrinji ali pa v kambri ʻmanjša sobaʼ. Nek ʻnekje, ponekodʼ so imeli tudi zidne omare, not(ri) pa sveče, žegnano ʻblagoslovljenoʼ vodo, orodje, šnops ʻžganjeʼ, domača

Analiza

Sekcija	Opis	Primeri
1	Naglas	
1.1 ✕	Tonemski; kolikostna nasprotja so ohranjena, kratki samoglasniki so naglašeni dinamično:	vəlì:ke, zí:dane; nọ̀:t : səˈmọ, zˈdej

Slika 4: Skrbniški pogled aplikacije – urejevalnik narečnih besedil in pogled urejanja vnosov.

Ker za vse kraje nimamo vedno pripravljenih vseh podatkov, lahko transkripcijo, poknjižitev ali analizo pustimo tudi neizpolnjeno.

4. Zaključek

Predstavljena aplikacija je prva tovrstna aplikacije za slovenska narečna besedila. Poskušali smo izdelati uporabniku prijazno spletno aplikacijo, ki bi bila dovolj enostavna tudi za širšo uporabo. Rezultat je interaktivna spletna karta slovenskih narečnih besedil, ki omogoča pregled vseh slovenskih narečnih skupin, narečij in podnarečij ter omogoča poslušanje zvočnih zapisov narečnih govorov, ogled njihovih fonetičnih zapisov, prevodov v knjižno slovenščino in analiz narečnih posebnosti. Poleg tega skrbnikom z Oddelka za slovenistiko omogoča dodajanje novih vnosov in tako zagotavlja vsebinsko vedno bogatejši spletni vir ne le za jezikoslovce in narečjeslovce, ampak tudi za učence in dijake, njihove profesorje ter vse, ki jih zanima odkrivanje bogastva in posebnosti slovenske narečne govorice.

Interaktivno karto narečnih besedil nameravamo še nadgraditi, predvsem za izboljšanje uporabniške izkušnje. Aplikacijo bomo prenesli na strežniško infrastrukturo CJVT[7] in tako poskrbeli tudi za njeno dolgoročno vzdrževanje in delovanje. S tem bomo omogočili tudi trajnostno upravljanje z zbranimi gradivi.

Načrtujemo tudi vsebinsko dopolnitev aplikacije z dodatnimi 30 kraji, za katere imamo že pripravljene zvočne posnetke, transkripcije besedil in pripadajoče poknjižitve, za nekatere govore pa so izdelane tudi diahrone analize.

Zbrane podatke o narečnih govorih, vključenih v aplikacijo, nameravamo ponuditi tudi v obliki spletne zbirke podatkov, saj predstavljena aplikacija ne omogoča enostavnega izvoza in ponovne uporabe teh podatkov. Tako bodo zbrani podatki enostavno dosegljivi tudi drugim raziskovalcem, predvsem pa bosta omogočena njihovo strojno branje in obdelava.

5. Literatura

Tine Logar in Jakob Rigler. 1983. *Karta slovenskih narečij*. Izdelal Geodetski zavod SRS, kartografski oddelek. Ljubljana, DDU Univerzum (stenski zemljevid).

Ivan Lovrić. 2018. *Interaktivna spletna aplikacija za slovenska narečna besedila*. Diplomsko delo, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani. http://eprints.fri.uni-lj.si/4117/.

Fran Ramovš. 1931. *Dialektološka karta slovenskega jezika*. Rektorat Univerze kralja Aleksandra I – J. Blasnika nasl. Ljubljana, Univerzitetna tiskarna.

Vera Smole in Mojca Horvat. 2016. *Stara kmečka hiša, Narečna besedila z analizo I*. Znanstvena založba Filozofske fakultete Univerze v Ljubljani.

Vera Smole. 2016. Drugi zvezek Slovenskega lingvističnega atlasa (SLA) in narečna besedila o stari kmečki hiši. V: Remnëva Marina Leont'evna, ur., *Slovenskij jazyk, literatura i kul'tura v slavjanskom i evropejskom kontekste: tezisy Meždunarodnoj naučnoj konferenciji*. MGU im. V. Lomonosova, Filologičeskij fakul'tet, 28-29 nojabra 2016 goda, str. 91–94. MAKS Press, Moskva.

Jože Toporišič. 1987. Slovensko narečjeslovje. V: Toporišič Jože, *Portreti, razgledi, presoje*, str. 217–256. Založba Obzorja, Maribor.

Peter Weiss. 2004. ZRCola. *Jezikovni zapiski*, 10(1):145–152.

[7] Center za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT), dostopno na https://www.cjvt.si/.

Učinkovit izračun frekvenčnih statistik za slovenske jezikovne korpuse

Aleksander Ključevšek*, **Simon Krek**†°, **Marko Robnik-Šikonja***

*Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana
† Univerza v Ljubljani, Filozofska fakulteta, Aškerčeva 3, 1000 Ljubljana
° Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana
aleksander.kljucevsek@gmail.com, simon.krek@guest.arnes.si, marko.robnik@fri.uni-lj.si

Povzetek

Veliki besedilni korpusi vsebujejo številne informacije o jeziku in njegovi rabi, nekatere lahko iz njih izluščimo s statistično analizo. Večina obstoječih orodij je razvitih in prilagojenih za obdelavo angleških besedil. V prispevku predstavljamo razvoj orodja za statistično analizo velikih slovenskih jezikovnih korpusov, ki upošteva značilnosti slovenščine kot morfološko bogatega jezika. Današnji besedilni korpusi lahko vsebujejo tudi več milijard besed, zato je bil velik del pozornosti namenjen razvoju učinkovitih paralelnih algoritmov, s katerimi bo moč tako obsežne zbirke v razmeroma kratkem času obdelati tudi na običajnih računalnikih. Orodje omogoča analizo na več nivojih: na nivoju besednih nizov, nivoju besed, n-gramov, predpon in končnic ter tudi oblikoskladenjskih oznak v slovenščini. Trenutno so podprti korpusi Gigafida, ccGigafida, Kres, ccKres, GOS in Šolar, vendar je dodajanje novih korpusov enostavno zaradi abstrakcije vhodnih podatkov.

Efficient calculation of frequency statistics for Slovene language corpora

Large text corpora hold a vast amount of information about language and its use, some of which can be extracted with statistical analysis. Most of existing tools are prepared for English texts. We present an application for statistical analysis of large Slovene text corpora, which takes into account rich morphology of Slovene. Since modern text corpora contain billions of words, we developed efficient parallel algorithms capable of processing these collections effectively using desktop computers. Our tool can analyze corpora on multiple levels: as strings, words, n-grams, through prefixes, suffixes, and POS tags. Currently the tools supports Gigafida, Kres, GOS, and Šolar, but adding support for new corpora is simple due to abstraction of input data.

1. Uvod

Čeprav obstajajo številni priročniki, ki formalizirajo pravila uporabe jezika, se jezik spreminja s časom in kontekstom, kar s časom privede do prilagoditve pravil. V zadnjem času se z vse večjimi možnostmi komunikacije razvijajo tudi nove specifične vrste komunikacije: jezik, uporabljen v kratkih sporočilih, se razlikuje od tistega, ki ga isti najstnik uporablja med pisanjem eseja, tako kot se jezik profesionalnega novinarskega članka razlikuje od jezika romanov. S statistično analizo besedil lahko dobimo vpogled v same temelje jezika: kako so sestavljeni stavki, kako pogosto se uporabljajo posamezne besede in besedne vrste, v kakšnih kombinacijah, v besedotvorne procese v različnih tipih besedil oz. v različnih tipih komunikacije itd. S tem spremljamo razvoj in spremembe v jeziku, odkrijemo pa lahko tudi nove jezikovne pojave. V članku opišemo učinkovito prosto dostopno orodje za statistično analizo besedil, ki podaja odgovore na ta in podobna vprašanja.

Velike množice besedil danes hranimo v zbirkah, imenovanih korpusi, večinoma zapisanih v formatu XML, ki omogoča enostavno dodajanje metapodatkov. Statistična analiza jezika je še posebej zanimiva na dovolj velikih korpusih, ki neredko dosežejo več milijard besed. Obdelava tolikšne količine podatkov lahko postane računsko zahtevna in dolgotrajna, še posebej če se ne uporabijo dovolj optimizirani algoritmi in primerna strojna oprema.

Da bi bilo naše orodje uporabno smo si zadali nekaj zahtev:

1. orodje mora biti zmožno učinkovito obdelati korpuse velikosti več milijard besed,
2. delovati mora tudi na enem samem, povprečnem računalniku,
3. sposobno mora biti izkoristiti razpoložljive pomnilniške in procesorske vire računalnika.

Prispevek je razdeljen na šest razdelkov. V 2. razdelku pripravimo kratek pregled obstoječih del. V 3. razdelku na kratko predstavimo zapis korpusov in podprte korpuse. V 4. razdelku predstavimo podprte funkcionalnosti, zgradbo našega orodja, težave, s katerimi smo se soočili, in analizo časovne ter prostorske zahtevnosti nekaterih uporabljenih algoritmov. V 5. razdelku predstavimo manjši vzorec zanimivih analiz. Prispevek zaključimo s 6. razdelkom, v katerem podamo glavne sklepe in navedemo možne izboljšave.

2. Sorodna dela

Eno prvih večjih statističnih analiz sta v 60-ih letih prejšnjega stoletja opravila Mayzner in Tresselt (1965), ki sta iz 100 različnih angleških virov (časopisi, revije, knjige itd.) ročno izločila po 200 zaporednih besed za skupno 20 000 besedni korpus. Te besede sta ročno prenesla na luknjane kartice in jih s pomočjo naprave za procesiranje kartic analizirala glede na dolžino in pozicijo črk znotraj besed. Norvig (2013) je ponovil analizo na angleških unigramih iz zbirke Google books Ngrams (Michel et al., 2011; Google, 2012). Uporabili so 97 565 različnih besed, ki so se v korpusu pojavile vsaj 100 000 −krat in so bile skupaj omenjenje 743 842 922 321 −krat, kar pomeni, da je bil korpus 37 milijonkrat obsežnejši kot originalni iz dela (Mayzner in Tresselt, 1965). Za izračun tako obsežnega korpusa bi računalnik, ki ga je uporabljal Mayzner, potreboval 700 let. Med drugim je analiza razkrila, da število omemb besed glede na dolžino besede sledi Poissonovi distribuciji, pri čemer je $55.95\,\%$ najpogosteje uporabljenih besed dolžine $2-4$ črke. Povprečna dolžina besede je $4,79$ črk na besedo. Besede dolžine več kot 15 črk so uporabljene $834\,000\,000-$krat.

Lauer (1995) je pokazal, da lahko tudi preposte korpusne statistike dodajo informacije o sintaksi samostalniških fraz. Predstavil je štiri algoritme, ki za sintaktično analizo samostalniških fraz uporabljajo korpuse in delovanje algoritmov preizkusil na zbirki 244 izrazov. Trije algoritmi temeljijo na modelu sosednjosti (angl. adjacency model), medtem ko zadnji temelji na modelu odvisnosti (angl. dependency model). V vseh primerih se je za statistično značilno boljšega izkazal slednji.

Statistična analiza korpusov se je za pomembno izkazala tudi pri analizi podobnosti stavkov v kombinaciji s semantičnimi mrežami, kjer dobro odraža uporabo besed in izrazov v vsakdanji rabi (Li et al., 2006). Schiffman et al. (2001) so korpusne statistike uporabili pri avtomatskem generiranju povzetkov biografij. Njihov cilj so bili povzetki za enostavne biografije ljudi iz podatkov, ki so bili o njih objavljeni v medijih. Razvito metodo so uporabili na t. i. "Clintonovem korpusu" - korpusu sodnih zapisnikov predsednika Clintona na obravnavi po razkritju njegove afere. V zapisniku je bilo omenjeno veliko število ljudi in njihovih aktivnosti. Yang in Wilbur (1996) sta s pomočjo korpusnih statistik izračunala pomembnost posameznih besed in s tem skrčila seznam besed uporabljenih pri opisih namenjenih tekstovni kategorizaciji za $87\,\%$, kar je pripomoglo k $63\,\%$ krajšemu času obdelave takšnih seznamov in $74\,\%$ zmanjšanju porabe pomnilnika. Sočasno se je točnost napovedi v povprečju povečala za $10\,\%$ v primerjavi z neskrčenimi seznami.

Za kompleksnejše procesiranje slovenskega jezika v jezikoslovne namene je možno uporabiti odprtokodno orodje NooJ (Silberztein, 2016). Orodje omogoča analizo tekstov v več kot 20 jezikih, tudi slovenščini, na pravopisnem, leksikalnem, oblikoslovnem, sintaktičnem in semantičnem nivoju. Dobrovoljc (2014) je predstavila uporabo orodja za slovenščino. Orodje omogoča oblikovanje korpusnih poizvedb po površinski in označeni strukturi besedila, denimo za luščenje podatkov iz površinsko skladenjsko razčlenjenih korpusov, govornih korpusov ali drugih korpusov, ki poleg slovničnih lastnosti besednih oblik vsebujejo tudi druge vrste in ravni jezikoslovnih oznak. NooJ odlikuje vmesnik za preprost opis raznolikih jezikovnih pojavov v obliki grafov.

3. Korpusi

Za enostavno uporabo so korpusi shranjeni v strukturirani obliki. Če format to dovoljuje, so posameznim besedam dodani metapodatki, ki omogočajo uporabo dodatnih lastnosti besed pri analizi, npr. leme in oblikoskladenjske lastnosti besed. Vsi korpusi, ki jih naše orodje trenutno podpira, so v formatu XML.

Standard jezika XML se je zelo uveljavil v jezikovnotehnološki skupnosti, saj omogoča dobro berljivost ter shranjevanje in predstavitev strukturiranih jezikovnih podatkov. Osnovni standard jezika XML je moč prilagoditi specifičnim potrebam z dodatkom strožjih omejitev strukture in označevanja podatkov. Za namen shranjevanja označenih jezikovnih korpusov je konzorcij TEI (Text Encoding Initiative) izdal priporočila za označevanje besedil, ki obsegajo zapise različnih zvrsti besedil in jezikoslovnih korpusov. Prva dva nivoja etiket dokumentov, pripravljenih po standardu TEI, sta i) informacija o uporabljeni različici standarda XML in kodiranju ter ii) korenski element TEI, ki vsebuje kolofon in besedilo.

V kolofonu so shranjeni metapodatki o besedilu, npr. bibliografski podatki, struktura dokumenta, uporabljena taksonomija, opis značilnosti vsebovanega besedila itd. Samo besedilo se primerno strukturirano nahaja v naslednjem elementu. TEI predpisuje mnogo možnih oznak, ki jih lahko uporabimo pri strukturiranju in opisovanju besedil. Pri gradnji korpusov so najpogosteje uporabljani elementi za odstavke <p>, stavke <s>, besede <w>, ločila <c>in presledke.

V korpusih, ki jih trenutno podpira orodje, vsaka beseda poleg zapisa besede iz obravnavanega besedila vsebuje še atributa *lemma* in *msd*. V atributu *lemma* je shranjena lema (geselska iztočnica) besede, medtem ko je v atributu *msd* shranjena oblikoskladenjska oznaka besede, ki sledi specifikacijam MULTEXT-East različica 4.0 (Erjavec, 2012). Oznake za slovenski jezik so bile razvite v okviru projekta JOS (Erjavec in Krek, 2008) in zajemajo več kot 1900 oznak, ki so lahko izražene v slovenščini ali angleščini. V nadaljevanju na kratko predstavimo podprte korpuse.

3.1. Gigafida in KRES

Gigafida je referenčni korpus, ki v trenutni različici 1.0 (Holdt et al., 2012) vsebuje skoraj 1,2 milijarde besed (natančneje 1 187 002 502 besed), zajetih iz besedil, ki so v tiskani obliki ali na internetu izšla v obdobju med leti 1990 do 2011 (Erjavec in Logar Berginc, 2012). Korpus je zapisan v formatu XML TEI P5 (format XML z dodatnimi specifikacijami namenjenimi strukturirani predstavitvi besedil), je lematiziran in oblikoskladenjsko označen. Gigafida vključuje referenčni korpus FidaPLUS iz leta 2006 (621 milijonov besed) (Arhar et al., 2007) in tudi prvi slovenski referenčni korpus FIDA (1997 - 2000) (Erjavec et al., 1998).

Korpus Gigafida s svojo velikostjo in raznolikostjo besedilnih zvrsti predstavlja celovito podobo slovenskega jezika, ni pa uravnotežen, saj je kar $77\,\%$ besed iz periodičnih virov in samo $6\,\%$ besed iz knjig. KRES je uravnotežen podkorpus Gigafide, ki vsebuje skoraj 100 milijonov besed (natančneje 99 831 145 besed) iz besedil izdanih med letoma 1990 in 2011, pri čemer je bil spletni del korpusa zbran in izdelan leta 2010. Ker je KRES podkorpus Gigafide, ji je po strukturi identičen.

3.2. GOS

Korpus GOvorjene Slovenščine GOS vsebuje več kot milijon besed zapisanih iz okrog 120 ur posnetkov (Verdonik et al., 2011). Vsa besedila so transkripcije posnetkov različnih vsakodnevnih situacij, ki so bili večinoma posneti med letoma 2008 in 2010. Nekatere od situacij so: radijske in televizijske oddaje, predavanja, sestanki, svetovanje, zasebni pogovori med družinskimi člani ali prijatelji itd. Posnet govor je v korpusu shranjen v dveh različicah: pogovorni in standardizirani, pri čemer je standardizirani zapis tudi lematiziran in oblikoskladenjsko označen. Za reprezentativnost korpusa je bilo poskrbljeno tako z vključitvijo številnih različnih diskurzov kot tudi z reprezentativnostjo govorcev iz različnih regij, spolov, starosti in izobrazbenih ravni.

3.3. Šolar

Korpus Šolar je namenjen raziskovanju pisne jezikovne zmožnosti šolajoče se populacije (Rozman et al., 2012). Zajema skoraj milijon besed (967 477 besed) zajetih iz 2703 pisnih izdelkov (eseji oz. spisi, obnove, prošnje, odgovori na vprašanja itd.) šolarjev zadnjega triletja osnovne šole in srednješolcev. Poleg oblikoskladenjskih oznak vsebuje tudi označene učiteljske popravke. Za razliko od prej omenjenih korpusov Šolar ni skladen s TEI. Ne vsebuje taksonomije, vsebuje pa dodatne metapodatke, po katerih lahko filtriramo besedila: regija, predmet, razred, leto, šola in tip besedila.

4. Zgradba orodja

V tem razdelku predstavimo najprej interno predstavitev korpusnih podatkov, ki omogočajo iskalnim algoritmom enoten dostop, in zgradbo programske rešitve.

4.1. Interna predstavitev podatkov

Format TEI P5 XML, v katerem so zapisani vsi omenjeni korpusi z izjemo Šolarja, je primeren za shrambo jezikovnih korpusov in pripadajočih metapodatkov. Za enostavnejše procesiranje je potrebno podatke iz takšnega formata preneseti v podatkovno strukturo, s katero je moč enostavno upravljati v programski kodi. Za razvoj programa smo uporabili programski jezik java, zato smo izkoristili njegovo objektno naravo. Ustvarili smo dva tipa objektov, enega za stavke in drugega za besede, pri čemer je objekt *stavek* sestavljen iz množice objektov *beseda* in atributa za podatek o taksonomiji in ostalih lastnostih stavka. Objekt *beseda* je sestavljen iz znakovnih nizov za besedo, lemo besede in kode MSD (morfosintaktična oznaka). Posamezne etikete XML zapisa so tako analogne objektom v Javi, atributi etiket pa atributom objekta.

4.2. Zasnova programa

Za čim enostavnejšo razširljivost z novimi korpusi je program razdeljen na več ločenih nivojev, ki med seboj komunicirajo preko smiselnih programskih klicev: grafični vmesnik, opis strukture in metapodatkov korpusov, branje podatkov in računanje statistik. Struktura in pripadajoče lastnosti korpusov so interno že vnaprej definirane, ker je s tem omogočena hitrejša obdelava, kot če bi te podatke generirali sproti ob izračunu vsake statistike. Rezultati se shranijo v tekstovni tabelarični obliki, ločeni z vejico v formatu CSV (angl. comma separated values). S tem je poenostavljeno dodajanje novih korpusov v prihodnosti, saj je v primeru novega korpusa potrebno zgolj definirati taksonomijo in ostale atribute ter dodati metodo, ki korpus prebere v prej opisano strukturo.

Program v prvem koraku učinkovito analizira metapodatke podanega korpusa, zazna za kateri korpus gre in na podlagi tega ponudi statistike, ki jih je moč izračunati. Vsi korpusi namreč ne vsebujejo vseh vrst podatkov, npr. pogovorni del korpusa GOS ni oblikoskladenjsko označen, kar močno omeji nabor možnih izračunov. Uporabnik nato izbere statistiko, ki ga zanima, in program jo izračuna.

4.3. Učinkovita raba pomnilnika

Največja težava procesiranja velikih jezikovnih korpusov je, da jih zaradi njihove velikosti ni mogoče hraniti v pomnilniku. Korpus Gigafida zasede 83.5 GB pomnilnika, povprečen računalnik pa ima med 4 GB in 8 GB pomnilnika. Z velikostjo korpusa je povezano tudi počasno branje podatkov z diska . V nasprotju z računanjem statistik, ki ga na večjedrnih in večnitnih procesorjih lahko paraleliziramo, branje s trdega diska ostaja ozko grlo, zaradi česar program veliko časa porabi za branje podatkov.

Za rešitev te težave smo uporabili paketno obdelavo vhodnih podatkov:

1. Program prebira vhodne podatke, dokler ne prebere določenega števila stavkov, nakar začasno prekine branje. Količina prebranega besedila je določena glede na velikost razpoložljivega pomnilnika v računalniku.

2. Na prebranih podatkih delno izračunamo zahtevano statistiko.

3. Prebrani podatki se zbrišejo, s tem se sprosti prostor za nov paket in program nadaljuje z branjem novih podatkov, s čimer se vrnemo na točko 1.

4. Ko podatkov zmanjka, program izpiše kumulativen rezultat za zahtevano statistiko.

Postopek je natančneje prikazan v algoritmu 1. Časovna zahtevnost takšnega pristopa je linearna - $\mathcal{O}(n)$, kjer n predstavlja velikost vhodnih podatkov.

Algoritem 1 Paketno procesiranje korpusa

```
while v korpusu so še neprebrani stavki do
    $subcorpus \leftarrow stavek$
    if $subcorpus.size \geq limit$ then
        FORK-JOIN($subcorpus$)
        $subcorpus = \emptyset$
    end if
end while
```

Tudi tako optimiziran pristop hitro naleti na omejitev: vsak objekt v javi ima določeno režijo (angl. overhead). Posamezen objekt tipa *beseda*, ki hrani besedo, lemo besede in nekaj črk MSD kode, zavzame 136 bajtov. Tako lahko v 1 GB pomnilnika v najboljšem primeru shranimo 7 352 941 besed, kar predstavlja 0.62 % besed v korpusu Gigafida. K temu je potrebno dodati še podatkovne strukture, v katerih hranimo računane statistike.

Zaradi podpore sočasnemu dostopu in konstantemu času vstavljanja in posodabljanja vrednosti smo kot osnovno podatkovno strukturo za hranjenje statistik izbrali razpršeno tabelo tipa HashMap. To je podatkovna struktura, v kateri je vsak vnos shranjen v obliki "ključ: vrednost". HashMap ne more vsebovati podvojenih ključev. Za dostop do ključa se uporablja razpršilna funkcija, kar omogoča, da se vnos in preverjanje, ali določen element že obstaja v podatkovni strukturi, izvedeta v konstantnem času $\mathcal{O}(1)$. Struktura HashMap podpira več sočasnih dostopov in je tako prilagojena paralelnemu procesiranju. Kljub temu pa je za velike korpuse, kot je Gigafida, pomnilniška poraba precejšnja in lahko na računalniku s 16 GB pomnilnika naenkrat računamo samo eno statistiko tipa frekvenca besed. Za nekatere statistike, ki zgenerirajo obsežne tabele rezultatov pa tudi to ni dovolj pomnilnika, zato program omogoča sprotno shranjevanje rezultatov na disk, kar sprosti pomnilnik na račun počasnejšega delovanja.

4.4. Paralelizacija

Vsi moderni procesorji imajo več jeder in podpirajo večnitno procesiranje. S poganjanjem algoritmov na več jedrih oz. nitih se ustrezno skrajša čas izvajanja. Če želimo izračunati frekvenco vseh besed to pomeni, da bomo za vsako besedo najprej preverili, ali smo nanjo že naleteli, nakar bomo število pojavitev te besede povečali za 1. Za korpus Gigafida to pomeni več kot milijardo operacij za vsako tovrstno statistiko, ki jo želimo izračunati.

Pri reševanju tega problema smo preverili več opcij: tokove (angl. streams) v Javi 8 ter paralelne arhitekture Fork-Join, Map Reduce in Akka. Večina obstoječih rešitev je namenjena obdelavi podatkov, ki se nahajajo na več ločenih sistemih (npr. porazdeljeno računanje na način Map Reduce) in s poganjanjem na samo enem sistemu ne morejo izkoristiti vgrajenih optimizacij (Stewart in Singer, 2012; Ranger et al., 2007). Odločili smo se za implementacijo Fork/Join paralelizacije (De Wael et al., 2014; Ponge, 2011; Lea, 2000), ker so se tokovi v Javi 8 izkazali za nezanesljive in so v določenih pogojih lahko tudi nekajkrat počasnejši kot Fork-Join (Langer, 2015; Zhitnitsky, 2015).

Fork-Join je v osnovi paralelna verzija principa deli in vladaj, kjer začetni problem rekurzivno delimo na manjše naloge, dokler ne dosežemo dovolj majhne velikosti podproblemov. Te rešimo neposredno, nakar rešitve združimo nazaj v rešitev začetnega problema. Pri vzporednih algoritmih se manjši deli ločeno in istočasno računajo s samostojnimi nitmi. Pogoj za takšno delitv je neodvisnost posameznih podproblemov. Osnovno delovanje je prikazano v algoritmu 3. Podatkovnih struktur pri takšni obdelavi ne spreminjamo ali prepisujemo v nove, manjše, temveč je podproblem definiran kot pogled na omejen del celotne podatkovne strukture. V Javi 8 je bil dodan princip ForkJoinPool, ki med drugim omogoča krajo opravil (angl. work stealing): če ena nit zaključi s svojim delom in mora čakati, da s svojim zaključi še druga, lahko v vmesnem času prevzame nedokončana opravila druge niti. Če so podproblemi dovolj majhni, je tako procesiranje bolj učinkovito, ker lahko učinkoviteje izkoriščamo celotno kapacitete procesorja.

Algoritem 3 Fork-Join algoritem.

```
procedure REŠI($problem$)
    if problem je dovolj majhen then
        reši problem direktno/sekvenčno
    else
        razdeli na podproblemA in podproblemB
        fork REŠI($podproblemA$)
        fork REŠI($podproblemB$)
        join rešitvi podproblemov A in B
    end if
end procedure
```

4.5. Časovna zahtevnost postopkov

Ob zaključku izračuna ene statistike na celotnem korpusu izračunane frekvence uredimo po padajočem zaporedju, za kar uporabimo javansko metodo TimSort, ki ima v najslabšem primeru časovno zahtevnost $O(n \log n)$ (Korniichuk, 2015). Skupno časovno zahtevnost celotnega programa dobimo s seštevanjem časovnih zahtevnosti posameznih delov:

- branje podatkov: $\mathcal{O}(n)$ +
- Fork-Join: $\mathcal{O}(k)$ +
- računanje statistike: $\mathcal{O}(n)$ +
- TimSort: $O(n \log n)$

Pri tem k pri koraku Fork-Join predstavlja število korakov delitve večjega problema na manjše (načeloma reda $\log_t n$), kjer je t število vzporedno delujočih niti. Končna časovna zahtevnost celotnega programa je tako enaka $O(n \log n)$.

4.6. Izračun statistik

Algoritme, ki statistiko izračunajo na opisan način smo razdelili v naslednje razrede:

LetterCount - vsebuje metode za računanje distribucij posameznih črk in njihovih zaporedij,

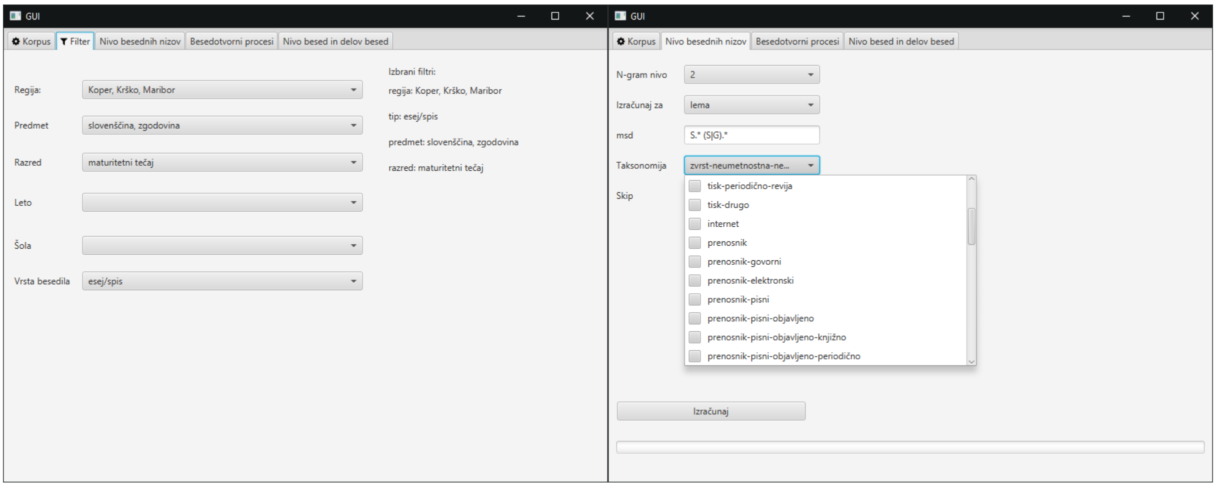
NGrams - vsebuje metode za računanje n-gramov besed in lem,

WordCount - vsebuje metode za računanje distribucij besed in lem,

WordLengthCount - vsebuje metode za računanje distribucij dolžin besed.

Metode v navedenih razredih kličemo z imenom metode in tremi parametri:

- podkorpusom oz. pogledom na del celotnega korpusa,
- referenco na HashMap, v katerega zapisujemo frekvence - ta ima v primeru računanja distribucije dolžin besed kot ključ objekt tipa Integer (celo število)
- z omejitvijo - to je lahko določena oznaka v taksonomiji, oblikoskladenjska oznaka iz tabele oznak JOS ali podatek, da določeno statistiko računamo za besedo, njeno lemo ali poljubno kombinacijo omenjenih kriterijev.



Regija: Koper, Krško, Maribor
Predmet: slovenščina, zgodovina
Razred: maturitetni tečaj
Leto:
Šola:
Vrsta besedila: esej/spis

Izbrani filtri:
regija: Koper, Krško, Maribor
tip: esej/spis
predmet: slovenščina, zgodovina
razred: maturitetni tečaj

N-gram nivo: 2
Izračunaj za: lema
msd: S.* (S|G).*
Taksonomija: zvrst-neumetnostna-ne...
Skip

- □ tisk-periodično-revija
- □ tisk-drugo
- □ internet
- □ prenosnik
- □ prenosnik-govorni
- □ prenosnik-elektronski
- □ prenosnik-pisni
- □ prenosnik-pisni-objavljeno
- □ prenosnik-pisni-objavljeno-knjižno
- □ prenosnik-pisni-objavljeno-periodično

Slika 1: Prikaz zavihka za izbiro filtrov pri korpusu Šolar (levo) in zavihka Nivo besednih nizov (desno).

4.7. Grafični vmesnik

Grafični vmesnik orodja je strukturiran v zavihke in ilustriran na sliki 1:

Korpus - uporabnik izbere lokacijo korpusa in lokacijo kamor se bodo shranjevali rezultati. V primeru korpusa GOS se pokaže dodatna opcija, s katero označimo, če želimo statistike računati za govorni zapis.

Filter - zavihek je viden samo, če smo izbrali korpus Šolar, ki vsebuje dodatne označbe kot npr. regija, predmet, vrsta besedila itd.

Nivo besednih nizov - na tem zavihku lahko računamo statistike na nivoju črk ali n-gramov. V obeh primerih lahko za računanje uporabimo različnice ali leme, v primeru n-gramov lahko dodatno računamo še za oblikoskladenjske oznake, v primeru n-gramov stopnje 2 ali več pa je možno tudi računanje skip-gramov. Omogočeno je filtriranje po oblikoskladenjskih oznakah, ki podpira uporabo regularnih izrazov, s čimer lahko npr. računamo samo število pojavitev bigramov, kjer je prva beseda pridevnik v dvojini in druga lastno ime ženskega spola "P...d.. Slz.*", bigrami, kjer je prva beseda samostalnik, druga pa samostalnik ali glagol "S.* (S—G).*", pojavitve samostalnikov v imenovalniku "S...i" ali ekvivalenten "S.{3}i" itd. Možno je tudi filtriranje glede na taksonomijo besedila v korpusu oziroma, v primeru korpusa Šolar, filtriranje glede na regijo, predmet, razred, leto, šolo ali vrsto besedila. Na nivoju črk je možno tudi računanje frekvenc kombinacij samoglasnikov in soglasnikov poljubne dolžine.

Besedotvorni procesi - za pregibne besedne vrste program izračuna distribucijo oblik glede na tabelo oznak JOS. Pogoj za to je, da korpus vsebuje oznake msd. Izračuna se frekvenca vseh kombinacij lasnosti, s čimer dobimo odgovore na vprašanja kot so npr. "Se v korpusu pojavi več pridevnikov ženskega ali moškega spola?", "Kakšna je distribucija samostalnikov po sklonih?" ali tudi bolj specifična vprašanja kot npr. "Se večkrat pojavijo zanikani ali nezanikani dovršni glagoli velelne oblike?". Računanje lahko omejimo glede na taksonomijo.

Nivo besed in delov besed - program izračuna kolikokrat se v korpusu pojavi specifična predpona ali pripona na podlagi metode najdaljšega ujemajočega se podniza. Seznam predpon in pripon je podan programu kot dodaten vhod, program nato izračuna frekvenco. Računanje lahko omejimo glede na taksonomijo.

5. Rezultati

Kot primere izračunov, ki jih izvede naše orodje, navajamo nekaj statistik za slovenščino, izračunanih na korpusih Gigafida in KRES. Ker namen našega članka ni analiza teh frekvenc, pač pa demonstracija orodja, ki je prosto dostopno na `http://github.org/cjvt-ul/corpusStatistics`, predstavljamo primere izračunov za razrede statistik iz razdelka 4.6. Za prve tri razrede statistik navajamo najpogostejših 10 primerov posamezne izračunane statistike (tabele 5., 5., 5. in 5.), rezultate za WordLengthCount pa v obliki histograma navajamo na sliki 5..

	KRES		Gigafida	
	črka	%	črka	%
1.	a	10.12	a	10.01
2.	e	9.99	e	9.74
3.	o	9.07	o	9.03
4.	i	8.78	i	8.73
5.	n	6.74	n	6.69
6.	r	5.17	r	5.26
7.	s	4.57	t	4.47
8.	t	4.48	s	4.45
9.	l	4.46	l	4.36
10.	j	4.17	v	4.11

Tabela 1: Razred statistik LetterCount - 10 najpogosteje uporabljenih črk.

	KRES		Gigafida	
	števka	%	števka	%
1.	0	0.28	0	0.39
2.	1	0.27	1	0.33
3.	2	0.19	2	0.24
4.	3	0.12	3	0.15
5.	9	0.11	5	0.15
6.	5	0.11	4	0.12
7.	4	0.10	9	0.12
8.	8	0.08	6	0.10
9.	6	0.08	8	0.09
10.	7	0.07	7	0.09

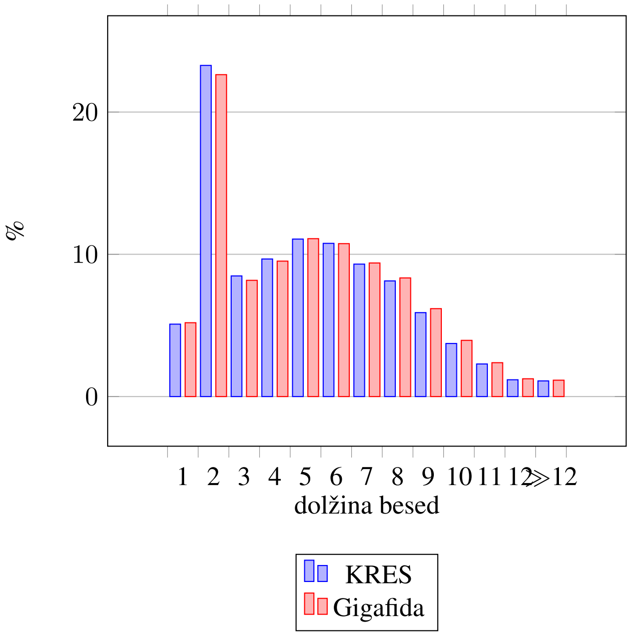
Tabela 2: Razred statistik LetterCount - pogostost uporabe števk.

	KRES		Gigafida	
	bigram	%	bigram	%
1.	Slmei- Slmei-	0.75	Slmei- Slmei-	1.09
2.	Dm Sozem-	0.74	Dm Sozem-	0.74
3.	Vd Gp-ste-n	0.64	Vd Gp-ste-n	0.64
4.	Dm Somem-	0.62	Ppnzer- Sozer-	0.64
5.	Ppnzei- Sozei-	0.61	Dm Somem-	0.63
6.	Ppnzer- Sozer-	0.59	Ppnzei- Sozei-	0.63
7.	Rsn Rsn	0.55	Kag—- Kag—-	0.58
8.	Dt Sozet-	0.52	Ppnmeid Somei-	0.54
9.	L Rsn	0.50	L Rsn	0.51
10.	Kag—- Kag—-	0.50	Dt Sozet-	0.50

Tabela 3: Razred statistik NGrams - najpogostejši bigrami oblikoskladenjskih oznak. Pomen oznak je definiran v standardu MULTEXT-East (Erjavec, 2012) in dosegljiv na `http://nl.ijs.si/ME/V4/msd/html/msd-sl.html`.

	KRES		Gigafida	
	lema	%	lema	%
1.	biti	7.60	biti	7.34
2.	in	2.84	v	2.63
3.	v	2.47	in	2.56
4.	se	1.87	se	1.59
5.	na	1.51	na	1.58
6.	z	1.39	z	1.33
7.	da	1.28	za	1.31
8.	on	1.24	da	1.23
9.	za	1.19	ki	1.02
10.	ta	1.06	ta	1.01

Tabela 4: Razred statistik WordCount - 10 najpogosteje uporabljenih lem.



Slika 2: Razred statistik WordLengthCount - kako pogosto se pojavijo besede določene dolžine. Povprečna dolžina besede v korpusu KRES je 5, 11 črk in 5, 18 črk v korpusu Gigafida.

6. Zaključki

Predstavili smo orodje za učinkovit izračun frekvenčnih statistik na velikih korpusih. Orodje z večnitnostjo učinkovito izkorišča večjedrne procesorje in frekvence izračunava vzporedno, pri tem pa izračune deli tako, da ne preseže razpoložljivega pomnilnika.

Kot nadaljno delo vidimo možnost podpore še drugim slovenskim in tujim korpusom in drugim vhodnim formatom. Večji izziv predstavlja avtomatska detekcija formata in vrste korpusa iz formata XML. Takšna razširitev bi načeloma bila zmožna obdelati poljuben korpus. Možne razširitve programa so dodatne statistike, ki bi vključevale regularne izraze na besedah in oblikokladenjskih oznakah, ter različnim tipom uporabnikov in jezikom prilagojen uporabniški vmesnik.

Zahvala

Raziskavo je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije skozi projekt J6-8256 (Nova slovnica sodobne standardne slovenščine: viri in metode) in raziskovalni program P2-0209 (Umetna inteligenca in inteligentni sistemi).

7. Literatura

Špela Arhar, Vojko Gorjanc in Simon Krek. 2007. Fidaplus corpus of slovenian: the new generation of the slovenian reference corpus: its design and tools. V: *Proceedings of the Corpus Linguistics conference*.

Mattias De Wael, Stefan Marr in Tom Van Cutsem. 2014. Fork/Join Parallelism in the Wild: Documenting Patterns and Anti-patterns in Java Programs Using the Fork/Join Framework. V: *PPPJ'14, International Conference on Principles and Practices of Programming on the Java*

Platform: Virtual Machines, Languages, and Tools, str. 39–50, Cracow.

Kaja Dobrovoljc. 2014. Procesiranje slovenskega jezika v razvojnem okolju NooJ. V: Tomaž Erjavec in Jerneja Žganec Gros, ur., *Zbornik 9. konference Jezikovne tehnologije, Informacijska družba - IS 2014*, str. 79–84.

Tomaž Erjavec, Vojko Gorjanc in Marko Stabej. 1998. Korpus fida. V: *Proc. of the Intl. Multi-Conf. Intl. Society'98*, str. 124–127.

Tomaž Erjavec in Simon Krek. 2008. The JOS morphosyntactically tagged corpus of Slovene. V: *6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, May 26 - June 1, 2008. LREC 2008: proceedings*, str. 322–326.

Tomaž Erjavec in Nataša Logar Berginc. 2012. Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES. *Proceeding of the Eighth Language Technologies Conference*, str. 57–63.

Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for central and eastern european languages. *Language resources and evaluation*, 46(1):131–142.

Google. 2012. Google books ngram corpus. http://books.google.com/ngrams.

Špela Arhar Holdt, Iztok Kosem in Nataša Logar Berginc. 2012. Izdelava korpusa gigafida in njegovega spletnega vmesnika. V: *Proceedings of 8th Eighth Language Technologies Conference IS-LTC*, zvezek 12.

Volodymyr Korniichuk. 2015. Timsort Sorting Algorithm. http://www.infopulse.com/blog/timsort-sorting-algorithm/.

Angelika Langer. 2015. Java performance tutorial – How fast are the Java 8 streams? https://jaxenter.com/java-performance-tutorial-how-fast-are-the-java-8-streams-118830.html.

Mark Lauer. 1995. Corpus statistics meet the noun compound: some empirical results. V: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, str. 47–54. Association for Computational Linguistics.

Doug Lea. 2000. A java fork/join framework. V: *Proceedings of the ACM 2000 conference on Java Grande*, str. 36–43. ACM.

Yuhua Li, David McLean, Zuhair A Bandar, James D O'shea in Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering*, 18(8):1138–1150.

Mark S Mayzner in Margaret Elizabeth Tresselt. 1965. Tables of single-letter and digram frequency counts for various word-length and letter-position combinations. *Psychonomic monograph supplements*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig in Jon et al. Orwant. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Peter Norvig. 2013. English letter frequency counts: Mayzner revisited or etaoin srhldcu. http://www.norvig.com/mayzner.html.

Julien Ponge. 2011. Fork and join: Java can excel at painless parallel programming too! http://www.oracle.com/technetwork/articles/java/fork-join-422606.html.

Colby Ranger, Ramanan Raghuraman, Arun Penmetsa, Gary Bradski in Christos Kozyrakis. 2007. Evaluating mapreduce for multi-core and multiprocessor systems. V: *High Performance Computer Architecture, 2007. HPCA 2007. IEEE 13th International Symposium on*, str. 13–24. Ieee.

Tadeja Rozman, Irena Krapš Vodopivec, Mojca Stritar in Iztok Kosem. 2012. *Empirični pogled na pouk slovenskega jezika*. Trojina, zavod za uporabno slovenistiko.

Barry Schiffman, Inderjeet Mani in Kristian J Concepcion. 2001. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. V: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, str. 458–465. Association for Computational Linguistics.

Max Silberztein. 2016. *Formalizing Natural Languages: The NooJ Approach*. John Wiley & Sons.

Robert Stewart in Jeremy Singer. 2012. Comparing fork/join and mapreduce. *Cite-seer, Tech. Rep.*, str. 1–20.

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

Yiming Yang in John Wilbur. 1996. Using corpus statistics to remove redundant words in text categorization. *JASIS*, 47(5):357–369.

Alex Zhitnitsky. 2015. How Java 8 Lambdas and Streams Can Make Your Code 5 Times Slower. http://blog.takipi.com/benchmark-how-java-8-lambdas-and-streams-can-make-your-code-5-times-slower/.

Kolokacijski slovar sodobne slovenščine

**Iztok Kosem,*† Simon Krek,† Polona Gantar,* Špela Arhar Holdt,*‡
Jaka Čibej,*†‡ Cyprian Laskowski***

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
iztok.kosem@ff.uni-lj.si
polona.gantar@guest.arnes.si
spela.arharholdt@ff.uni-lj.si
jaka.cibej@ff.uni-lj.si
cyprianadam.laskowski@ff.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
simon.krek@guest.arnes.si
‡ Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana

Povzetek

V prispevku predstavljamo Kolokacijski slovar sodobne slovenščine, nov leksikalni vir za slovenščino. Vir temelji na uporabi sodobnih leksikografskih metod, ki vključujejo avtomatsko luščenje leksikalnih podatkov iz korpusov, množičenje in hitro odzivnost na spremembe v jeziku. Med pomembnejšimi lastnostmi korpusa je prikazovanje gesel v različnih fazah izdelave, kar je novost v slovenskem in tudi mednarodnem prostoru in nadgrajuje idejo rastočega slovarja, pri čemer je eden glavnih razlogov za vpeljavo tega pristopa upoštevanje potreb uporabnikov. Poleg predstavitve vira in metodologije njegove izdelave se prispevek osredotoča na vmesnik, ki uvaja številne novosti prikaza kolokacijskih podatkov, pa tudi slovarskih podatkov nasploh. Prispevek zaključuje predstavitev načrtov za nadaljnje delo.

Collocations Dictionary of Modern Slovene

The paper presents a new lexical resource for Slovene, namely the Collocations Dictionary of Modern Slovene. The resource is being compiled using state-of-the-art lexicographic methods such as automatic extraction of lexical data from corpora, crowdsourcing, and quick responsiveness to language change. An important aspect of the dictionary compilation is that all entries (whether automatically generated, post-processed, finalized by lexicographers, etc.) are immediately published, while dictionary users are provided with the information on their status, i.e. the stage in the compilation process. After the presentation of the Collocations Dictionary and the methodology of its compilation, the paper focuses on the interface, which introduces several innovations in the presentation of collocational information. The paper concludes with an overview of future plans.

1. Uvod

Na mednarodni ravni se v zadnjih letih kaže porast zanimanja za izdelavo kolokacijskih virov. Slovarji kolokacij so nastali oz. nastajajo npr. za estonski (Estonian Collocations Dictionary; Kallas et al., 2015), nemški (German Collocations Dictionary; Roth, 2013; Häcki Buhofer et al., 2014) in španski jezik (DiCE; Vincze et al., 2011; Vincze in Alonso Ramos, 2013). Kolokacijski slovarji v tujini so bili vsaj do zdaj skoraj vedno izdelani za tuje govorce določenega jezika, praksa pa je pokazala, da so kolokacijski podatki zelo koristni tudi za materne govorce, kar navsezadnje potrjuje tudi vse večja tendenca splošnih enojezičnih slovarjev po vključevanju kolokacijskih podatkov. Vseeno pa omenjeni kolokacijski slovarji in podobni viri še vedno ne izkoriščajo vseh prednosti, ki jih ponujajo digitalni mediji; ravno nasprotno, nekateri avtorji, npr. nemškega kolokacijskega slovarja, so pri zasnovi v (pre)veliki meri upoštevali omejitve tiskane različice.

V prispevku predstavljamo Kolokacijski slovar sodobne slovenščine (KSSS), pri čemer največ pozornosti posvečamo metodologiji in postopkom priprave podatkov ter vmesniku, preko katerega bo slovar na voljo uporabnikom. Slovar je rezultat avtomatskih postopkov luščenja kolokacijskih podatkov iz korpusov, ki so bili za slovenščino v zadnjih letih razviti in nenehno izboljševani (npr. Kosem et al., 2013a, 2013b; Gantar et al., 2016). Glavni namen je nasloviti potrebo slovenskih govorcev po jezikovnih virih, usmerjenih v izboljševanje jezikovne produkcije, hkrati pa jezikovnotehnološki skupnosti in ostalim zainteresiranim deležnikom ponuditi obsežne računalniško procesljive podatke o sodobni slovenščini. Poleg tega smo želeli storiti pomemben korak naprej na področju prikazovanja kolokacijskih podatkov in izkoristiti čim več prednosti digitalnih medijev. Tako je bil eden od izzivov izdelati vmesnik, ki bi zadovoljil potrebe različnih uporabnikov, tako maternih kot tujih govorcev slovenščine.

2. Vzorčna baza kolokacij sodobne slovenščine

Vzorčna baza kolokacij je nastala na podlagi poskusnega projekta Baze kolokacijskega slovarja slovenskega jezika (Krek et al. 2016) in vsebuje avtomatsko izluščene kolokacijske podatke (kolokacije z zgledi) za 2.500 gesel, razdeljene po skladenjskih relacijah. Poskusna gesla so na voljo prek posebnega vmesnika (http://bkssj.cjvt.si), o katerem je bila leta 2016 opravljena evalvacijska študija med uporabniki, ki je ponudila tudi informacije o tem, kakšen odnos imajo uporabniki do avtomatsko izluščenih kolokacijskih podatkov. Rezultati študije so pokazali, da nekatere uporabnike neizčiščeni rezultati oz. neustrezni kolokacijski kandidati ter nestrukturiranost in (pre)velika količina podatkov mestoma motijo, vendar pa se jim scela takšen vir zdi koristen in uporaben. Dejansko je nekatere vprašane celo bolj kot prevelika količina podatkov zmotila premajhna količina gesel.

3. Kolokacijski slovar sodobne slovenščine

Na podlagi strokovnih analiz, pa tudi uporabniških komentarjev vmesnika BKSSJ, smo se lotili avtomatskega izvoza podatkov za veliko večji nabor iztočnic in izdelavo kolokacijskega slovarja (Kosem et al. 2018), in sicer prek Sketch Engine API (Gantar et al., 2015; 2016). Podatki so bili izluščeni iz referenčnega korpusa Gigafida (Logar et al. 2014). Prvotno smo ocenjevali, da bo izvoz zajel približno 50.000 iztočnic, vendar pa se je številka po čiščenju šuma v frekvenčnem seznamu, izločitvi lastnoimenskih iztočnic in iztočnic s prenizko frekvenco in posledično pomanjkanjem koristnih kolokacijskih podatkov skrčila na 35.989 iztočnic, ki vsebujejo skoraj 8 milijon kolokacij in malo manj kot 37 milijonov pripadajočih korpusnih zgledov. Pri izvozu smo uporabili enake nastavitve kot za 2.500 iztočnic BKSSJ (gl. Krek et al., 2016), nekoliko smo izboljšali le konfiguracijo GDEX[1] za slovenščino (Kosem, 2015), npr. kaznovali smo stavke, ki se končajo s podpičjem ipd. Izvožene podatke smo v postopku postprocesiranja dodatno prečistili (deduplikacija zgledov, odstranjevanje kolokacij z vsemi enakimi zgledi ipd.) in prilagodili (pripis iztočnice v ustrezni obliki, zapis kolokatorja v ustrezni obliki glede na podatke v oblikoslovnem leksikonu Sloleks ipd.).

Dandanes v praksi najdemo dva prevladujoča načina objave slovarjev. Prvi, ki ostaja zvest tradicionalnim metodam, je objava slovarja, ko so vsa gesla dokončana. Drugi način, ki je postal standard za spletne slovarje, pa je objava novih slovarskih gesel v rednih intervalih (ponavadi enkrat letno, mogoče celo pogosteje) – temu načinu Klosa (2013) pravi rastoči slovar. Za naše namene noben od omenjenih načinov objave ni bil ustrezen. Tudi pri rastočem slovarju lahko traja več let, preden količina gesel doseže dejansko uporabno vrednost za uporabnike.[2] Posledično smo se odločili za pristop, ki smo ga v slovenskem prostoru prvi predlagali v Krek et al. (2013) in pri katerem je čimveč jezikovnih podatkov odprto ponujenih uporabnikom takoj, ko je z jezikoslovnega vidika ocenjeno, da uporabna vrednost za jezikovno skupnost ustrezno odtehta podatkovni šum; tako pripravljeni podatki morajo vsebovati jasno informacijo o stopnji jezikoslovne pregledanosti.

V KSSS smo se odločili za uporabo naslednjih petih stopenj oz. faz slovarskih gesel:

(1) Avtomatsko izluščeni podatki, ki so postprocesirani (deduplikacija zgledov ipd.), dodano je tudi avtomatsko gručenje kolokatorjev glede na semantične lastnosti, tj. glede na semantični tip, npr. ustanove, predmeti ipd.

(2) Podatki po implementaciji leksikalnogramatičnih oz. statističnih »filtrov«. Na primer, pri vseh strukturah smo odstranili kolokator *biti*, saj se je glagol v veliki večini analiziranih primerov pojavljal v neustreznih kolokacijah oz. je bil pomensko izpraznjen. V bodoče načrtujemo izdelavo obsežnejših seznamov kandidatov za izločanje (angl. stoplist). Kot drugo smo iz avtomatsko izluščenih podatkov izločili vse predložne strukture, ki jih ne potrjuje Slovenski pravopis (Toporišič ur., 2001), saj se je v večini primerov izkazalo, da gre za napačno prepoznane strukture zaradi napak v označevanju.[3]

(3) Podatki z zgolj potrjenimi kolokacijami, ki pa še niso razporejene po pomenih. Poudariti velja, da na tej stopnji ne izločamo samo nekolokacij, temveč tudi statistično šibkejše oz. semantično manj relevantne kolokacije.[4] Upoštevati je namreč treba razliko med statističnimi kolokacijami, tj. statistično relevantnimi sopojavitvami dveh (ali več) besed, in semantičnimi kolokacijami, ki opravljajo določeno semantično funkcijo in so posledično relevantne za kolokacijski slovar. Kot primer lahko navedemo kolokacije tipa *bolnišnica + v +* samostalnik v mestniku (npr. *bolnišnica v Ljubljani, bolnišnica v Izoli*), ki so sicer prepoznane kot statistično relevantne, a za kolokacijski slovar niso zanimive. Vseeno so naši kriteriji za vključitev gradiva, ki se bodo sproti dopolnjevali tudi na podlagi ugotovitev temeljnega raziskovalnega projekta Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki (KOLOS; J6-8255), manj strogi kot kriteriji nekaterih znanih kolokacijskih slovarjev – npr. avtorji kolokacijskega slovarja založbe Macmillan so izločali že iztočnice, kot so *hiša*, *kupiti* in *dober*, ker naj ne bi imele statistično zelo relevantnih kolokatorjev. Vendar pa že pri *hiša* najdemo kolokacije, kot so *stanovanjska hiša, medijska hiša, gradnja hiše*, katerih jakost je zelo visoka.

(4) Pomensko členjena gesla: kolokacije in pripadajoči zgledi so razporejene po pomenih (več o tem v nadaljevanju).

(5) Dokončno pregledano in z morebitnimi manjkajočimi podatki (npr. oznake) opremljeno geslo.

Precej razmisleka je bilo vloženega v snovanje pomenskih opisov (Kosem et al., 2017). Kot prvo smo se odločili, da bomo uporabili samo pomene, ne pa tudi podpomenov, saj je ta rešitev zaradi zelo majhnih razlik med posameznimi podpomeni bolj smiselna in posledično prijazna uporabnikom. Poleg tega podpomene pogosto še bolj učinkovito kot razlage ponazarjajo kolokacijski nizi.

Pri pomenskih opisih smo se odločili za uporabo kratkih indikatorjev namesto daljših razlag, saj predvidevamo, da uporabniki pomene besed, ki jih iščejo, bodisi poznajo ali pa ne potrebujejo natančnih razlag, temveč le osnovne pomenske namige za prepoznavo ustreznega pomena. Indikatorjem podobne mehanizme že dolgo uporabljajo slovarji za tuje govorce, predvsem angleščine (npr. Longman Dictionary of Contemporary English), zadnje

[1] GDEX (Good Dictionary EXamples; Kilgarriff et al., 2008) je del korpusnega orodja Sketch Engine, s katerim rangiramo kandidate za dobre slovarske zglede.

[2] Dober primer omenjene problematike je eSSKJ (https://fran.si/201/esskj-slovar-slovenskega-knjiznega-jezika, pri katerem je bilo v prvih dveh letih priprave objavljenih 611 gesel (v prvem letu celo manj kot 100 gesel).

[3] Med izločenimi strukturami so sicer mogoče tudi takšne, ki bi bile lahko legitimne, vendar pa moramo pred njihovo ponovno vključitvijo opraviti podrobnejšo analizo podatkov.

[4] Potrjene statistične kolokacije, ki jih ne vključimo v KSSS, sicer ostajajo v interni bazi, saj so relevantne za leksikografske (npr. izdelavo splošnih enojezičnih slovarjev) in jezikovnotehnološke namene.

čase pa tudi splošni enojezični slovarji (npr. Veliki slovar poljskega jezika).[5]

V KSSS uporabljamo več različnih tipov indikatorjev, npr. sinonime iztočnic, nadpomenke kolokatorjev določenega pomena, najbolj tipični kolokator pomena, področje rabe ipd. Težimo k jedrnatosti, tj. indikatorji so praviloma eno- ali dvobesedni.

Ker večino pomenske informacije prinašajo kolokacije same in njihovi zgledi, je ključna vloga indikatorjev predvsem v vzpostavljanju jasnih razlik med pomeni. Razločevalnost ima pri oblikovanju indikatorjev prednost med sistematičnostjo, kar pomeni, da lahko pri posamezni iztočnici za različne pomene izberemo različne tipe indikatorjev, npr.:

prevajati *(glagol)*

1. jezike
2. energijo
3. dražljaje
4. v računalništvu
5. v drugačno obliko

briljanten *(pridevnik)*

1. o občudovanju
2. iz briljantov
3. bleščeč

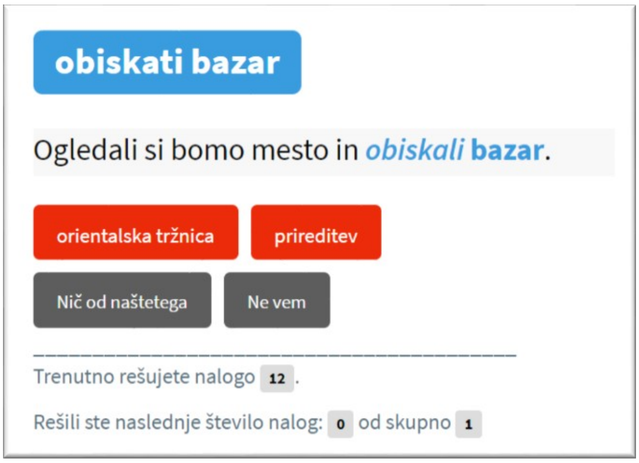
Zaradi že prej omenjene stopenjskosti gesel bo imel KSSS lastnosti odzivnega slovarja (Krek et al., 2017; Arhar Holdt et al., 2018), saj se bo odzival na spremembe v jeziku tako, da bodo na podlagi analiz novih podatkov, npr. nove verzije referenčnega korpusa slovenskega jezika, posodabljana tudi že objavljena gesla.

Izziv, ki ga prinaša odzivnost, pa je kratek čas za pripravo podatkov, saj sodobni uporabniki pričakujejo, da so jim slovarske informacije na voljo zelo hitro (Müller-Spitzer, 2014). To potrebo v prvi meri pokrivamo z vključitvijo podatkov v različnih fazah obdelave, pri čemer je končni cilj seveda ponuditi leksikografsko pregledane in redno ažurirane slovarske informacije. Ker je količina kolokacijskih podatkov za pregledovanje zelo velika, poleg tega pa smo tudi kadrovsko in finančno omejeni, smo za pomoč pri čim hitrejši izdelavi gesel KSSS v postopek vpeljali tudi metode množičenja, ki smo jih zasnovali in preizkusili že pri pripravi Predloga za izdelavo Slovarja sodobnega slovenskega jezika (Krek et al., 2013). Odločitev za vpeljavo množičenja se zdi še toliko bolj samoumevna, saj digitalni svet zdaj omogoča tovrstno podporo leksikografskih delotokov.

3.1. Množičenje podatkov za KSSS

Glavni namen vključitve neleksikografov v proces izdelave slovarja je razbremeniti leksikografe rutinskih nalog in njihovo znanje in energijo usmeriti v zahtevnejše leksikografske naloge, kot sta npr. pomenska členitev in pri pripravi indikatorjev. Ena od nalog, ki se nam je zdela primerna za množičenje, je uvrščanje zgledov pod pomene. Zgled, ki v našem primeru ponazarja konkretno kolokacijo, mora množičnik uvrstiti v enega od ponujenih pomenov.

Takšno nalogo smo izvedli na 3.295 kolokacijah iz 88 gesel KSSS, pri čemer smo ponudili po dva zgleda na kolokacijo (skupaj 6.590 mikronalog).[6] Nalogo smo pripravili v lokalni inštalaciji platforme Pybossa (Slika 1).[7] Za nalogo smo uporabili štiri označevalce, študente jezikoslovja, za vsako mikronalogo smo želeli dobiti tri odgovore. Poleg zgleda so označevalci imeli na voljo informacijo o kolokaciji, ki jo je zgled ponazarjal, ter pomene iztočnice, katere kolokacijo so označevali. Poleg pomenov so označevalci lahko izbrali tudi odgovor "Nič od naštetega", s čimer naj bi opozorili, da gre za pomen, ki ga ni na seznamu ponujenih, in "Ne vem", če niso vedeli, katerega od ponujenih pomenov izbrati.



Ogledali si bomo mesto in *obiskali* **bazar**.

Trenutno rešujete nalogo 12.

Rešili ste naslednje število nalog: 0 od skupno 1

Slika 1: Naloga uvrščanja zgledov pod pomene v Pybossi.

Ujemanje označevalcev je bilo precej visoko, strinjali so se v 79-86 % kolokacij (v povprečju v 83 % kolokacij, povprečna Cohenova kapa je bila 0,83). V 65 % kolokacij oz. 4.258 kolokacijah so se v odgovoru strinjali vsi trije označevalci. V 1.387 primerih (21 %) sta se strinjala po dva označevalca, le 147 primerov (2 %) pa je bilo povsem brez ujemanja. Večina primerov brez ujemanja (106 primerov) je bila označena z "Ne vem" ali z "Nič od naštetega" (54 primerov).[8] Na podlagi teh preizkusnih rezultatov lahko zaključimo, da je raba množičenja vsaj za takšno vrsto slovarsko vezane naloge, precej koristna.

Rezultati so pokazali še dodatno korist naloge, in sicer pridobivanje povratnih informacij o ustreznosti ubeseditve indikatorjev in pomenske členitve, pa tudi o morebitnih manjkajočih pomenih. Tako smo recimo pri glagolu *prihraniti* prvotno imeli ločena pomena za "manj porabiti" (npr. *prihraniti pri stroških*, *prihraniti denar*) in "varčevati" (npr. na banki), analiza odgovorov množičenja pa je pokazala, da bi bilo treba bodisi spremeniti ubeseditev enega ali celo obeh indikatorjev ali pa pomena združiti v en sam pomen. S takšno množičenjsko nalogo tako že dobivamo uporabniške povratne informacije, ki jih ponavadi raziskovalci oz. založniki pridobivajo, če sploh, šele v študijah po objavi slovarja oz. slovarskih gesel.

[5] http://wsjp.pl

[6] Število vseh kolokacij v 88 geslih je sicer še večje, a smo izločili (potrjene) kolokacije, ki smo jih našli v Leksikalni bazi za slovenščino (Gantar in Krek, 2011; Gantar et al., 2012).

[7] https://pybossa.com

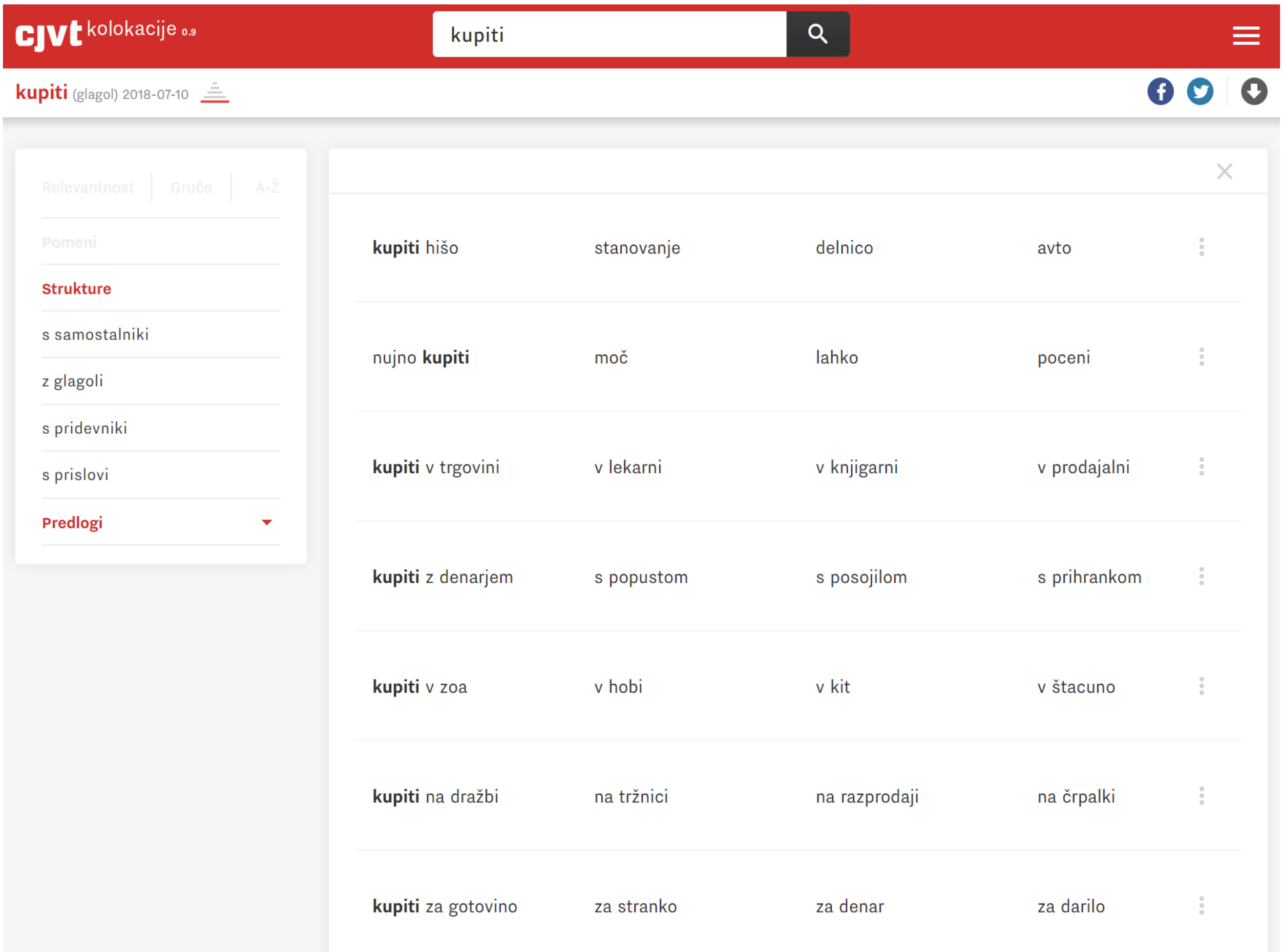
[8] 26 primerov je bilo označenih tako z "Ne vem" kot z "Nič od naštetega".

4. Vmesnik KSSS

Veliko pozornosti smo posvetili tudi zasnovi vmesnika KSSS.[9] Pri pripravi funkcionalnosti vmesnika smo izhajali predvsem iz informacij, pridobljenih pri uporabniški evalvaciji vmesnika poskusnih 2.500 gesel (Krek et al., 2016). Prva ključna odločitev je bila, da se vmesnik v jedrnem delu posveča kolokacijam, medtem ko so ostali tipi informacij, npr. pomeni, skladenjske relacije ipd., podani kot filtri. Na ta način se odmikamo od pomensko temelječega podajanja kolokacij, ki ga uporabljajo predvsem tiskani (kolokacijski slovar založbe Oxford University Press), pa tudi digitalno zasnovani kolokacijski slovarji (npr. kolokacijski slovar estonskega jezika).

Daleč največji izziv je bil, kako uporabnikom na jasen in nevsiljiv način posredovati informacijo o različnih stopnjah izdelanosti gesel. Čeprav je ta implicitno razvidna iz razpoložljivih funkcionalnosti vmesnika, npr. odsotnost filtra Pomeni nakazuje, da pomenska analiza na podatkih še ni bila opravljena, smo hoteli posamezne stopnje izdelave gesla nakazati tudi eksplicitno. Po daljšem razmisleku in diskusijah smo se odločili za uporabo ikone v obliki petstopenjske piramide, saj najbolje ponazarja postopek izdelave gesla: na začetku je podatkovno bogata, a vseeno že precej zanesljiva avtomatsko izluščena osnova, ki jo z vsakim korakom (proti vrhu) čistimo oz. pilimo. Poleg tega je sestavni del vsakega gesla tudi informacija o datumu zadnje posodobitve.

Pod vrstico, ki vsebuje informacije o iztočnici, tj. besedno vrsto, datum zadnje posodobitve in fazo izdelave gesla, se vmesnik deli na dva dela: na desni je osrednje okno s kolokacijami oz. kolokatorji, na levi pa (ožji) stolpec s filtri in funkcijo razvrščanja.[10] Ob odprtju gesla se uporabniku prikaže neke vrste kolokacijski profil iztočnice, saj mu ponudimo po nekaj kolokacij na skladenjsko strukturo (Slika 2). Praviloma je vsaki strukturi namenjena ena vrstica v vmesniku, če pa določena struktura močno prevladuje oz. vsebuje izredno velik delež relevantnih kolokacij iztočnice, lahko obsega več kot eno vrstico in posledično več kolokacij. V tem uvodnem prikazu uporabnik lahko izbere posamezno strukturo in si ogleda vse kolokacije v njej ali pa že izbere konkretno kolokacijo in si ogleda korpusne zglede.



kupiti

kupiti (glagol) 2018-07-10

kupiti hišo	stanovanje	delnico	avto
nujno **kupiti**	moč	lahko	poceni
kupiti v trgovini	v lekarni	v knjigarni	v prodajalni
kupiti z denarjem	s popustom	s posojilom	s prihrankom
kupiti v zoa	v hobi	v kit	v štacuno
kupiti na dražbi	na tržnici	na razprodaji	na črpalki
kupiti za gotovino	za stranko	za denar	za darilo

Slika 2: Uvodna stran gesla (primer glagola *kupiti*).

Namen levega stolpca je uporabniku omogočiti, da čim hitreje pride do želenih podatkov. Na vrhu stolpca je možnost razvrščanja kolokacij, ki je na voljo šele, ko uporabnik odpre posamezno strukturo. Privzeto so kolokacije razvrščene po relevantnosti oz. statistični jakosti, ostali možnosti sta Gruče (razvrščanje kolokatorjev v skupine glede na semantično podobnost) in A-Ž (po abecednem vrstnem redu).

[9] http://viri.cjvt.si/kolokacije/slv/. Povezava bo aktivna od sredine oktobra 2018, ko bo slovar uradno objavljen.

[10] V različici za mobilne telefone so filtri in razvrščanje na voljo prek menija na priklic na vrhu zaslona.

Možnostim razvrščanja sledijo trije tipi filtrov. Prvi filter, Pomeni, prikazuje pomene iztočnice. Pomembna lastnost tega filtra je, da so vsi pomeni iztočnice ves čas prikazani uporabniku – pomeni, v katerih se izbrana kolokacija ali skladenjska struktura ne pojavlja, so namreč zgolj osiveni, ne pa odstranjeni.

Pod Pomeni sledi filter Strukture, ki omogoča filtriranje kolokacij glede na besedno vrsto (prva raven) ali glede na relevantne oblikoskladenjske lastnosti kolokatorja, kot so npr. sklon, stopnja, število ipd.

Zadnji filter v levem stolpcu vmesnika je Predlogi in omogoča filtriranje predložnih struktur. Ker predlogi niso omejeni na določeno besedno vrsto, smo filter podali ločeno in tako uporabnikom omogočili kombinirano uporabo teh dveh filtrov.

Pri filtrih Strukture in Predlogi ima uporabnik lahko vedno izbrano samo eno možnost, ne more npr. hkrati gledati struktur iztočnice s samostalniki in glagoli. Za tako rešitev smo se odločili, ker je glavni namen filtrov omejiti količino podatkov v desnem oknu.

Pomembna lastnost filtrov Pomeni, Strukture in Predlogi je, da se prilagajajo tudi takrat, ko uporabnik manipulira s podatki v osrednjem oknu, npr. ko izbere posamezno strukturo ali kolokacijo. Na ta način filtri opravljajo informativno vlogo o konkretni kolokaciji ali strukturi. Tak način filtriranja slovenski uporabniki že poznajo, saj je bil uporabljen že v vmesnikih korpusov Gigafida, Kres in Gos.[11]

Poleg filtrov v levem stolpcu so na voljo tudi filtri v osrednjem oknu. Stalno aktivni filter je Pogostost, pri katerem uporabnik vleče drsnik proti "redko" oz. "pogosto". Filter je ponujen na vrhu desnega, glavnega okna, ker se nanaša na pogostost oz. redkost kolokatorjev oz. besed v celotnem korpusu, ne pa na pogostost kolokacij. Glavni namen filtra, ki ga bomo (kot ostale funkcionalnosti) še testirali med uporabniki, je omogočiti določenim skupinam uporabnikov dodatno izločanje nerelevantnih kolokatorjev, npr. učiteljem slovenščine kot tujega jezika izločanje redkejših kolokatorjev oz. besed v jeziku.

Dodatni filtri v osrednjem oknu so ponujeni glede na lastnosti iztočnice (ali iztočnice in kolokatorja) zgolj na ravni posamezne strukture, pa še to samo takrat, ko je njihova uporaba glede na lastnosti iztočnice smiselna. Tako je npr. pri pridevniških iztočnicah na voljo filter za moški, ženski in srednji spol.

Po izboru posamezne kolokacije znotraj strukture se uporabniku prikažejo tudi navigacijski gumbi, ki omogočajo enostavno premikanje med sosednjimi kolokacijami. Na ta način se odpravlja potreba po nenehnem drsenju navzdol in klikanju na naslednje kolokacije, ki si jih uporabnik želi ogledati.

Pomembno vlogo v vmesniku ima tudi iskalno okno, ki omogoča iskanje po iztočnicah, kmalu pa je predvidena tudi možnosti iskanja po kolokacijah. Tako bo v primeru, ko bo uporabnik poiskal konkretno kolokacijo, ponujen prikaz, ki bo nekoliko drugačen od prikaza iste kolokacije znotraj posamezne iztočnice – podane bodo namreč informacije o pomenih (v kolikor bodo na voljo) in povezave na kolokacije znotraj iste strukture, in sicer o obeh iztočnicah, ki sestavljata kolokacijo.

KSSS je del portala virov Centra za jezikovne vire in tehnologije Univerze v Ljubljani, s katerim je kljub samostojnemu vmesniku nenehno ohranjena povezava, saj so podatki o morebitnih zadetkih iskanja, ki ga je uporabnik prvotno izvedel v KSSS, v ostalih virih na portalu na voljo prek klika na gumb ob iskalnem oknu.

Vmesnik KSSS je zasnovan za različne digitalne medije, tj. računalnike, tablice in mobilne telefone, z ustreznimi prilagoditvami, kot je npr. omejitev funkcionalnosti pri mobilnih telefonih na račun večje uporabniške prijaznosti.

5. Zaključek in nadaljnje delo

KSSS prinaša v slovenski prostor pomembno novost, in sicer novo različico odzivnega slovarja, katerega značilnost so podatki na različni stopnji izdelanosti, tj. od avtomatsko izluščenih do leksikografsko pregledanih. Na ta način KSSS sledi metodologiji, zastavljeni v Krek et al. (2013) in Gorjanc et al. (2015).

V prihodnjih letih načrtujemo razvoj tako na metodološki in vsebinski kot na predstavitveni ravni. Z vidika metodologije bomo v okviru projekta KOLOS raziskali morebitne izboljšave pri luščenju kolokacij, kot je npr. uporaba metod distribucijske semantike, preizkusili pa bomo tudi luščenje na skladenjsko razčlenjenih korpusih. Načrtujemo tudi vpeljavo novih metod množičenja, predvsem prek igrifikacije. Poleg dodajanja novih gesel KSSS bomo podatke vsebinsko posodabljali z novim korpusnim gradivom, v prvi vrsti iz korpusa Gigafida 2.0, ki bo objavljen konec leta 2018.

Na predstavitveni ravni se bomo posvetili predvsem testiranju vmesnika z različnimi tipi uporabnikov. Tako je v okviru projekta KOLOS že v teku raziskava, v kateri kombiniramo vprašalnike z intervjuji, preverili pa bomo predvsem, katere informacije bi uporabniki želeli oz. potrebovali na prvi strani gesla. Izsledki bodo pokazali, kako izboljšati vmesniško izkušnjo, npr. odkrili morebitne manjkajoče (ali odvečne) dele uporabniškega vmesnika.

Za jezikovnotehnološki razvoj bodo kolokacijski podatki iz KSSS na voljo kot baza podatkov v repozitoriju CLARIN.SI pod licenco Creative Commons 4.0 CC-BY.

6. Zahvala

Prispevek izhaja iz dveh temeljnih raziskovalnih projektov: Kolokacije kot temelj jezikovnega opisa: semantični in časovni vidiki (J6-8255) in Nova slovnica sodobne standardne slovenščine: viri in metode (J6-8256), ki ju financira Agencija za raziskovalno dejavnost Republike Slovenije.

Avtorji se tudi zahvaljujemo podpori infrastrukturnih programov ARRS, in sicer Centru za jezikovne vire in tehnologije Univerze v Ljubljani in Centru za uporabno jezikoslovje pri zavodu Trojina (I0-0051), ter mednarodnemu projektu ELEXIS (European Lexicographic Infrastructure), ki ga finančno podpira evropski program za raziskave in inovacije Obzorje 2020.

Vmesnik predstavljenih virov je razvil Studio Kruh v sodelovanju z Leonom Noetom Jovanom.

[11] http://www.gigafida.net, http://www.korpus-kres.net, http://korpus-gos.net

7. Literatura

eSSKJ: Slovar slovenskega knjižnega jezika 2016–2017, www.fran.si, dostop 28. 08. 2018.

Annelies Häcki Buhofer, Marcel Dräger, Stefanie Meier in Tobias Roth. 2014. *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag*. Tübingen: Francke.

Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Vojko Gorjanc, Bojan Klemenc, Iztok Kosem, Simon Krek, Cyprian Laskowski, Marko Robnik-Šikonja. 2018. Thesaurus of Modern Slovene: By the Community for the Community. V: J. Čibej, V. Gorjanc, I. Kosem in S. Krek, ur., *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, str. 401-410. Ljubljana, Ljubljana University Press, Faculty of Arts. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/2991-1.pdf (dostop 25. 8. 2018).

Polona Gantar in Simon Krek. 2011. Slovene lexical database. V: D. Majchráková in R. Garabík, ur., *Natural language processing, multilinguality*, str. 72–80. Brno, Tribun EU.

Polona Gantar, Simon Krek, Iztok Kosem, Mojca Šorli, Katja Grabnar, Olga Pobirk, Petra Zaranšek in Nina Drstvenšek. 2012. *Leksikalna baza za slovenščino*. Ljubljana, Ministrstvo za izobraževanje, znanost, kulturo in šport. http://www.slovenscina.eu/spletni-slovar/leksikalna-baza, http://hdl.handle.net/11356/1030 (dostop 8. 4. 2018).

Polona Gantar, Vojko Gorjanc, Iztok Kosem in Simon Krek. 2015. Going semi-automatic and crowdsourced: collocation dictionary of Slovene. V: I. Kosem, ur., *Electronic lexicography in the 21st century: linking lexical data in the digital age. eLex 2015, knjiga povzetkov*, str. 37. Ljubljana, Trojina, Institute for Applied Slovene Studies; Brighton, Lexical Computing.

Polona Gantar, Iztok Kosem in Simon Krek. 2015. Leksikografski proces pri izdelavi spletnega slovarja sodobnega slovenskega jezika. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 280–97. Ljubljana, Univerza v Ljubljani, Filozofska fakulteta.

Polona Gantar, Iztok Kosem in Simon Krek. 2016. Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography*, 29(2):200–225.

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek, ur. 2015. *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana, Univerza v Ljubljani, Filozofska fakulteta.

Annette Klosa. 2013. The lexicographical process (with special focus on online dictionaries). V: R. H. Gouws, U. Heid, W. Schweickard in H. E. Wiegand, ur., *Dictionaries. An international Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography,* str. 517–24. Berlin in Boston, de Gruyter.

Adam Kilgarriff, Miloš Husák, Katy McAdam, Michael Rundell in Pavel Rychly. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. V E. Bernal in J. DeCesaris, ur., *Proceedings of the Thirteenth EURALEX International Congress*, str. 425–32. Barcelona, Spain, Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Adam Kilgarriff, Miloš Husak in Miloš Jakubíček. 2013. Automatic collocation dictionaries. Predstavitev na konferenci eLex 2013, Tallinn, Estonija. Dostopno na: https://youtu.be/b3KyhPBeoLU.

Iztok Kosem, Polona Gantar in Simon Krek. 2013a. Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. V: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets in M. Tuulik, ur., *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*, str. 32–48. Ljubljana, Trojina, Institute for Applied Slovene Studies; Tallinn, Eesti Keele Instituut.

Iztok Kosem, Polona Gantar in Simon Krek. 2013b. Avtomatizacija leksikografskih postopkov. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave,* 1(2):139–164. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_07.pdf (dostop 8. 4. 2018).

Iztok Kosem. 2015. Slovarski zgledi. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 320–38. Ljubljana, Znanstvena založba Filozofske fakultete UL.

Iztok Kosem, Polona Gantar in Simon Krek. 2017. Sense menus in collocations dictionary of Slovene. V: *Electronic lexicography in the 21st century: lexicography from scratch,* str. 43. Leiden, Dutch Language Institut; Brno, Lexical Computing; Ljubljana, Trojina Institute for Applied Slovene Studies.

Iztok Kosem, Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, Cyprian Laskowski. 2018. Collocations Dictionary of Modern Slovene. V: J. Čibej, V. Gorjanc, I. Kosem in S. Krek, ur., *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, str. 989-97. Ljubljana, Ljubljana University Press, Faculty of Arts. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/2939-1.pdf (dostop 25. 8. 2018).

Simon Krek, Polona Gantar, Iztok Kosem, Vojko Gorjanc in Cyprian Laskowski. 2016. Baza kolokacijskega slovarja slovenskega jezika. V: T. Erjavec in D. Fišer, ur., *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 29. september - 1. oktober 2016*, Filozofska fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija = *Proceedings of the Conference on Language Technologies & Digital Humanities, September 29th - October 1st*, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia, str. 101–105 Ljubljana, Znanstvena založba Filozofske fakultete: = Ljubljana University Press, Faculty of Arts.

Simon Krek, Cyprian Laskowski, Marko Robnik Šikonja, Iztok Kosem, Špela Arhar Holdt, Polona Gantar, Jaka Čibej, Vojko Gorjanc, Bojan Klemenc in Kaja Dobrovoljc. 2017. *Sopomenke 1.0: Slovar sopomenk sodobne slovenščine.* Znanstvena založba Filozofske fakultete Univerze v Ljubljani. Dostopno na: viri.cjvt.si/sopomenke (dostop 13. 04. 2018).

Nataša Logar Berginc, 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 in Fakulteta za družbene vede.

Carolin Müller-Spitzer, ur. 2014. *Using Online Dictionaries*. Berlin in Boston, de Gruyter.

Tobias Roth. 2013. Going Online with a German Collocations Dictionary. V: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets in M. Tuulik, ur., *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*, str. 152–63. Ljubljana/Tallinn, Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Jože Toporišič, ur. 2001. Slovenski pravopis. Ljubljana: Založba ZRC, ZRC SAZU.

Orsolya Vincze, Estela Mosqueira in Margarita Alonso Ramos. (2011). An online collocation dictionary of Spanish. V: I. Boguslavsky in L. Wanner, ur., *Proceedings of the 5th International Conference on Meaning-Text Theory*, str. 275–86. Barcelona.

Orsolya Vincze in Margarita Alonso Ramos. (2013). Testing an electronic collocation dictionary interface: Diccionario de Colocaciones del Espanol. V: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets in M. Tuulik, ur., *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013 Conference, 17-19 October 2013, Tallinn, Estonia*, str. 328–37. Ljubljana, Trojina, Institute for Applied Slovene Studies; Tallinn, Eesti Keele Instituut.

A Rule-Based Syllabifier for Serbian

Aniko Kovač,* Maja Marković†

* Department of Language Science and Technology, Saarland University
Campus A2 2, 66123 Saarbrücken, Germany
anikok@coli.uni-saarland.de

† Department of English Language and Literature, Faculty of Philosophy, University of Novi Sad
Dr Zorana Đinđića 2, 21000 Novi Sad, Serbia
majamarkovic@ff.uns.ac.rs

Abstract

In this paper, we present an automatic rule-based syllabification algorithm for Serbian based on prescriptive rules from traditional grammar. We explore the problems and limitations of the existing rule set and present the statistical data related to the distribution of syllables and their structure in Serbian.

1. Introduction

Syllables have been considered — although not unequivocally (cf. Koehler, 1996) — to be one of the basic units in phonology constituting the minimal units of pronunciation, and to play a role in prosody, phonotactics, and phonological processing (Ladefoged and Johnson, 2014). The role of the segmentation of words into syllables and their distributional properties began to see an increase in importance in language technology in the 1990s (Iacoponi and Savy, 2011), most notably in the areas of speech recognition (SR) and text-to-speech synthesis (TTS).

The two generally distinguishable approaches to automatic syllabification are rule-based versus data-driven approaches (Marchand et al., 2009). While data-driven approaches have taken over many aspects of natural language processing, and there are a number of data-driven models of syllable segmentation using artificial neural networks (e.g. Daelemans and van den Bosch, 1992; Hunt, 1993; Stoianov et al., 1997; Landsiedel et al., 2011), the unavailability of segmented data for Serbian makes rule-based approaches the only viable option for automatic syllabification in Serbian.

2. The goal of the paper

In this paper, we present a rule-based automatic syllabifier for Serbian. We based our starting set of rules on *Gramatika srpskoga jezika* by Stanojčić and Popović (2005), a prescriptive textbook for Serbian grammar that presents a set of rule descriptions for the segmentation of words into syllables. However, as the formulation of some of these descriptions proved to be redundant, we devised an algorithm for syllabification aimed to produce an output consistent with the rules prescribed in *Gramatika srpskoga jezika*, rather than a verbatim implementation of the formalized rules, with three added modifications related to the treatment of nasals and the alveolar sonorant /*r*/ based on Kašić (2014) and the treatment of alveolar sonorants /*l*/ and /*n*/ based on Zec (2000).

The goal of the paper is threefold: i) to develop a system for automatic rule-based syllabification for Serbian based on the formalization of existing rule descriptions, ii) to provide an analysis of the outcomes of the automatic syllabification process in order to address possible theoretical considerations and serve as a basis for the development of future syllabifiers, and iii) to present statistical data related to the distribution of syllables and their structure in Serbian.

3. The descriptive rule set

Stanojčić and Popović (2005) establish syllables as speech units of the language which can be produced with a single articulatory movement. While there is no consensus on a universal definition of the syllable or what principles should govern the segmentation of words into syllables, there is general agreement that each syllable consists of a syllable-carrying element called *nucleus* which can be preceded by zero or more consonants constituting the *onset* and followed by zero or more consonants making up the *coda*.

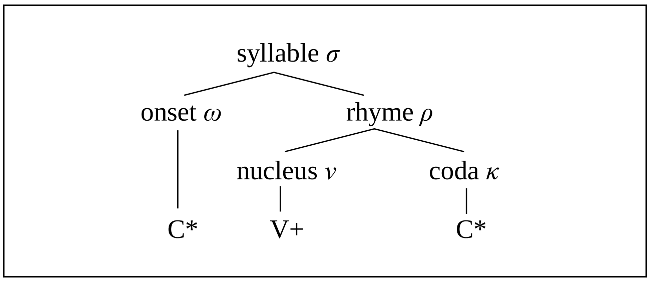


Figure 1: Tree diagram of syllable structure

In accordance with this, Stanojčić and Popović state that syllables in Serbian can be made up of a single phoneme, provided that that phoneme is a vowel. In syllables consisting of multiple phonemes — the nucleus in combination with consonants in the onset and/or coda — the sonorants /*r*/, /*l*/ and /*n*/ can also act as syllable carrying nuclei in Serbian.

Regarding syllable boundaries, Stanojčić and Popović (2005:37) establish the following general rule (1).

(1) *In words made up of multiple phonemes, consonants, sonorants and vowels, the syllable boundary comes after the vowel and before the consonant* (e.g. či-ta-ti [*to read*]).

In addition to this general rule, they list the following rules — (2), (3), (4), (5) and (6) — that further specify medial syllable boundaries depending on consonant manner of articulation.

(2) *Medially, in a consonant cluster which has an affricate or fricative sound in its initial position, the syllable boundary will be before that consonant cluster* (e.g. po-šta [*post*], ma-čka [*cat*]).

(3) *The syllable boundary will be before a consonant cluster if, in a consonant cluster found medially in a word, the second position in the cluster is occupied by one of the sonorants* v, j, r, l *or* lj *preceded by any other consonant besides a sonorant* (e.g. sve-tlost [*light*]).

(4) *If a consonant cluster consists of two sonorants, the syllable boundary will be between them so that one sonorant belongs to the preceding, and one sonorant belongs to the following syllable* (e.g. lom-ljen [*broken*]).

(5) *If a consonant cluster consists of a plosive in its initial position and some other consonant except the sonorants* j, v, l, lj *and* r, *the syllable boundary will be between the consonants* (e.g. lep-tir [*butterfly*]).

(6) *If in a cluster of two sonorants, the second position is occupied by the sonorant* j *from* je *corresponding to the ijekavica dialect to* e *in the ekavica dialect, the syllable boundary will be before that group* (e.g. čo-vjek [*man*]).

The initial member of a consonant cluster in the rule descriptions presented above is understood as the first consonant following a vowel based on the general rule presented under (1). However, a more precise definition would be that the initial member of a consonant cluster is the first consonant following a syllable nucleus — which in the case of Serbian also includes the sonorants /r/, /l/ and /n/ in certain positions. The general rule under (1) should be then revised as follows.

(1*) *In words made up of multiple phonemes, consonants, sonorants and vowels, the syllable boundary comes after the vowel or sonorants* r, l *and* n *in syllable bearing positions and before the consonant* (e.g. či-ta-ti [*to read*], tr-ča-ti [*to run*]).

Stanojčić and Popović (2005: 32) introduce the rule descriptions (7) and (8) to define when the sonorants /r/, /l/ and /n/ constitute syllable nuclei.

(7) *The sonorant* r *can be a syllable carrier in standard Serbian when:*
   a. *it is found medially between two consonants* (e.g. tr-ča-ti [*to run*]),
   b. *it is found initially before a consonant (*e.g. r-va-ti se [*to wrestle*]),
   c. *it is found after a vowel in compounds* (e.g. za-r-đa-ti [*to rust*]),
   d. *before* o *that is realized as an* l *in other members of the paradigm* (e.g. o-tr-o (m.) *from* o-tr-la (f.) [*wiped*]).

(8) *The other two alveolar sonorants,* l *and* n *can be syllable carriers in dialectal toponyms (e.g.* Stlp, Vlča glava, Žlne*) or foreign toponyms (e.g.* Vltava, Plzen*) but also in other personal names (e.g. English* Idn *or Arabic* Ibn-Saud*) and in the word* bicikl [*bicycle*].

3.1. A note on modifications of the original rule set

In addition to our expansion of the general rule presented under (1) to include the syllable bearing sonorants /r/, /l/ and /n/ (1*), the rule descriptions in Stanojčić and Popović (2005) needed to be further modified in the following cases.

While formalizing the rule descriptions via finite-state automata, rules (2) and (3) proved to be redundant as they produced identical outcomes to the general rule (1). Because of this, these rules were disregarded in our syllabification algorithm.

During our early testing of the verbatim implementation of the rule descriptions of Stanojčić and Popović (2005), we noticed that the existing rule descriptions treated a consonant cluster consisting of a nasal in initial position followed by a consonant that is not one of the sonorants /j/, /v/, /l/, /lj/ and /r/ as a part of the following syllable onset, producing outcomes such as: *gu-ngula [commotion]*, *mo-mci [guys]*, *ka-ncelarije [offices]*, *su-nce [sun]*, etc. However, other authors (e.g. Kašić, 2014) argue that nasals should be treated analogously to plosives during syllabification because there is a complete occlusion in the oral cavity during their production. If this principle were to be employed, rule (5) should be revised as follows.

(5*) *If a consonant cluster consists of a plosive or nasal in its initial position and some other consonant except the sonorants* j, v, l, lj *and* r, *the syllable boundary will be between the consonants.*

Following rule (5*), the examples above would then be segmented as: *gun-gula [commotion]*, *mom-ci [guys]*, *kan-celarije [offices]*, *sun-ce [sun]*, etc. As this approach also respects the limitations put forward by the Sonority Hierarchy — even though this version of our syllabifier is not based on the Sonority Sequencing Principle (SSP) — we follow the treatment of nasals by Kašić (2014) in our implementation.

3.1.1. Alveolar sonorant nuclei

One of the most problematic areas of the rules put forward by Stanojčić and Popović (2005) was their treatment of syllable bearing alveolar sonorants under (7) and (8).

We decided against the treatment of /r/ as a syllable nucleus following a vowel in compounds as specified in rule description (7c) as taking morpheme boundaries into consideration would not be a phonological, but rather a morphological criterion of syllabification. We also decided to treat the alveolar sonorant /r/ as non-syllabic before the vowel /o/ that is realized as /l/ in some members of the paradigm (7d) following Kašić (2014) who states that /r/ is no longer systematically treated as a separate syllable in these instances, and that it is pronounced as non-syllabic in words such as *umro [died]*, *groce [throat]* and *otro* *[wiped]*. This means that these words should no longer be segmented as *um-r-o*, *gr-o-ce* and *ot-r-o* as suggested in Stanojčić and Popović (2005), but rather as *um-ro*, *gro-ce* and *ot-ro*.

We have also amended rule (7) for syllable bearing /r/, by further specifying it to exclude /r/ followed by the sequence *je* from being treated as a syllable nucleus as this would be in violation of the rule description under (6) which puts the syllable boundary before a sonorant cluster in words from the ijekavski dialect thus keeping the consonant cluster together.

In order to formalize the rule description under (8) of Stanojčić and Popović (2005) which gave no formal criteria defining when /l/ and /n/ were syllable carriers, we drew on generalizations based on their examples for syllable bearing /l/ (*Stlp, Vlča glava, Žlne, Vlava, Plzen*) and /n/ (*Idn, Ibn-Saud*) and implemented rule (8*) in analogy to the rules defined for the syllable carrying alveolar /r/.

(8*) *The other two alveolar sonorants,* l *and* n, *can be syllable carriers if they are found medially between two consonants, initially before a consonant, or finally after a consonant.*

However, this resulted in outcomes such as: *Be-rn, Ka-rl, erla-jn, Kla-jn, kasa-rn-skim, Linko-ln, Va-jl-dom* etc. In these examples, the sonority of /l/ and /n/ identified as syllable nuclei is lower than the sonority of a consonant in their immediate context — /r/ and /j/ are more sonorous than /n/ and /l/, and /l/ is more sonorous than /n/. Because of this, native speakers do not perceive as there being a syllable constituted around /l/ and /n/ in these contexts. [1] According to Zec (2000), alveolar sonorants can be syllable carriers in Serbian only in contexts in which there is no segment of a higher level of sonority in their immediate vicinity. Because of this, we need to further specify rule (8*) as follows.

(8**)*The other two alveolar sonorants,* l *and* n, *can be syllable carriers if they are found medially between two consonants of lower sonority, initially before a consonant of lower sonority, or finally after a consonant of lower sonority.*

Interestingly, this principle applied to the syllable bearing /r/ could also account for our extension of rule (7) keeping the consonant cluster of the ijekavica dialect unsegmented in initial position — because /j/ is more sonorous than /r/, and then /r/ should not be treated as a syllable nucleus initially in words such as *rjeka [river]*. However, our rule extension has a more general scope than the sonority rule as it also accounts for medial clusters (e.g. in *isko-rje-nilo [eradicated]*).

4. Our algorithm[2]

Our syllabification algorithm consists of the following steps:

i. Identify vowels in the word and mark their positions as positions capable of constituting syllable nuclei.

ii. If a word contains the letters *l*, *n* or the letter *r* not followed by the sequence *je* in the center of a consonant cluster consisting of elements of lower sonority or at the beginning or a word followed by a consonant of lower sonority, or the letters *l* or *n* at the end of a word preceded by a consonant of lower sonority, treat those positions in the word as capable of constituting syllable nuclei.

iii. For each position identified as capable of constituting a syllable nucleus:
    a. If it is followed by a sequence of two sonorants, mark the syllable boundary between the two sonorants, except if the second sonorant is *j* and it is followed by *e*. If the second sonorant is *j* followed by *e*, mark the syllable boundary before the sonorant cluster.
    b. If it is followed by a sequence of a plosive or nasal and a plosive, fricative, affricate or nasal, mark the syllable boundary between the two consonants.
    c. In all other cases mark the syllable boundary after the syllable nucleus.

5. Results

In this section, we present the statistical distribution data for syllables in Serbian based on our syllabification process applied to the Serbian Lemmatized and PoS Annotated Corpus *SrpLemKor* (Popović, 2010; Utvić, 2011). We chose *SrpLemKor* for our analysis, because its annotation allowed us to filter out numbers, Roman numerals, abbreviations and non-Serbian words or suffixes in compounds (at least to some extent) and thus reduce noise in the data.

The following results show the syllable distribution statistics based on 3,607,450 word-forms in *SrpLemKor*. From a total of 4,681,713 entities in our version of the corpus, 113,679 (2.43%) entities of texts #260, #4505 and #4517 were excluded because the files contained faulty encoding. Based on corpus tags, we excluded 947,666 (20.24%) entities tagged *PUNCT* (punctuation), *SENT* (sentence separator full-stops), *RN* (Roman numerals), *NUM* (numbers), *ABB* (abbreviations) and *?* (non-Serbian words and other uncategorized entries). An additional 551 (0.01%) entities that contained the characters *w* and *q* were removed in an attempt to further reduce noise stemming from foreign words, as not all foreign words were tagged as such in the corpus. In the process of syllabification, an additional 12,910 (0.28%) entities were removed as they were solely made up of consonant clusters with no available syllable nucleus candidate.

5.1. Syllable type distributions in Serbian

In the 3,607,450 word-forms from *SrpLemKor*, a total of 8,147,679 syllables were identified. Table 1 presents the

[1] We thank Miloš Košprdić for his insight and helpful discussion on this topic.

[2] Our implementation of the algorithm can be found at https://github.com/versi-regular/rule-based_syllabifier_sr, licensed under the GNU General Public License v3.0.

syllable type distribution based on our syllabification algorithm.

Syllable structure	No. of instances	Percent
CV	5034567	61.791
CCV	1009791	12.394
V	863631	10.6
CVC	771143	9.465
CCVC	215267	2.642
VC	131021	1.608
CCCV	69577	0.854
CCCVC	21151	0.26
CVCC	17210	0.211
CCVCC	6487	0.08
CCCCV	4292	0.053
VCC	1526	0.019
CCCCVC	708	0.009
CVCCC	705	0.009
CCCVCC	391	0.005
VCCC	66	0.001
CCVCCC	32	0
CCCCVCC	23	0
CCCVCCC	14	0
CCCCCV	3	0
CCCCCVC	2	0
Other	73	0.001
Total	8147679	100

Table 1: Syllable structure distribution for syllables in the *SrpLemKor* corpus

These results show the distribution of syllables in a somewhat noisy data. We found that there are still foreign words annotated as non-foreign in the corpus constituting some of the less-frequent syllable structures listed as "Other" in Table 1. For example, we found one instance of the structure CCCCVCCC from the German word *Fleischmarkt* [*meat market*], one example of the structure CCCCCCVC from the German *Nachtschatten* [*nightshade*], a single entry CCCCCCCV from the German word *Storchschnabel* [*Crane's-bill*], one instance of the structure CCCCCCVCC from the English *healthystuff*, 4 examples of the structure VCCCCC from two occurrences of the German words *Peitscht* [*lashes*], one instance of *staruch* (typo or possibly Polish [*old man*]) and one instance of the English word *knights*. We also found 10 instances of the structure VCCCC from the German *Ernst* [*seriousness*], *Deutsch* [*German*], and strings such as *ikvbv*, which we assume stand for unfiltered acronyms, and strings we could not associate with any meaning such as *ehmc* and *rhutm*. We have also identified one example of the sequence CVCCCCCCCC to stand for the onomatopoeic vulgarism *mršššššššš* [*go away*].

Besides these, we found 6 types of syllable structure that differed from the structures found by Meštrović et al. (2015) for Croatian. The structures CCCCCVC (e.g. *mo-na-rhstvom [with the monarchy])*, CCCCV (e.g. *se-rbska [Serbian]*, *ca-rstva [kingdoms]*, *sta-ra-te-ljstva [custody]*) and CCCCVC (e.g. *se-rbskom [Serbian]*, *de-jstvom [with effect]*, *vo-đstvom [leadership]*, *spo-rtskim [sport]*, *a-lpskog [alpine]*) represented Serbian entities and are in accordance with the syllabification rules, but present some theoretical issues which we discuss in section 6. In the case of the structure CCCCCV, we separated the counts to include *se-rbstvo [Serbian]* as a problematic but valid entry, but exclude counts resulting from typos (e.g. *ri-va-ststva*, *su-žnjstva*, *šttske*) and foreign words (e.g. *ba-ckstre-et*) which were counted as "Other". The structure CCCCVCC found in foreign origin names (e.g. *Go-ldštajn, Rot-hchild*, *Ar-mstrong*), and the structure CVCCCC, a result of typos (e.g. *slav-janskh*, *cr-no-gorskg*), were also counted under "Other" in Table 1.

5.2. Syllable type positional distributions in Serbian

We also examined the syllable type frequencies with respect to their position in a word. Four positional frequencies are presented in Table 2: syllable type frequencies in monosyllabic words, and syllables type frequencies in the initial position, in medial positions, and in the final position of polysyllabic words.

Syllable structure	Monosyllabic words		Polysyllabic words					
	MONO		INITIAL		MEDIAL		FINAL	
	No. of instances	Percent	No. of instances	Percent	No. of instances	Percent	No. of instances	Percent
CV	612244	50.784	1398930	58.244	1486143	69.499	1537250	64.002
CCV	54417	4.514	376527	15.676	351099	16.419	227748	9.482
V	301295	24.991	379122	15.785	62176	2.908	121038	5.039
CVC	128321	10.644	121162	5.045	155947	7.293	365713	15.226
CCVC	35434	2.939	44923	1.87	47315	2.213	87595	3.647
VC	64037	5.312	57451	2.392	6210	0.29	3323	0.138
CCCV	177	0.015	20012	0.833	24708	1.155	24680	1.028
CCCVC	1490	0.124	3715	0.155	3950	0.185	11996	0.499
CVCC	4666	0.387	0	0	0	0	12544	0.522
CCVCC	1638	0.136	0	0	0	0	4849	0.202
CCCCV	9	0.001	19	0.001	750	0.035	3514	0.146

VCC	1100	0.091	0	0	0	0	426	0.018
CCCCVC	4	0	0	0	46	0.002	658	0.027
CVCCC	568	0.047	0	0	0	0	137	0.006
CCCVCC	104	0.009	0	0	0	0	287	0.012
VCCC	42	0.003	0	0	0	0	24	0.001
CCVCCC	12	0.001	0	0	0	0	20	0.001
CCCCVCC	1	0	0	0	0	0	22	0.001
CCCVCCC	1	0	0	0	0	0	13	0.001
CCCCCV	0	0	0	0	0	0	3	0
CCCCCVC	0	0	0	0	0	0	2	0
Other	36	0.003	1	0	16	0.001	20	0.001

Table 2: Syllable structure distribution for syllables in the *SrpLemKor* corpus categorized by position

Based on *SrpLemKor*, the most frequent monosyllabic syllable structures in Serbian are CV (51%), V (24%) and CVC (11%). The most frequent syllable structures in the initial position of polysyllabic words are CV (58%), V (16%) and CCV (16%). In medial positions in polysyllabic words, the most frequent syllable structures are CV (70%), V (16%) and CVC (7%). The most frequent syllable structures in the final position of polysyllabic words are CV (64%), CVC (15%) and CCV (10%).

It is interesting to note the asymmetry that the syllable structures CVCC, CCVCC, VCC, CVCCC, CCCVCC, VCCC, CCVCCC, CCCCVCC and CCCVCCC occurred only in monosyllabic words and in the final position of polysyllabic words, while the syllable structure CCCCVC occurred in all positions except the initial position in polysyllabic words. The rare (and problematic) structures CCCCCV, CCCCCVC occurred only in the final positions of polysyllabic words.

5.3. Syllable nuclei statistics in Serbian

The distribution of different syllable nuclei in Serbian based on the *SrpLemKor* corpus is presented under Table 3.

Nucleus	TOTAL		Monosyllabic words		Polysyllabic words					
			MONO		INITIAL		MEDIAL		FINAL	
	No. of instances	Percent	No. of instances	Percent	No. of instances	Percent	No. of instances	Percent	No. of instances	Percent
a	2166178	26.586	327721	27.183	604299	25.160	585064	27.360	649094	27.025
o	1747318	21.446	167750	13.914	671083	27.940	385403	18.023	523082	21.778
i	1725046	21.172	228055	18.916	394426	16.422	599859	28.052	502706	20.930
e	1620813	19.893	300701	24.942	430654	17.930	393488	18.401	495970	20.649
u	797667	9.790	178664	14.820	234319	9.756	155017	7.249	229667	9.562
r	88233	1.083	1966	0.163	66435	2.766	19383	0.906	449	0.019
n	1411	0.017	411	0.034	602	0.025	50	0.002	348	0.014
l	1014	0.012	328	0.027	44	0.002	96	0.004	546	0.023

Table 3: Syllable nuclei statistics and positional frequencies for syllables in the *SrpLemKor* corpus

Based on the positional nucleus distribution data, it can be seen that overall /*a*/ and /*o*/ constitute the most frequent nuclei in Serbian. However, there is some positional variation. While the most frequent nuclei in final position are also /*a*/ and /*o*/, and /*o*/ and /*a*/ represent the most frequent nuclei in the initial position of polysyllabic words, in monosyllabic words, the most frequent nuclei are /*a*/ and /*e*/, while in the medial positions in polysyllabic words, the most frequent nuclei are /*i*/ and /*a*/.

6. Discussion

In the previous section, we mentioned that the 3,607,450 word-forms extracted from *SrpLemKor* used for the calculation of statistical data related to the distribution of syllables and their structure in Serbian still contained some noise such as foreign words, acronyms, typos, and possibly random character strings. Based on 500 random samples taken from the syllable output data checked by a human evaluator, the estimate of the amount of such noise in the data is <2%.

While our syllabifier is suitable for the segmentation of words into syllables following the set of provided rule descriptions, we argue that the prescriptive rules themselves need revising as they seem to violate basic phonetic and phonotactic principles of the language.

In their automatic syllabification system for Croatian based on the Onset Maximization Principle, Meštrović et al. (2015) limit possible onsets in medial and final clusters to those onsets which occur in word-initial positions with some extensions to the allowed onsets following the principle of analogy by place of articulation and taking into account voiced and voiceless consonant pairs. While we remain uncertain whether initial occurrences should be used as a criterion for medial and final onsets, we are interested in exploring the possibility of an onset maximization segmentation based on Meštrović et al. (2015), but limited by the prescriptive rules used in the syllabifier presented in the paper. For example, this would mean that some questionable onsets such as /*pn*/ in *va-pno* [*lime*], which they allow for because /pn/ constitutes a valid onset in the word *pneumatski* [*pneumatic*], would be disallowed and segmented as *vap-no* in such a system because of rule (5) that defines a syllable boundary between a plosive and subsequent consonant that is not one of the sonorants /*j*/, /*v*/, /*l*/, /*lj*/ and /*r*/.

In order to verify the syllabic status of different clusters, it would be interesting to conduct a series of monitoring studies modeled after Mehler et al. (1981), who have shown that reaction times to a word are faster if the word is primed by a sequence corresponding to a syllable in the word when compared to priming with a string that does not constitute a syllable. Bradley et al. (1993) argue that these effects produce mixed results in some languages which contain a large number of ambisyllabic segments, so these studies may also reveal whether and to what extent syllables play a role in pre-lexical processing in Serbian.

One of the main problems that we have identified with a syllabifier based on the set of prescriptive rules presented in section 3 is that even with the revised rule set, the results are often problematic when taking into account the viewpoint that the structure of syllables should be in accordance with the Sonority Sequencing Principle. Namely, if we assumed that syllables are structured in such a way that there is a rising sonority of elements in the onset leading up to the nucleus, examples such as some of the problematic cases presented in section 5 (e.g. *se-rbska [Serbian]*, *de-jstvom [with effect]*) clearly violate the Sonority Hierarchy as alveolar sonorants have a higher sonority level than plosives and fricatives.

One way in which we attempted to remedy this was to introduce a limit of onset length to three-syllable clusters, which is the maximum length of non-syllabic consonant clusters word initially in Serbian (Kašić, 2014). While this — in combination with rules (5) and (6) — would indeed resolve the issues in the examples we encountered — they would be segmented as *serb-ska* and *dej-stvom* — medial clusters with a syllabic consonant would still present a problem. For example, the word *najstrpljiviji* [*most patient*], which contains a syllabic /*r*/ at the beginning of the hypothesized three-syllable maximum onset, would result in a boundary at *najst-rpljiviji* which is incorrect when taking into account the syllabic status of /*r*/. It would be interesting to see whether an added rule to separate elements with sonority violations might amend the existing rule set and resolve these problems, and compare the results stemming from this rule to the results of a rule limiting the range of possible onsets.

7. Conclusion

In this paper, we presented a rule-based syllabifier modelled after the rule descriptions found in Stanojčić and Popović (2005) and extended by rule specifications from Kašić (2014) and Zec (2000).

An implementation of the existing prescriptive rules for the segmentation of words into syllables allowed us to gain an insight into the problem areas of the rule descriptions, and propose a number of revisions and amendments to the existing rules. We have also gained an insight into the distribution of different syllable structures and syllable nuclei following this approach, which will be useful for comparison with the performance of alternative syllabification systems.

In the future, we plan to improve our system by developing an onset-maximization-based syllabifier as well as a sonority-based syllabifier for Serbian, and then test a combination of these with the prescriptive rules to see if we can create a hybrid system that will produce outputs consistent with the intuition of native speakers of Serbian.

We also believe that, while phonological criteria present a basis for syllabification, in the future we might also need to test whether subsequent approaches coincide with morphological boundaries, or whether the phonological rules need to be amended to respect morphological boundaries as well.

In addition to these issues, the question of the treatment of foreign origin words and transcribed foreign words might be an additional point to consider. As an extension of a syllabifier, a language detection algorithm might be employed to properly segment the former, while the latter might not need special treatment as the process of transcription should in itself contain a degree of phonological adaptation.

8. References

Dianne C. Bradley, Rosa M. Sánchez-Casas, and José E. García-Albea. 2007. The status of the syllable in the perception of Spanish and English. *Language and Cognitive Processes*, 8(2): 197–233.

Andrew Hunt. 1993. Recurrent Neural Networks for Syllabifiation. *Speech Communication* 13(3–4):323–332.

Walter Daelemans and Antal van den Bosch. 1992. Generalization performance of backpropagation learning on a syllabification task. In: *Connectionism and natural language processing: Proceedings of the third Twente Workshop on Language Technology, TWLT3*, pages 27–38, Enschede, the Netherlands. https://pure.uvt.nl/portal/files/760578/generalization.pdf .

Luca Iacoponi and Renata Savy. 2011. Sylli: Automatic Phonological Syllabification for Italian. In: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, pages 641–644, Florence, Italy. http://eden.rutgers.edu/~li51/php/papers/interspeech2011.pdf.

Zorka Kašić. 2014. *Opšta lingvistika 2 (Fonologija)*. Lecture Materials, Faculty of Philosophy, University of Belgrade.

Kenneth J. Koehler. 1996. Is the syllable a phonological universal? *Journal of Linguistics*, 2:207–208.

Peter Ladefoged and Keith Johnson. 2014. *A Course in Phonetics*. Wadsworth Publishing.

Christian Landsiedel, Jens Edlund, Florian Eyben, Daniel Neiberg, and Björn Schuller. 2011. Syllabification of conversational speech using Bidirectional Long-Short-Term Memory Neural Networks. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5256–5259, Prague, Czech Republic. http://ieeexplore.ieee.org/abstract/document/5947543.

Yannick Marchand, Connie R. Adsett, and Robert I. Damper. 2009. Automatic syllabification in English: A comparison of different algorithms. *Laguage and Speech* 52(1):1–27.

Jacques Mehler, Jean Yves Dommergues, and Uli Frauenfelder, Juan Segui. 1981. The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior,* 20(3): 298–305.

Ana Meštrović, Sanda Martinčić-Ipšić, Mihaela Matešić. 2015. Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik. *Govor,* 32:3–34.

Zoran Popović. 2010. Taggers Applied on Texts IN Serbian. *INFOtheca* 11(2):21a–38a.

Živojin Stanojčić and Ljubomir Popović. 2005. *Gramatika srpskoga jezika*. Zavod za udžbenike i nastavna sredstva Beograd.

Ivelin Stoianov, John Nerbonne, and Huub Bouma. 1997. Modelling the phonotactic structure of natural language words with Simple Recurrent Networks. In: *Computational Linguistics in the Netherlands 1997: Selected Papers from the Eight Clin Meeting*, pages 77–95.

Miloš Utvić. 2011. Annotating the Corpus of Contemporary Serbian. *INFOtheca*, 12(2):36a–37a.

Draga Zec. 2000. O strukturi sloga u srpskom jeziku. *Južnoslovenski filolog*, 56(1-2):435–448.

Debating Evil

Using Word Embeddings to Analyze Parliamentary Debates on War Criminals in The Netherlands

Milan M. van Lange, Ralf D. Futselaar
NIOD, Institute for War, Holocaust and Genocide Studies
Herengracht 380, 1016CJ Amsterdam, The Netherlands
m.van.lange@niod.knaw.nl, r.futselaar@niod.knaw.nl

Abstract

We are proposing a method to investigate changes in historical discourse by using large bodies of text and word embedding models. As a case study, we investigate discussions in Dutch Parliament about the punishment of war criminals in the period 1945-1975. We will demonstrate how word embedding models, trained with Google's Word2Vec algorithm, can be used to trace historical developments in parliamentary vocabulary through time.

Keywords: War Criminals, Penal History, Word2Vec, Word Embedding Models

1. The case: War Criminals

Soon after German forces in the Netherlands surrendered in May of 1945, the question arose how the hundreds of suspected war criminals and thousands of Nazi collaborators in Dutch custody were to be treated. For the next five decades, this question caused a series of heated political controversies. The debates in Dutch parliament about the punishment, penalty reduction, or release of these people are not only among the longest debates in Dutch parliamentary history, but are generally considered to have been the most emotionally charged (Bootsma and van Griensven, 2003; Futselaar, 2015; Tames, 2013).

1.1. Discourse and controversy

In this paper, we use an implementation of word embedding models (WEMs) to analyze parliamentary discussions concerning incarcerated war criminals and Nazi collaborators after the end of the German occupation. At peak, in the summer of 1945, more than a hundred thousand people were incarcerated. They were accused of a variety of crimes, all committed during the occupation of the country: political and military collaboration, war crimes, and (complicity in) genocide. The overwhelming majority of these prisoners were released quickly, but a small and dwindling number remained in prison until 1989. After the 1960s, all remaining prisoners were former German officials and officers whose initial death sentences had been commuted to life in prison. As long as they remained behind bars, political controversy about plans for their release continued to resurface (Tames, 2013; Piersma, 2005).

We map the language used in Dutch parliament to discuss this specific case during a relatively short historical period. The results will enable us to track the preferred vocabularies in these discussions through time. In other words, we use the words spoken in plenary sessions of the Dutch parliament as a reflection of the vocabulary used. This vocabulary, in turn, we assume reflects the changing discourse about incarcerated war criminals in Dutch society. Thus, we aim to link these developments in parliamentary vocabulary to actual historical events, developments concerning the post-war dealing with war criminals, and discursive shifts in Dutch society (Olieman et al., 2017). Specifically, we aim to investigate the changing political attitude towards incarcerated war criminals and use our findings to test established notions prevalent in Dutch historiography.

The published proceedings provide us with a dataset comprising of all the words spoken in plenary sessions in both houses of parliament. The completeness of the parliamentary dataset allows us to investigate the changing parliamentary vocabulary through time, and in the context of different discussions. This vocabulary changed, and we use these changes to investigate, ultimately, the changing discourse in postwar Dutch society.

We here focus on two questions directly related to the treatment of these delinquents in the Dutch penal system. The first of these concerns the focus on the identification of the wronged party: did politicians focus on crimes against the dutch nation as a whole, or against specific groups of individual victims? The second concerns the appropriateness of harsh punishments, specifically whether or not life imprisonment was considered a just alternative for the death penalty. These questions both derive directly from historiography and serve to answer an overarching question: can we assess the validity of traditional scholarship using unsupervised text mining?

2. Parliamentary proceedings

In this investigation, we rely entirely on parliamentary proceedings, known in Dutch as the *Handelingen der Staten-Generaal*. The Handelingen are available in machine-readable form. The minutes of both houses of parliament for the period 1814-1995 were first digitised by the Royal Library of the Netherlands and made available to the public in 2010. The dataset for the period since 1946 was dramatically improved in the *political mashup* project that ran from 2012 to 2016. This improved and enriched dataset is freely available, on request, from DANS, the Dutch national repository of research data. The dataset consists of a large collection of XML files containing the complete minutes of all the meetings of the lower and upper chambers of parliament, separated by date, speaker, political affiliation, etc. This makes it an excellent corpus for various forms of automated text analysis.[1]

3. Word Embedding Models and Historical Research

We investigate the vocabularies used in parliament to discuss a broad category of inmates that could be described as political delinquents, as well as the changes of these vocabularies through time. This is a fairly normal investigation undertake in traditional historical research, that is to say without computational analyses. Historians typically work by reading the relevant texts. This approach has several disadvantages. In this particular case the corpus to be assessed is enormous, making manual encoding of text problematic. More importantly, the traditional research process is highly vulnerable to the biases of the reader/researcher. When studying ethically charged controversies in the relatively recent past, this vulnerability to bias is evidently problematic.

3.1. Words in vector space

A WEM provides a possible solution to these problems. WEMs are techniques to investigate words, and relations between words, in large text corpora. More specifically, WEMs are based on the calculation of the average distance of unique words to all other unique words in a corpus. This results in a list of numerical values, that make up the 'vector' for each word. In principle, the number of values, also referred to as 'coordinates', or 'dimensions' of the vector, is the same as the the number of unique words in the text, minus one. The complete trained corpus, or 'spatial model', is often referred to as a vector space. Within this space, the position of a specific word relative to all other words, is described by its vector.

Since the position of unique words relative to other words is an average calculated on the basis of all occurrences in the text, WEMs are exceptionally effective at investigating relations between relatively frequent words in a sufficiently large text corpus. The method does not prioritize any particular words; the position of each unique word is investigated. Obviously, many close relationships occur only once or a few times. Other relationships appear frequently. Some words are synonyms or near synonyms, have very similar usages (tea and coffee, for example) or often appear in combination (New and York). The analytical possibilities of WEMs, as we will demonstrate below, go far beyond mere closeness. With WEMs we are able to identify associations between words that are not self-evident.

3.2. Limitations of WEMs

WEMs also have an important downside that is particularly relevant to historical research. Since the training of the model determines the position of a word relative to all other words in that specific corpus, its vector is meaningless in any other model. Word vectors, hence, can only be compared with other word vectors within the same spatial model. For historians, this means that comparisons between different moments in time are likewise impossible, because each period in time would result in a different 'bag of words' and hence a different, and incomparable, spatial model. This means that, while WEMs are perfectly adequate tools for fulfilling the first of our aims, investigating vocabularies, they are virtually useless for the second aim, investigating change through time. Since change through time is the core of virtually all historical research (including this investigation), this presents us with a major problem; how can we compare outcomes for different WEMs, for different periods in time? We have, however, developed a workaround to enable us to use WEMs to investigate changing ways to talk about certain topics through time, about which more below.

3.3. Word2Vec

For this investigation, we have have used the relatively popular Word2Vec implementation of WEMs to train and analyze word embedding models. Word2Vec was developed by a team of Google engineers and published in 2013. It has been shown to be a particularly effective implementation. This algorithm, however, was developed with a different aim than the one for which we are using it. Initially, Word2Vec was a tool to investigate natural language itself, for example to identify (near) synonyms. In our, historical, investigation, the statistical modeling of language as such is not the objective. Rather than trying to identify linguistic regularities to investigate language, we focus on linguistic irregularities and patterns to identify the influence of political and historical change on changes in the language used in political speech.

For researchers using the R language, a package is readily available to analyse texts. This package, created and maintained by Benjamin Schmidt, has been used in this investigation as well (Schmidt, 2015). Our method, however, is in no way dependent on this particular platform and could also be used in Python or any other environment. Neither is the method reliant on the Word2Vec algorithm. It would work broadly in the same way with another implementation of word embeddings. Here, however, we have chosen to use a popular WEM implementation in a relatively user-friendly and accessible environment, with the added benefit of using open-source, free software.

4. Analytical process

Text analysis with WEMs involves two necessary steps. The first of these, the training of the corpus, creates the spatial model, the WEM itself. The second step is the analysis of the positions of specific words or word clusters within the virtual space of the model.

The corpus of the Handelingen is vast by the standards of historical research, but not very large for the kind of anal-

[1] Maarten Marx, Johan van Doornik, Andre Nusselder and Lars Buitinck, *Dutch Parliamentary Proceedings 1814-2012, non-semanticized*, (October 10, 2012), Distributed by DANS EASY, https://doi.org/10.17026/dans-xk5-dw3s

ysis we are undertaking. For the purpose of WEMs, the size is barely adequate. Therefore we have trained our dataset with a Skip-Gram Word2Vec model, which has anecdotally been shown to yield better results on smaller samples (Gelbukh, 2015).

Within the model, the vectors of different words can be compared by using cosine similarity. Within a vector space, any two vectors by can be described, by definition, as lying within a horizontal plane. Cosine similarity calculates the angle between these vectors. Perfectly overlapping vectors would result in a cosine similarity of 1, a perfectly opposite relationship -1. In practice, WEMs consist only of positive space, which means that scores fall between 0 (low, or no similarity) and 1 (high, or perfect) similarity (Singhal, 2001).

4.1. Training the models

The first step of our workaround is to train two WEMs (more than two is equally feasible), based on two corpora (in this case 1945-1955 and 1965-1975). Each of these corpora contains ten years of parliamentary speeches. (When using this approach, it is necessary to use relatively similar corpora, both in terms of size and in terms of language use. For historical research into relatively short periods of parliamentary history, this is not particularly problematic.) For reasons of efficiency, we have limited ourselves to unique words that appear at least five times in the corpus and we have limited the number of dimensions of each vector to one hundred. This allows this investigation to be undertaken, and repeated, using fairly normal office-grade hardware. We have experimented with more dimensions (several hundreds), but more vectors appear only to be useful with larger files and require far more computational power.

4.2. Analyzing word vectors

Within each spatial model, we have identified the 250 words with the highest cosine similarity to the Dutch terms for 'war criminal' (singular and plural, see table 1). With these 250 nearest neighbors, we have defined the time-specific vocabulary used in the discussion of war criminals. Obviously, these are not the same 250 words in each model.

To identify changes in the discussions surrounding our topic, we calculated the cosine similarity of each of the 250 nearest-neighbor words in each model to two different terms that are present in each of the two corpora. This allows us to compare the position of the vocabulary of the discussion on our topic (war criminals) in relation to, in this case, two stable concepts. The selection of these concepts is crucial for our investigation and for this method. It is here that we translate our research question into a formal, computational inquiry.

For now, we have chosen a two-dimensional implementation of this technique. This is not theoretically necessary, but it allows us to visualize and analyze results more easily in two dimensions. What is important is that concepts used to investigate the relative position of each investigated word are the same in each of the models to be compared. It is also necessary that the concepts are relatively stable through time. Since concepts are represented by words in the corpus itself, words that shift meaning dramatically, such as the English word 'gay' are less suitable than 'cheerful' or 'homosexual', which have not undergone such dramatic change over time.

When discussing concepts, the number of possible words referring to the same concept is often greater than one. Since our investigation focuses on concepts that may be described with multiple words, we need to create a so-called combined vector. We used synonyms and plurals to create a cluster of words with the shared meaning of the concept of interest. This cluster was used as a combined vector in the model by calculating the mean of all the vectors of the cluster words. That is to say that this word set was treated as a single term, resulting in a vector of similar length to a single-word vector. This combined vector allows us to investigate our corpus using all synonyms and near-synonyms of terms as if they were a single term, with a single vector.

After selecting two concepts that are present in each of the two corpora, we can calculate the relative similarity of other terms in the corpus to each of them. Although vectors between the two trained WEMs are not comparable, the relative distance to two or more other vectors can be compared very well across several models, provided the underlying concepts are historically stable. When the terms used to estimate the relative position of vocabularies are related and dissimilar, or even perfectly opposite, an historically meaningful analysis becomes viable.

Concept	**Concept represented by combined vector of the Dutch words:**
Death penalty	'doodstraf' and 'doodstraffen'
Life imprisonment	'levenslang', 'levenslange', 'vrijheidsstraf', 'gevangenisstraffen', 'gevangenisstraf', 'opsluiting', 'hechtenis'
Treason/traitor	'landverrader', 'landverraders', 'verrader', 'verraders' and 'landverraad'
Victim	'slachtoffer' and 'slachtoffers'
War criminal	'oorlogsmisdadiger' and 'oorlogsmisdadigers'

Table 1: Word sets used in Debating Evil

Using two concepts allows us to plot our 'vocabulary', that is the top 250 war-criminal-related words in each of the two periods, in a two-dimensional space. Figure 1 and 2 show the similarity scores of each of the 250 word vocabularies relative to one concept that serves as the y axis, and another on the x axis. Each point represents one of the 250 words that form the war-criminal vocabulary for a specific time period. They are plotted based on their cosine similarity score to the combined vector of the concept 'victim' (x) and 'treason' (y) in figure 1, and to 'life imprisonment' (x) and 'death penalty' (y) in figure 2. The average scores of all 250 war criminal words on the two dimension are shown as horizontal and vertical lines. Thus, we have arrived at a visual representation that allows for a comparison of word embedding results for more than one corpus and hence for a comparison through time. (In this case, between two distinct periods.)

5. Results

Here, we present only two examples using four concepts and two time periods (1945-1955 and 1965-1975). Specifically, we try to identify differences in the way incarcerated war criminals and collaborators were discussed in the immediate aftermath of the Nazi occupation of the Netherlands, and at the height of controversies surrounding the intended release of a number of German war criminals from Dutch prisons - namely Kotälla, Aus der Fünten, and Fischer (Piersma, 2005).

Obviously, the discussions in the two periods refer to different groups of perpetrators. In the immediate aftermath of the Nazi occupation the population of inmates was large and diverse, consisting of small-time war profiteers, minor collaborators and their families, but also mass murderers. In the second period, only a handful of elderly foreigners were left, whose crimes were relatively similar and also similarly egregious.

For this investigation, however, our primary aim is not to unearth radically new insights into postwar penal policy in the Netherlands, but to confront the results of an unsupervised, 'distant' reading of parliamentary records to an established historiography. Such an historiography is available for the case at hand; Dutch historians have identified a number of trends in the thinking about political delinquents that (if true) should be reflected in these discussions. Two changes have been identified in particular:

1. The shifting focus from the nature of the crime committed and the person of the perpetrator towards the lasting, psychological damage endured by the victims (de Haan, 1997; van der Heijden, 2011).

2. A decline in the support, both public and political, for harsh, vengeful punishments, exemplified here in the discussions about the propriety of the death penalty. Although the death penalty was (again) abolished in the 1950s, it remained a point of discussion with regard to war criminals in custody. (Futselaar, 2015; Smits, 2008).

5.1. Historical case

Over the course of three decades, attitudes to incarcerated war criminals, as represented by the vocabularies used to discuss them, changed. In the first period the emphasis lay on crimes against the collective, whereas the focus changed to the plight of individual victims. As can be seen in figure 1, the initial emphasis on crimes against the nation (treason) in debates about war criminals clearly declined. The average cosine similarity between war-criminal words and treason words (horizontal lines) decreased significantly when we compare 1945-1955 to 1965-1975. At the same time, we observed increased levels of closeness in vector space between war criminal related words to words associated with (individual) victims, as can be seen in figure 1.

This observation is in line with the relevant historiography. Several authors have emphasized the sharp rise of interest into the mental health of individual war victims and and their families as a decisive factor in policy making and the formation of political opinion. This also indicates a shift in discourse from focusing on the initial crimes, committed by the war criminals, to the consequences of their deeds for individual people involved (de Haan, 1997; van der Heijden, 2011; Smits, 2008; Withuis, 2002).

This development can not be considered a mere discursive change: the observed shifts in parliamentary vocabulary represent actual historical developments in the postwar dealing with war criminals. In the early 1970s, the only war criminals remaining in Dutch prisons were German nationals. Whereas in 1945, main part of the more than hundred thousand incarcerated war criminals were Dutch citizens. Evidently, the accusation of treason was only applicable to the latter group. Hence, if we compare the two periods, it is not surprising that the discursive element of 'treason' evaporates from the war criminal vocabulary in Dutch parliamentary debate between 1965 and 1975.

It remains imperative to remain aware of the possible pitfalls of this type of investigation. This is evident in the sharp rise of references to the death penalty in war criminal vocabulary that we observed (see figure 2). During the second period under scrutiny, capital punishment had long been discontinued in the Netherlands and could not have been discussed as a serious penal option. Closer scrutiny of the data revealed that in many discussions, capital punishment was not advocated, but merely used as a reference point. The war criminals in question had originally been condemned to die, but their punishment had been commuted into life imprisonment. Several members of parliament felt that a pardon would mean that the original verdict (death penalty) would be watered down twice. In these discussions, capital punishment was often referenced, even when its use was not a viable (or even legal) option.

6. Conclusion

This paper outlines a method for studying discursive changes in history. We trained WEMs and calculated cosine similarities between two opposite or related concepts for specific periods. This enabled us to compare WEMs for different periods. This opens the door for the use of word embeddings as a tool for historical research, because it enables us to investigate change through time in sufficiently large and consistent historical datasets. Parliamentary records are perhaps the best example of such datasets. As such, this method holds considerable promise in a period when parliamentary proceedings and other historical sources are increasingly made available in machine readable form.

We have shown how developments in vocabulary can be considered reflective of discursive changes. These changes coincide with related historical events and developments in the post-war dealing with war criminals in Dutch society. The war criminal vocabulary shifted from focusing on the act of crime committed by war criminals towards the consequences of these deeds for victims and relatives. We also showed how actual historical developments regarding the

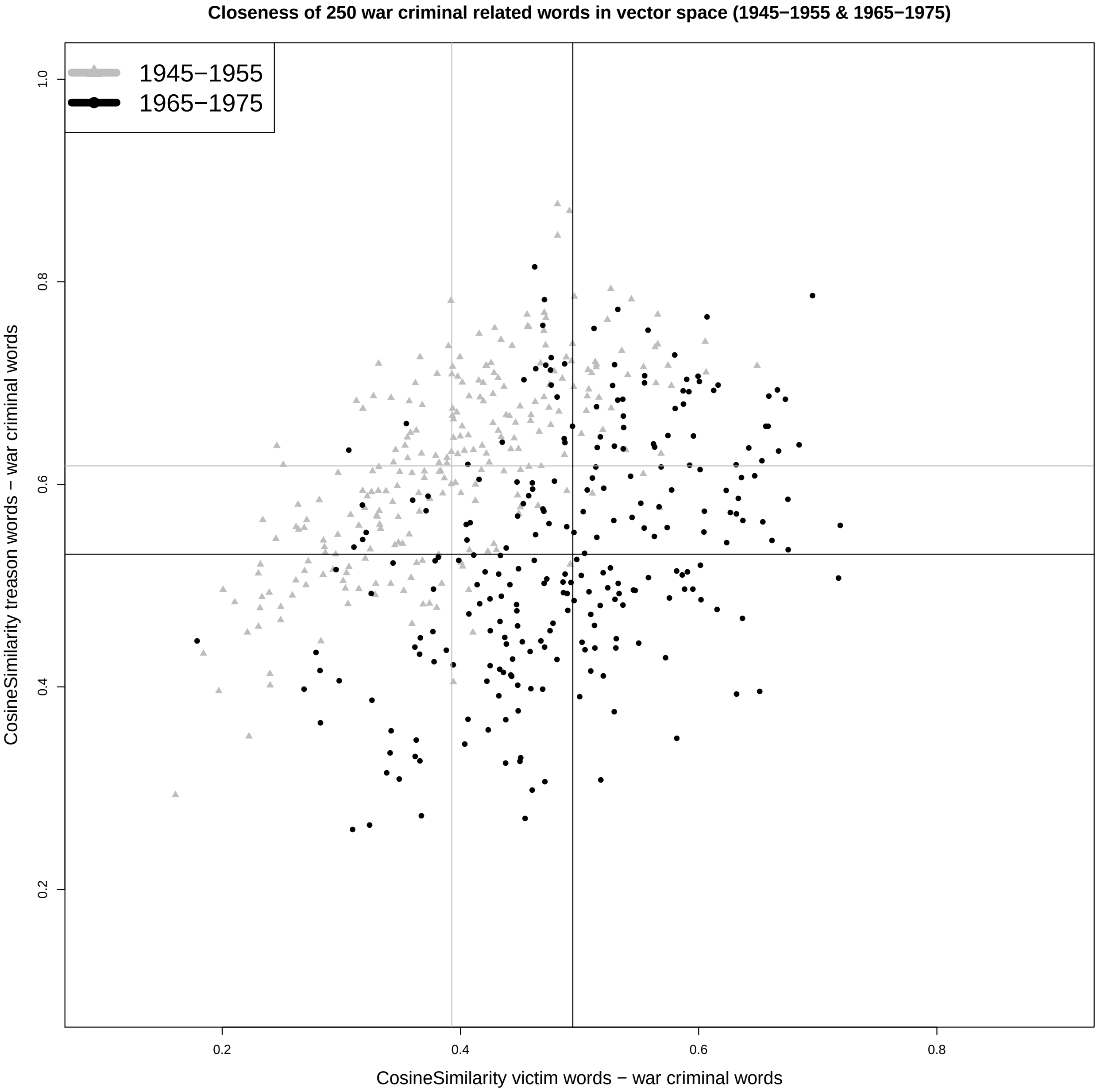


Figure 1: Top 250 war criminal related words 1945-1955 (grey) and 1965-1975 (black) plotted by their cosine similarity to victim (x) and traitor (y) words.

type of war criminals incarcerated in the Netherlands were reflected by a discursive shift, in which closeness to 'treason' declines and gave way to an increasing focus on victims in debating war criminals.

We have also encountered examples of pitfalls of an overly enthusiastic reliance on word embeddings as an analytical tool. Capital punishment was mentioned particularly frequently in the 1970s, but not because the possibility of executing the war criminals was seriously entertained. Distributional semantics are a powerful new tool for historians, but they do not not remove the need for hermeneutical awareness.

In this paper, the method is itself the main object of inquiry. We believe we have shown that it possible, feasible, and useful to develop and implement a coherent and widely applicable method for investigating historical change using WEMs.

7. Discussion

7.1. Method evaluation

For this paper, we have used two corpora of ten years to train our WEMs on. More interesting, from a research perspective, would be to find out how stable our results are when using smaller, overlapping windows of corpora over time, say with one-year steps. It is likely (but not certain) that using more fine-grained windows will reveal similar developments and shifts in language use over time. Repeating the analysis with more data points has the potential to gain more insights in the graduality and the pace of the

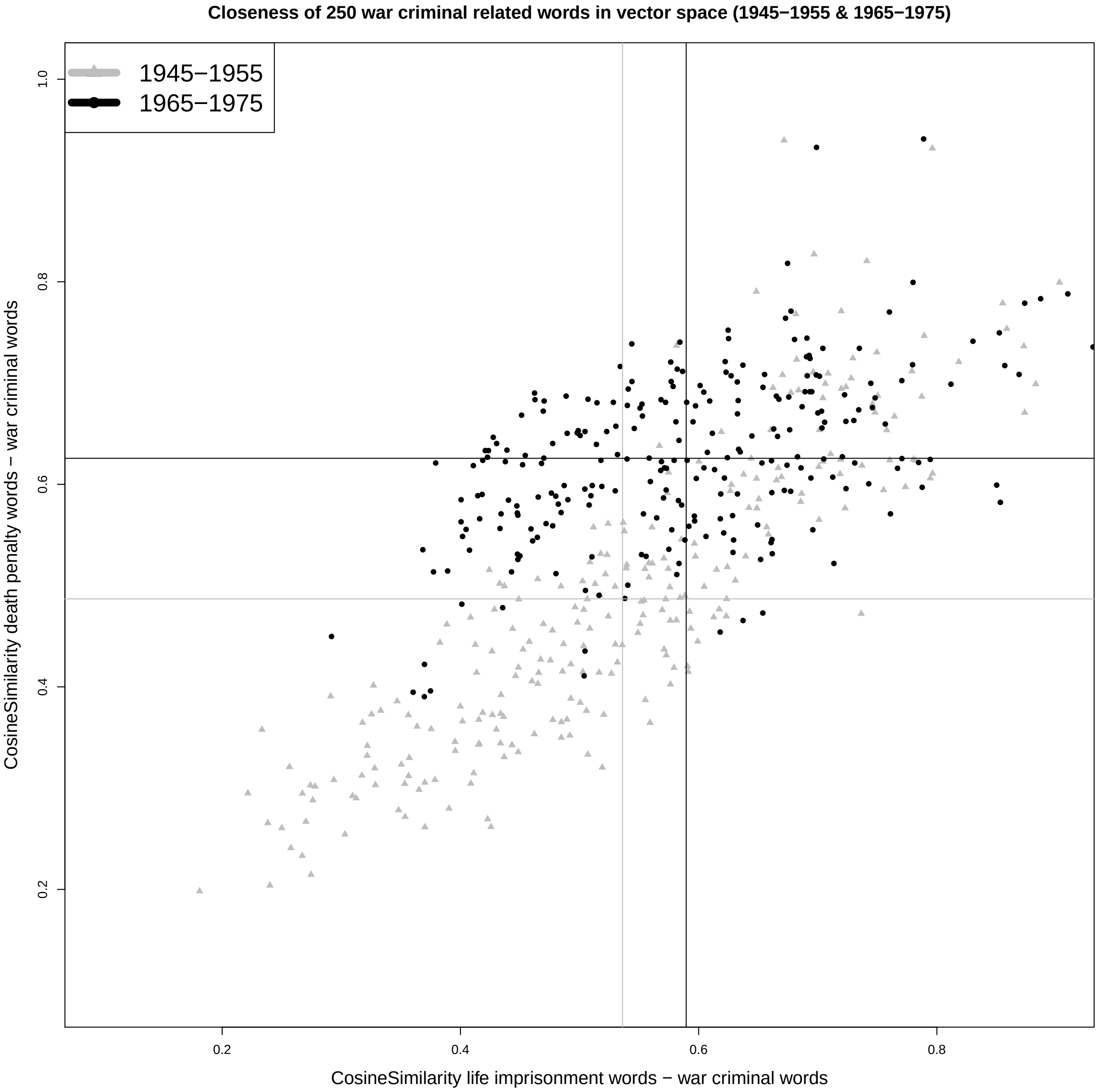


Figure 2: Top 250 war criminal related words 1945-1955 (grey) and 1965-1975 (black) plotted by their cosine similarity to life imprisonment (x) and death sentence words (y).

observed shifts in language used. That said, there is a potential trade-of between detail and precision given that the size of the corpora available to historians are mostly modest in size.

A second ambition is to look more seriously into the distribution of the cosine similarity scores, and the changes in these distributions over time. It will be interesting to measure, visualize, and statistically evaluate these distributions more closely, and to see whether they can be linked to, for example, unanimity and/or homogeneity in parliamentary discussions.

7.2. Historical evaluation

Another remaining ambition is to compare the parliamentary vocabularies used to discuss 'domestic' collaborators and foreign (usually German) war criminals. Furthermore, we also hope to position the war criminal debates in a broader context: how distinct are they from other war-related debates, and from other discussions about penal law or criminals in a more general sense?

Just as a closer investigation of different categories of perpetrators is viable and useful, different groups of war victims who were discussed in parliamentary debates also license further investigation. These may have included first and second generation victims of wartime violence and persecution, former forced labourers, holocaust survivors and the children of holocaust victims, etc. Given the emphasis on the protection of war victims mentioned above, we are interested to see if there have been changes in the groups emphasized in political speech about the topic.

8. Acknowledgements

We are grateful to the participants of our Text Mining workshop at the Luxembourg Centre for Contemporary and Digital History (C2DH) in Esch-sur-Alzette (June 2018), for their comments, input, and criticism.

9. Bibliographical References

Peter Bootsma and Peter van Griensven. 2003. ‘teleurstelling is mijn opperste emotie’: Vragen over emotie in de politiek aan a.a.m. van agt. In C. van Baalen et al., editor, *Jaarboek Parlementaire Geschiedenis, 2003. Emotie in de Politiek*. SDU Uitgevers, Den Haag.

Ido de Haan. 1997. *Na de ondergang: de herinnering aan de Jodenvervolging in Nederland 1945-1995*. Nederlandse cultuur in Europese context. Sdu Uitgevers.

Ralf Futselaar. 2015. *Gevangenissen in oorlogstijd 1940-1945*. Boom uitgevers Amsterdam.

Alexander Gelbukh. 2015. *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings*. Number dl. 1 in Lecture Notes in Computer Science. Springer International Publishing.

Alex Olieman, Kaspar Beelen, Milan van Lange, Jaap Kamps, and Maarten Marx. 2017. Good applications for crummy entity linkers? the case of corpus selection in digital humanities. *CoRR*, abs/1708.01162.

Hinke Piersma. 2005. *De Drie Van Breda: Duitse oorlogsmisdadigers in Nederlandse gevangenschap, 1945-1989*. Balans, Amsterdam, 1st edition.

Benjamin Schmidt. 2015. Vector space models for the digital humanities. *Ben’s Bookworm Blog*, oct.

Amit Singhal. 2001. Modern information retrieval: a brief overview. *BULLETIN OF THE IEEE COMPUTER SOCIETY TECHNICAL COMMITTEE ON DATA ENGINEERING*, 24:2001.

Hans Smits. 2008. *Strafrechthervormers en hemelbestormers: opkomst en teloorgang van de Coornhert-Liga*. Aksant, Amsterdam, 1st edition.

Ismee Tames. 2013. *Doorn in het vlees: foute Nederlanders in de jaren vijftig en zestig*. Balans, Amsterdam.

Chris van der Heijden. 2011. *Dat nooit meer: de nasleep van de Tweede Wereldoorlog in Nederland*. Atlas Contact, Uitgeverij.

Jolande Withuis. 2002. *Erkenning: van oorlogstrauma naar klaagcultuur*. De Bezige Bij, Amsterdam.

hr500k – A Reference Training Corpus of Croatian

Nikola Ljubešić,* Željko Agić,† Filip Klubička,‡ Vuk Batanović,§ Tomaž Erjavec*

*Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ijs.si, tomaz.erjavec@ijs.si

†Department of Computer Science, IT University of Copenhagen
Rued Langgaards Vej 7, 2300 Copenhagen S, Denmark
zeag@itu.dk

‡ADAPT Centre, School of Computing, Dublin Institute of Technology,
Kevin Street, Dublin, Ireland
filip.klubicka@adaptcentre.ie

§School of Electrical Engineering, University of Belgrade
Innovation Center, School of Electrical Engineering, University of Belgrade
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
vuk.batanovic@ic.etf.bg.ac.rs

Abstract

In this paper we present hr500k, a Croatian reference training corpus of 500 thousand tokens, segmented at document, sentence and word level, and annotated for morphosyntax, lemmas, dependency syntax, named entities, and semantic roles. We present each annotation layer via basic label statistics and describe the final encoding of the resource in CoNLL and TEI formats. We also give a description of the rather turbulent history of the resource and give insights into the topic and genre distribution in the corpus. Finally, we discuss further enrichments of the corpus with additional layers, which are already underway.

1. Introduction

Natural language processing techniques today are primarily based on supervised machine learning. Reference training corpora are therefore crucial for the development of NLP tools such as taggers, parsers, named entity recognizers etc.

In this paper we present hr500k – a reference corpus of Croatian which is currently annotated on the following levels: (1) token, sentence, and document segmentation, (2) morphosyntax, (3) lemmas, (4) dependency syntax, (5) semantic roles and (6) named entities. The corpus presents a significant extension of previous training corpora developed for Croatian, namely the SETimes.HR and the SETimes.HR+ corpora, and allows Croatian's basic language technologies to finally catch up to other well-equipped Slavic languages such as Slovene (Krek et al., 2018), Czech (Hajič et al., 2012) or Polish (Broda et al., 2012).

2. Description of the corpus

The hr500k corpus consists of 900 documents segmented into around 25 thousand sentences, making the average document length around 28 sentences or 563 tokens, while the average sentence length is around 20 tokens. A statistical overview of the corpus is given in Table 1. Each document is preceded by a tag indicating its name and the URL of the source, if available. Each tokenized sentence is preceded by a tag stating its original, untokenized text. The form of all such tags is compliant with the Universal Dependencies v2 specifications.[1]

Item	Count
Documents	900
Sentences	24 794
Tokens	506 457
Types	73 548
Lemmas	34 329
MSDs	768

Table 1: A statistical overview of the hr500k corpus

Regarding the genres covered in the corpus, around 55% of the content are news articles, blogs covering 20% of the content, forums 15% and 10% being covered by other genres. For the topical distribution, around 50% of content covers the general topic, music and medicine each covering around 10% of the content, while business, tech, lifestyle and education cover around 5% of the content each. A more detailed description of the genre and topic distributions, from the perspective of extending the corpus through time, can be found in Section 4..

2.1. Morphosyntax and lemmas

The entire hr500k corpus is annotated with morphosyntactic tags and lemmas for each token. Morphosyntax is encoded according to the MULTEXT-East V5 guidelines,[2] which specify 13 part-of-speech categories, with numer-

[1] http://universaldependencies.org/v2/

[2] http://nl.ijs.si/ME/V5/msd/html/

MTEv5 gloss	POS tag	Count	Percentage
Nouns	N	135 822	26.82%
Verbs	V	80 499	15.89%
Punctuation	Z	62 116	12.26%
Adjectives	A	50 982	10.07%
Adpositions	S	45 145	8.91%
Pronouns	P	40 591	8.01%
Conjunctions	C	36 685	7.24%
Adverbs	R	26 051	5.14%
Numerals	M	12 744	2.52%
Particles	Q	8 787	1.74%
Residuals	X	5 051	1.00%
Abbreviations	Y	1 666	0.33%
Interjections	I	318	0.06%

Table 2: MTEv5 part-of-speech tag distribution in the hr500k corpus

ous morphosyntactic attributes particular to each category. A list of these categories, alongside their frequencies in the hr500k corpus, is given in Table 2. In addition to the MTEv5 specification, we also provide POS tags in accordance with the Universal Dependencies v2 standard, which describes 17 part-of-speech categories. The frequency distribution of UD POS tags in the hr500k corpus is shown in Table 3.

MTE morphosyntactic tags can, for the most part, be automatically mapped into UD POS tags, and we provide the mapping table and code on the hr500k Github repository.[3] The only exception to this are abbreviations (MULTEXT-East tag Y), which have to be converted manually.

Given that the corpus was extended through time (more on the history of the corpus will be reported in Section 4.), there were multiple annotation rounds on the morphosyntactic and lemma annotation layers. To secure a consistent annotation of these phenomena, we have recently performed a series of global annotation consolidation procedures by which we expect for the annotation consistency to be high. However, up to this point, we have not measured this level of consistency.

2.2. Dependency syntax

The first two fifths of the hr500k, i.e. the first 197 028 tokens of the corpus, have been annotated with regard to dependency syntax. These annotations include both the older syntactic tags presented by Agić and Ljubešić (2014), as well as the newer Universal Dependency v2 syntactic relations.[4] Table 4 shows the distribution of the UDv2 syntactic tags in our corpus.

This annotation layer was annotated by a single annotator and there were no annotation consolidation procedures performed up to this point. One of our future goals is to harmonize the UD annotations between the Slovene (Krek et al., 2018), Croatian and Serbian (Samardžić et al., 2017) training corpora.

UD POS gloss	UD POS tag	Count	Percentage
Nouns	NOUN	113 674	22.44%
Punctuation	PUNCT	61 914	12.22%
Adjectives	ADJ	56 071	11.07%
Verbs	VERB	49 089	9.69%
Adpositions	ADP	45 144	8.91%
Auxiliary	AUX	31 413	6.20%
Adverbs	ADV	26 144	5.16%
Proper nouns	PROPN	23 160	4.57%
Coord. conj.	CCONJ	22 175	4.38%
Determiners	DET	21 012	4.15%
Pronouns	PRON	19 579	3.86%
Subord. conj.	SCONJ	14 510	2.86%
Particles	PART	8 941	1.76%
Numerals	NUM	7 813	1.54%
Other	X	5 262	1.04%
Interjections	INTJ	322	0.06%
Symbols	SYM	234	0.05%

Table 3: UD part-of-speech tag distribution in the hr500k corpus

2.3. Semantic roles

Semantic roles are currently annotated in the oldest part of the corpus, namely the documents coming from the original SETimes.HR corpus, without the original testing portion of the corpus. This part of the corpus contains 163 documents, 3 757 sentences and 83 630 tokens. On average there are 5 SRL labels applied to each sentence.

The SRL formalism was developed inside a bilateral Slovene-Croatian project, lead by Simon Krek on the Slovene side and Kristina Štrkalj Despot on the Croatian side. The formalism presents for the most part a simplification of the Prague Dependency Treebank formalism. Table 5 shows the distribution of the SRL tags in our corpus.

The Croatian SRL annotations were applied by a single annotator, with complex examples being discussed together by the Croatian and Slovene project partners.

2.4. Named entities

Named entity annotations cover the entire hr500k and are encoded in the IOB2 format. 36 735 tokens, or 7.25% of the total, are marked as belonging to a named entity, of which there are 23 186, which means there are around 26 named entities per document, or almost one per sentence, on average. Five NE types are considered – the standard categories for people (PER), locations (LOC), organizations (ORG), and miscellaneous entities (MISC) are augmented with a person derivative category (DERIV-PER), intended for marking personal (possessive) adjectives, enabling better information extraction and anonymization of personal data. The annotation guidelines applied are those that were developed while annotating the Slovene ssj500k and Janes-Tag datasets.[5]

The distribution of entities in hr500k between these categories is given in Table 6. The distribution of tokens belonging to a named entity is given in Table 7.

[3] http://github.com/nljubesi/hr500k/

[4] The meaning of the tags is explained here: http://universaldependencies.org/u/dep/index.html

[5] http://nl.ijs.si/janes/wp-content/uploads/2017/09/SlovenianNER-eng-v1.1.pdf

UD syntactic tag	Count	Percentage
punct	23 894	12.13%
case	18 936	9.61%
nmod	18 289	9.28%
amod	18 229	9.25%
nsubj	13 959	7.08%
obl	12 226	6.20%
conj	9 414	4.78%
root	8 889	4.51%
obj	8 719	4.42%
aux	8 532	4.33%
advmod	8 103	4.11%
cc	7 528	3.82%
mark	3 941	2.00%
acl	3 817	1.94%
det	3 518	1.78%
cop	3 478	1.76%
xcomp	2 949	1.50%
appos	2 894	1.47%
nummod	2 887	1.46%
flat	2 526	1.28%
compound	2 274	1.15%
parataxis	2 268	1.15%
expl	2 173	1.10%
discourse	2 040	1.04%
ccomp	1 766	0.90%
advcl	1 753	0.89%
fixed	707	0.36%
iobj	689	0.35%
csubj	359	0.18%
orphan	152	0.08%
goeswith	55	0.03%
list	24	0.01%
vocative	23	0.01%
dep	9	0.01% <
dislocated	8	0.01% <

Table 4: UD syntactic relation distribution in the hr500k corpus

The named entity annotation layer was applied by two annotators, with collisions in the annotations being resolved by a super-annotator. Regardless of the double-annotation procedure, we plan to perform a global lexical label consolidation procedure in the near future.

3. Corpus encoding and publishing

The working version of the corpus was encoded in a modified version of the tabular CoNLL-X format (Buchholz and Marsi, 2006), consisting of the following columns:

1. ID: sentence-local word index
2. FORM: token, i.e. word form or punctuation symbol
3. LEMMA: lemma of word form
4. POS: part-of-speech according to the MULTEXT-East specifications
5. MSD: morphosyntactic description according to the MULTEXT-East specifications
6. MSDFEAT: morphosyntactic features according to the MULTEXT-East specifications
7. SETDEPREL: dependency relation (head, label) according to the SETimes formalism
8. UDDEPREL: dependency relation (head, label) according to the UDv2 formalism

SRL gloss	SRL tag	Count	Percentage
Patient	PAT	4 860	26.10%
Agent	ACT	4 731	25.40%
Result	RESLT	2 860	15.36%
Time	TIME	1 344	7.22%
Recipient	REC	603	3.24%
Modality	MODAL	586	3.15%
Location	LOC	525	2.82%
Manner	MANN	472	2.53%
Duration	DUR	351	1.88%
Origin	ORIG	268	1.44%
Cause	CAUSE	242	1.30%
Aim	AIM	224	1.20%
Regard	REG	222	1.19%
Goal	GOAL	202	1.08%
Event	EVENT	170	0.91%
Means	MEANS	169	0.91%
Quantity	QUANT	158	0.85%
MW predicate	MWPRED	134	0.72%
Accompaniment	ACMP	106	0.57%
Condition	COND	87	0.47%
Contradiction	CONTR	82	0.44%
Frequency	FREQ	81	0.43%
Part of phraseme	PHRAS	74	0.40%
Source	SOURCE	50	0.27%
Restriction	RESTR	22	0.12%
Total		18 623	100%

Table 5: Distribution of SRL tags in the hr500k corpus

Named entity type	Count	Percentage
Person	6 802	29.34%
Person derivative	317	1.37%
Location	6 214	26.80%
Organization	6 354	27.40%
Miscellaneous	3 499	15.09%
Total	23 186	100%

Table 6: Distribution of named entities in the hr500k corpus

Named entity type	Token count	Percentage
Person	10 241	2.02%
Person derivative	319	0.06%
Location	7 445	1.47%
Organization	11 216	2.21%
Miscellaneous	7 514	1.48%
Total	36 735	7.25%

Table 7: Distribution of named entity tokens with regard to the whole hr500k corpus

9. UPOS+FEATS: universal part-of-speech tag with features from the universal feature inventory

10. UDSPEC: UDv2 language-specific features

11. NER: named entity annotations encoded through IOB2

12. SRLHEAD: heads of semantic roles, order of occurrence in a sentence defines in which columns semantic roles to specific heads are encoded (columns 13-23)

13. SRL: semantic roles, encoded to column 23

The CoNLL-type format was converted to TEI, i.e., to a schema following the Guidelines for Electronic Text Encoding and Interchange (TEI Consortium, 2017) to ensure (meta-)data persistence. Apart from the automatic conversion of the text and its annotations, this also involved writing the `teiHeader` element, which gives the meta-data of the corpus, containing its name, authors, license, source description, annotation vocabulary, tag usage, revision history etc.

As illustrated in Figure 1, each sentence in the TEI encoding is assigned a unique ID, as are all the tokens (words, `w` and punctuation symbols, `pc` in the sentence; white space in the sentence is also marked-up with `c`).

The lemma of the words is given in the `@lemma` attribute, while all tokens are given their MULTEXT-East MSD in the `@ana` attribute. The UD parts of speech and features are given in the `@msd` attribute, which is an attribute newly introduced into the TEI. Note that the double pipe symbol is used to separate the universal features from the (Croatian) language-specific ones. The reason why the MULTEXT-East MSDs are not given in the `@msd` attribute, as might be expected, is that while `@msd` can contain any string, the `@ana` is defined as a pointer, which MULTEXT-East MSDs can be, but UD features cannot; we explain below in more detail the functioning of TEI pointers for linguistic labels as used in the hr500k corpus.

Named entities are also encoded in-line, simply using the standard TEI `name` element, with the `@type` giving the type of the name.

The final three layers of annotation, namely the original syntactic dependencies, the UD dependencies and the semantic role labels are all encoded in stand-off annotation, using the `linkGrp` (link group) element, which specifies its type (so, annotation layer), the ordering of the arguments of the links, and then contains the links themselves; each of these gives the link label and pointers to the IDs of the link head and argument. It should be noted that in cases where a syntactic dependency has the (virtual) root as its head, the references to the sentence ID is given (so, in the example above that would be `#train-s2`).

As mentioned, the `@ana` attribute is a pointer, which usually contains a local reference to an ID (e.g. `#train-s2.1`) or a fully qualified URI. TEI has another option for its pointers, namely using a prefix before the ID and separated from it by a colon (e.g. `mte:Npnsn`). Such pointers are then resolved using the `prefixDef` element in the TEI header, which defines the prefixing schema used, showing how abbreviated URIs using the scheme may be expanded into full URIs. In the case of the hr500k corpus all the prefixes are simply expanded to local references, which are given in the TEI header, except for the MULTEXT-East MSD, which are defined in the `back` element of the TEI document. There, each MSD is defined as a feature-structure giving the decomposition of the MSD into its features. It is thus a simple matter, using just the TEI encoded corpus, to move from `mte:Mdo` to `Category = Numeral, Form = digit, Type = ordinal`.

The TEI encoded corpus, which is to be taken as the canonical version of the hr500k corpus, was then automatically converted to the so called vertical format which is used by CQP-based concordancers, in particular (no)Sketch Engine (Rychlý, 2007). The vertical format is able do encode hierarchical structures (e.g. sentences and names), and token annotations (e.g. lemmas and MSDs), but not links between tokens (e.g. dependencies and semantic role labels). To nevertheless preserve as much of this information as possible, the dependencies are annotated next to tokens, so that the argument token is annotated with the dependency label and head lemma.

Finally, the TEI, vertical and CoNLL encoded corpus were deposited to the CLARIN.SI repository,[6] where the corpus is available under the CC BY-SA license. The corpus is also available for exploration under the CLARIN.SI noSketch Engine and KonText concordancers; the links are given on the CLARIN.SI repository landing page.

4. History of the corpus

4.1. The SETimes.HR corpus

The prolific five-year period between the seminal shared tasks in dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007) and the emergence of the first cross-linguistically uniform morphological and syntactic annotation guidelines (De Marneffe and Manning, 2008; Petrov et al., 2011; McDonald et al., 2013) has thoroughly changed the landscape of multilingual NLP.

Back then, the field's positive momentum was not entirely mirrored by the developments for Croatian. In 2007 the Croatian dependency treebank (HOBS), deemed a central resource for training basic NLP models, was but a 50-sentence prototype (Tadić, 2007). By late 2012 the resource grew to around 75% of its full size of 4.6k sentences (Berović et al., 2012). Yet, it was not publicly available. At that point in time there were *no* freely available Croatian language resources to train NLP models whatsoever.

The consequences were dire. To illustrate, in 2012 one could not even tag Croatian for POS, let alone perform any syntactic parsing, while the rest of the field was involved in pursuing human-level accuracies in these basic tasks. The SETimes.HR corpus was thus developed to address the urgent need for a free-culture resource to build and evaluate basic NLP for Croatian.

The original instantiations of SETimes.HR are the experiments in POS tagging and dependency parsing by Agić et al. (2013a; 2013b), which also included a cross-lingual application to Serbian as a very closely related yet truly low-resource language. Agić and Ljubešić (2014) provide

[6] published at `http://hdl.handle.net/11356/1183`

```
<s xml:id="train-s2">
  <name type="loc">
    <w xml:id="train-s2.1" lemma="Kosovo" ana="mte:Npnsn"
       msd="UposTag=PROPN|Case=Nom|Gender=Neut|Number=Sing">Kosovo</w>
  </name>
  <c> </c>
  ...
  <w xml:id="train-s2.9" lemma="pritužba" ana="mte:Ncfpg"
     msd="UposTag=NOUN|Case=Gen|Gender=Fem|Number=Plur||SpaceAfter=No">pritužbi</w>
  <pc xml:id="train-s2.10" ana="mte:Z" msd="UposTag=PUNCT">.</pc>
  <linkGrp targFunc="head argument" type="SE-SYN">
    <link ana="se-syn:Sb" target="#train-s2.3 #train-s2.1"/>
    <link ana="se-syn:Adv" target="#train-s2.3 #train-s2.2"/>
    ...
  </linkGrp>
  <linkGrp targFunc="head argument" type="UD-SYN">
    <link ana="ud-syn:nsubj" target="#train-s2.3 #train-s2.1"/>
    <link ana="ud-syn:advmod" target="#train-s2.3 #train-s2.2"/>
    ...
  </linkGrp>
  <linkGrp targFunc="head argument" type="SRL">
    <link ana="srl:ACT" target="#train-s2.3 #train-s2.1"/>
    <link ana="srl:MANN" target="#train-s2.3 #train-s2.2"/>
    ...
  </linkGrp>
</s>
```

Figure 1: TEI encoding of a corpus sentence

a thorough documentation of the new corpus and introduce an annotation layer for named entities following Tjong Kim Sang and De Meulder (2003). The release was open to the public for all uses.

The corpus contained around 4,000 sentences (84k word forms) annotated using a modified MULTEXT-East V5 guideline for POS and morphology, while the syntactic dependencies followed a novel Prague-motivated lean tagset by Agić and Merkler (2013). It was introduced to the first version of Universal Dependencies (Agić et al., 2015; Nivre et al., 2016) together with compliant annotations that were applied manually.

As a grassroots effort, the SETimes.HR corpus was not devoid of flaws. For one, its training section consisted entirely of newspaper data from a single source,[7] while its test data were multi-domain. However, it played a crucial role in democratizing Croatian NLP resources, eventually drawing out even the dormant HOBS after years of public unavailability (Agić et al., 2014).

This section of the corpus is identifiable in the final hr500k corpus by document IDs starting with `set.hr` and consists of 164 documents (163 training and one testing document).

4.2. The SETimes.HR+ corpus

Further extensions of the SETimes.HR corpus were performed in 2014, in two main phases. The first phase consisted of adding texts collected in 2012 for a named entity recognition task. The second phase focused on selecting and annotating texts in a crowdsourcing framework inside a master course.

The first extension of the SETimes.HR corpus consisted of texts from a named entity recognition training corpus for Croatian that was built in 2012 during a student project and contained initially 59,212 tokens (Ljubešić et al., 2012). The data came from four different web domains belonging to the genres of general news, ICT news and business news. This section of the corpus is identifiable in the final hr500k corpus by document IDs starting with `news.hr`, consisting overall of 83 documents.

The second extension of the SETimes.HR corpus was performed as part of a master course. No specific topic domain was chosen, but rather a random sample of sentences from the general web which, through crowdsourcing efforts, were deemed as being of an acceptable linguistic standard. This dataset of 50,322 tokens was then automatically MSD-tagged, followed by employing crowdsourcing and a small team of experts to correct the annotations of tokens that were tagged differently by a tagger ensemble (Klubička and Ljubešić, 2014). This section of the corpus is organized in the final hr500k corpus into a single document with the document ID `web.hr`.

Both these corpus sections were later merged with the original SETimes.HR corpus into one corpus, internally referred to as SETimes.HR+, with approximately 190 thousand tokens in size. This new corpus was manually inspected for possible errors and inconsistencies. As for genre and register, the content of this corpus belonged mainly to news (>85%), and a little bit of the general web, which varied greatly by genre and topic, including the odd forum discussion or blog post, but mostly consisting of re-

[7] The now defunct Southeast European Times online news portal `setimes.com`, also basis of the SETimes parallel corpus.

	articles	blogs	forums	other
320k extension	40%	30%	20%	10%
hr500k	57.63%	20.6%	14.64%	7.13%

Table 8: Token percentage per web genre in the 320k extension added to the hr500k corpus and the final hr500k corpus

ports on politics, sports, religion, in addition to news and other informative articles.

The SETimes.HR+ corpus as described in this section currently also serves as the Croatian part of the current Universal Dependencies treebanks release (v2.2). It was manually annotated for UD-style syntactic dependencies, and its POS tags and morphological features were semi-automatically converted to UD.

4.3. hr500k

In 2015 we raised the bar to 500 thousand tokens (Ljubešić et al., 2016), motivated by results on morphosyntactic annotation of Slovene (Ljubešić and Erjavec, 2016) which showed that corpus supervision has a much higher impact on tagging accuracy than lexicon supervision. Thus, for the final phase of corpus assembly we manually selected 320k tokens worth of suitable documents from the hrWaC web corpus (Ljubešić and Klubička, 2014). The documents were automatically morphosyntactically annotated with a tagger learned on the 190k-sized SETimes.HR+ corpus, which was followed by having experts perform manual correction of the automatic annotations.

However, this time around we were somewhat spoiled for choice with regards to the content to be included in the corpus, as the hrWaC corpus boasts 1.3 billion tokens. This allowed us the luxury to include more varied content than the previous iterations had. Thus, our aim was to gather a representative sample of the Croatian language; one that expands beyond the confines of a particular genre, topic or register, and includes many different examples of linguistic expression that can be found on the web. With accordance to that, we divided the additional 320k token sample into 4 sections according to web genre, in the ratios shown in the first row of Table 8, the second row showing the distribution of web genres in the final hr500k resource.

This way, we covered registers used in different kinds of genres – articles, blogs, forums, reviews and advertisements – while at the same time covering a wide range of topics that were inadequately or not at all covered in the SETimes.HR+ corpus, which mainly consisted of general news articles. The web domains that we included cover topics ranging from medicine, education and technology, through music, sports and religion, all the way to listings and adverts, literature and political activism. Where possible, we also made the effort to include any user comments posted on their corresponding articles and blogs, so that, coupled with forum discussions, the corpus would also include a sample of the language used in direct communication among Internet users. Such meticulous selection results in considerable variety among documents, but given that documents were selected exclusively from a list of the top 200 most frequent domains in the hrWaC corpus, this varied sample is actually quite representative of the Croatian web.

topic	token ratio	topic	token ratio
general	35.01%	business	4.41%
music	13.55%	listings	3.88%
medicine	12.26%	religion	3.81%
tech	7.93%	sports	3.59%
lifestyle	7.38%	culture	2.36%
education	5.80%		

Table 9: Topic domain distribution in the 320k extension

topic	token ratio	topic	token ratio
general	51.89%	education	3.61%
music	8.43%	religion	2.87%
medicine	7.63%	sports	2.74%
business	6.93%	listings	2.42%
tech	6.92%	culture	1.97%
lifestyle	4.59%		

Table 10: Topic domain distribution in the hr500k corpus

An approximation of the distribution of web genres in the final hr500k corpus created by merging all the hitherto described corpora is presented in the second row of Table 8. An overview of the topic domains that enriched the corpus in the second phase of construction is presented in Table 9 and is based on the general topic of the web domains the sentences come from, while an approximation of topic domain distribution in the final 500k corpus is presented in Table 10. Compared to the approximate >85% of general news articles that comprise the initial SETimes.HR+ corpus, this is a vast improvement in terms of data diversity.

5. Conclusion

In this paper we presented the manually annotated reference corpus of Croatian, which is currently the largest and richest training dataset for Croatian and is made freely (CC BY-SA) available for download[8] and for on-line exploration.[9]

Future plans are primarily directed at (1) further consolidation of the presented annotation layers and (2) extension to new annotation layers, two being planned for the near future in form of layers of verbal multiword expression annotations.

We will also define training, development and test portions of the corpus for each task and benchmark the available language tools for the available tasks, thereby further fostering development of language technologies for Croatian and other languages.

6. Acknowledgements

The presented resource was initially developed through pure enthusiasm of a network of researchers, with recent extensions and consolidations being funded by the Regional Linguistic Data Initiative (ReLDI, Swiss National

[8] http://hdl.handle.net/11356/1183

[9] https://www.clarin.si/kontext/first_form?corpname=hr500k

Science Foundation grant IZ74Z0 160501) and the Slovenian Research Infrastructure CLARIN.SI. Filip Klubička is also supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. Vuk Batanović is also partially supported by the III44009 research grant of the Ministry of Education, Science and Technological Development of the Republic of Serbia.

7. References

Željko Agić and Nikola Ljubešić. 2014. The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Željko Agić and Danijela Merkler. 2013. Three syntactic formalisms for data-driven dependency parsing of croatian. In *International Conference on Text, Speech and Dialogue*, pages 560–567. Springer.

Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013a. Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proceedings of the Fourth Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57, Sofia, Bulgaria, August. Association for Computational Linguistics.

Željko Agić, Danijela Merkler, and Daša Berović. 2013b. Parsing Croatian and Serbian by using Croatian dependency treebanks. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2013)*.

Željko Agić, Daša Berović, Danijela Merkler, and Marko Tadić. 2014. Croatian dependency treebank 2.0: New annotation guidelines for improved parsing. In *Ninth International Conference on Language Resources and Evaluation (LREC 2014)*.

Željko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, et al. 2015. Universal dependencies 1.1. *LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague*, 3.

Daša Berović, Željko Agić, and Marko Tadić. 2012. Croatian dependency treebank: Recent development and initial experiments. In *Seventh International Conference on Language Resources and Evaluation (LREC 2012)*.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. Kpwr: Towards a free corpus of polish. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12*.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 149–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8. Association for Computational Linguistics.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. *Prague Czech-English Dependency Treebank 2.0*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. `http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4`.

Filip Klubička and Nikola Ljubešić. 2014. Using crowdsourcing in building a morphosyntactically annotated and lemmatized silver standard corpus of croatian. In Tomaž Erjavec and Jerneja Žganec Gros, editors, *Language technologies: Proceedings of the 17th International Multiconference Information Society IS2014*, Ljubljana, Slovenia.

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. 2018. *Training corpus ssj500k 2.1*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1181`.

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, Sweden. Association for Computational Linguistics.

Nikola Ljubešić, Marija Stupar, and Tereza Jurić. 2012. Building named entity recognition models for croatian and slovene. In Tomaž Erjavec and Jerneja Žganec Gros, editors, *Proceedings of the Eighth LANGUAGE TECHNOLOGIES Conference*, Ljubljana, Slovenia.

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*.

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'16)*. European Language Resources Association (ELRA).

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

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.

Tanja Samardžić, Mirjana Starović, Željko Agić, and Nikola Ljubešić. 2017. Universal dependencies for serbian in comparison with croatian and other slavic languages. In *The 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*.

Marko Tadić. 2007. Building the croatian dependency treebank: the initial stages. *Suvremena lingvistika*, 63(1):85–92.

TEI Consortium. 2017. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

The Parlameter corpus of contemporary Slovene parliamentary proceedings

Nikola Ljubešić,* Darja Fišer,†* Tomaž Erjavec,* Filip Dobranić‡

*Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
nikola.ljubesic@ijs.si, tomaz.erjavec@ijs.si

†Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si

‡Danes je nov dan
Parmova ulica 20, 1000 Ljubljana
filip@danesjenovdan.si

Abstract

The paper presents the Parlameter corpus of contemporary Slovene parliamentary proceedings, which currently covers the VIIth mandate of the Slovene Parliament (2014-2018). The Parlameter corpus offers rich speaker metadata (gender, age, education, party affiliation) which boost research in several digital humanities and social sciences disciplines. We analyze the linguistic production paired with the metadata from the perspective of communication and political studies. The corpus architecture allows for regular extensions of the corpus with additional Slovene data, as well as data from other parliaments, starting with Croatian and Bosnian.

1. Introduction

Parliamentary discourse is motivated by a wide range of communicative goals, from position-claiming, persuasion and negotiation to agenda-setting and opinion-building along ideological or party lines. It is characterized by role-based commitments and confrontation and the awareness of a multi-layered audience (Ilie, 2017). The unique content, structure and language of records of parliamentary debates are all factors make them an important object of study in a wide range disciplines in digital humanities and social sciences, such as political science (Van Dijk, 2010), sociology (Cheng, 2015), history (Pančur and Šorn, 2016), discourse analysis (Hirst et al., 2014), sociolinguistics (Rheault et al., 2016), and multilinguality (Bayley, 2004).

Despite the fact that parliamentary discourse has become an increasingly important research topic in various fields of digital humanities and social sciences in the past 50 years (Chester and Bowring, 1962; Franklin and Norton, 1993), it has only recently started to acquire a truly interdisciplinary scope (Bayley, 2004). Recent developments enable cross-fertilization of linguistic studies with other disciplines and in-depth exploration of institutional uses of language, interpersonal behavior patterns, interplay between language-shaped facts, and reality-prompted language ritualization and change (Ihalainen et al., 2016).

With an increasingly decisive role of parliaments and their rapidly changing relations with the public, mass media, executive branch and international organizations, further empirical research and development of integrative analytical tools is necessary in order to achieve a better understanding of parliamentary discourse as well as its wider societal impact, in particular with studies that represent diverse parts of society (women, minorities, marginalized groups) and cross-cultural studies (Hughes et al., 2013).

2. Parliamentary corpora

The most distinguishing characteristic of records of parliamentary debates is that they are essentially transcriptions of spoken language produced in controlled and regulated circumstances. For this reason, they are rich in invaluable (sociodemographic) meta-data. They are also easily available under various Freedom of Information Acts set in place to enable informed participation by the public and to improve effective functioning of democratic systems, making the datasets even more valuable for researchers with heterogeneous backgrounds.

This has motivated a number of national as well as international initiatives (for an overview, see Fišer and Lenardic (2018)) to compile, process and analyze parliamentary corpora. They are available for most countries within the CLARIN ERIC research infrastructure for language resources and technology, with the UK's Hansard Corpus being the largest (1.6 billion tokens) and spanning the longest time period (1803-2005) while corpora from other countries are significantly smaller (most comprise between 10 and 100 million tokens) and cover significantly shorter periods (mostly from the 1970s onwards).

The Slovene parliamentary corpus SlovParl 2.0 (Pančur, 2016) contains minutes of the Assembly of the Republic of Slovenia for the legislative period 1990-1992 when Slovenia became an independent country. The corpus comprises over 200 sessions, almost 60,000 speeches and 11 million words. It contains extensive meta-data about the speakers, a typology of sessions and structural and editorial annotations and is uniformly encoded to the Text Encoding Initiative (TEI) Guidelines, a de-facto standard for encoding and annotating textual data in Digital Humanities. It is available under the CC-BY licence in the CLARIN.SI repository of language resources and via the CLARIN.SI concordancers

(Pančur et al., 2017). SlovParl is thus an exemplary corpus but contains material from a quite limited, and not very recent, time period. This makes the corpus of limited use for the rich body of research on recent parliamentary activities.

Contemporary Slovenian parliamentary debates are monitored by the analytical tool Parlameter[1] which makes use of linguistic as well as non-linguistic data, such as MPs' attendance and voting results. While this is a wonderful tool for journalists and citizen scientists and gives valuable insight into contemporary parliamentary data, the material is confined to the functionalities of the tool and as such cannot be freely manipulated by scholars according to their specific research needs.

The goal of the research presented in this paper was to convert the Parlameter database into a corpus and to enrich the linguistic data with the session and speaker metadata. Section 3. gives the basic information on the corpus structure and size, Section 4. presents the analysis of the corpus according to the speaker metadata (gender, education, age, party affiliation), and Section 5. gives some conclusions and directions for further research. While the focus of the paper is the parliamentary language material which we analyze with standard corpus and natural language processing approaches, the aim of the analysis is to inform media and political studies.

3. Corpus compilation

The corpus covers the full VIIth mandate of the Slovene Parliament, ranging from August 1st 2014 to March 19th 2018. Currently the central entities in the corpus are the sessions and speeches given by the members of parliament and other speakers as this is the most interesting content to researchers in the areas of linguistics, communication and political science.

The data model currently used for encoding the corpus is presented in Figure 1. Parliamentary sessions are equipped with the mandate they belong to and the name and date of the session. Sessions are further broken down into individual speeches which, in addition to the content of the contribution, is annotated with speaker name and other speaker information (date of birth, gender, education and education level, party affiliation). The rich speaker data are available for the members of parliament and members of the government but not for all other speakers in the parliament (e.g., field experts, representatives of governmental agencies, non-governmental organizations or civil initiatives). This is why the analyses in Section 4. are performed based on the instances for which the metadata is available in the corpus.

Some basic statistics regarding the corpus are given in Table 1. The transcripts were also processed with the standard linguistic annotation pipeline for Slovene consisting of reldi-tokeniser, which segments the text string into tokens and sentences, and reldi-tagger, which adds morphosyntactic descriptions (MSDs) and lemmas to the word tokens (Ljubešić et al., 2016).

```
[
  {
  mandate,
  session_name,
  date,
  speeches:
    [
      {
      id
      content,
      speaker:
        {
        id
        birth_date
        education
        education_level
        gender
        name
        party
        }
      },
      ...
    ]
  },
  ...
]
```

Figure 1: The current JSON data model of the corpus

Level	Count
Sessions	362
Days	514
Speeches	218,398
Speakers	1,984
Sentences	3,070,314
Words	61,039,385
Tokens	70,874,201

Table 1: Basic statistics of the corpus

A basic analysis of the morphosyntactic annotations of the corpus in form of the most significant differences in frequency of MSD tags between the KRES balanced corpus of Slovene and the Parlameter corpus are given in Table 2.

The results show that the parliamentary speeches, as expected, contain more present tense word forms (`Vm.r[1-3][sp]`), especially in the fist person singular or plural (`Vm.r1[sp]`), demonstrative pronouns (`Pd-.*`), the first person singular personal pronoun (`Pp1-sn`), the first person auxiliary verbs (`Va-.1.-n`) and adverbs (`Rgp`) compared to general Slovene.

On the other hand, the parliamentary proceedings contain significantly fewer proper names (`Np.*`), numbers (`Md.`), verb participle forms (`Vm.p-.*`), personal pronoun dative cases (`Pp3..d.*`) and general and possessive adjective forms (`A.p.*`) than general Slovene.

4. Corpus analysis

This section persents a brief analysis of the corpus content given four main variables: gender, education, age and political affiliation of the speakers. In each session we disregard the speeches given by the most frequently occurring speaker because it is evident that this speaker was in charge of leading the session and, as a consequence, their content

[1] https://parlameter.si

KRES	Parlameter
Npmsn	Vmpr1s
Mdc	Pd-nsn
Ncmsn	Vmpr1p
Vmep-sm	Pp1-sn
Npfsn	Pd-nsa
Vmep-sf	Va-r1p-n
Pp3msd–y	Vmpr1p-n
Va-r3d-n	Va-f1p-n
Npmsl	Rgp
Vmpp-sm	Pd-nsg
Pp3fsd–y	Pr-nsn
Agpmsny	Vmbr3s
Npmsg	Vmbr2p
Vmem2s	Pd-msa
Ncmsi	Pd-fsa
Vmpp-sf	Pd-msl
Px—d–y	Va-r2p-n
Mdo	Pi-msa
Aspfsn	Va-f1s-n
Aspmsnn	Pd-fsn

Table 2: Most significant differences in morphosyntactic categories used in the KRES balanced corpus of Slovene and the Parlameter corpus

would skew the distributions we are interested in. We also disregard all speeches of speakers for whom we do not have the necessary metadata.

4.1. Gender

The basic statistics regarding the number of speakers per gender and their linguistic production are given in Table 3. In total, the gender information is available for 139 speakers. 85 or 61% of those are male while 54 or 39% are female. Male speakers delivered three quarters of the speeches while female speakers only one quarter. However, on average, the speeches given by female speakers were 20% longer than those by male speakers.

	Male	Female
# of speakers	85	54
% of speakers	61	39
# of speeches	88,896	31,072
% of speeches	74	26
Avg # of words per speech	355	424

Table 3: Basic statistics regarding the gender

A keyword analysis, based on the Log Likelihood score given the male and female subcorpus, is presented in Figure 3. Among the top 100 keywords from the female speeches, apart from the general (17%) and administrative (32%) vocabulary, the most prominent topics are health (17%), social issues (13%), family (8%), and environmental (8%) issues, followed by education (2%) and finance (2%). In terms of word types, by far the most prevalent are nouns (62%) and adjectives (25%). Among the top 100 keywords from the male speeches, there is much more general (51%) and administrative (30%) vocabulary, which mostly pertains to the meta-discussions of the procedures in parliamentary sessions, followed by proper names (11%). Specific topics are few and far between: transportation (6%), technology (1%) and finance (1%).

This analysis is very general as keywords were classified out of context and in cases of polysemous keywords, only the most predominant sense was considered, but still gives a valuable insight into the contributions by male and female speakers. That the nature and style of male speeches is quite different from the female ones can also be seen from the analysis of the types of words ranked as the most specific for male speeches. While nouns are the most frequent category here as well (42%), much more of those are used to address or refer to other people, e.g., *gospod, kolega, poslanec, predsednik*, and proper nouns, i.e., names of colleague MPs, ministers, parties and companies. Keywords from male speeches contain many more verbs (15%), adverbs, pronouns and particles, indicating a much more discursive and debating style than female speeches.

4.2. Level of education

In this section we present the basic statistics regarding the number of speakers per each education level and their linguistic production in Table 4. The codes for levels of education are the following:

- 5: secondary school
- 6/1: higher education degree
- 6/2: university bachelor degree
- 7: university master degree
- 8/1: scientific master degree
- 8/2 scientific doctorate degree

The statistics show that most of the members of parliament hold the old university or the new Bologna master degree (7), with a similar number of members holding the new Bologna bachelor degrees (6/1 and 6/2) and the old scientific masters or PhD (8/1 and 8/2) degrees.

Regarding the number of speeches given, the distribution roughly follows the distribution of speakers, with the least educated speakers speaking less frequently. These speakers, however, hold the longest speeches, which is an exception as overall the length of speeches tends to grow with the level of education.

4.3. Age

We organize the speakers' age by the decade in which they were born. The basic statistics regarding the number of speakers per each age group and their linguistic production in Table 5. The results of the analysis show that the most represented group are speakers born in the 1960s who were in their 40s and 50s in the mandate covered by the corpus. The most active group (roughly estimated as the difference between the percentage of speakers and the percentage of speeches given) are the youngest and the oldest members

	5	6-1	6-2	7	8-1	8-2
# of speakers	12	4	24	30	17	14
% of speakers	14	5	28	34	20	16
# of speeches	9,850	2,158	25,469	38,693	14,784	16,335
% of speeches	9	2	24	36	14	15
Avg # of words per speech	616	388	430	483	526	504

Table 4: Basic statistics regarding the education level

	1940s	1950s	1960s	1970s	1980s
# of speakers	8	34	44	36	13
% of speakers	6	25	33	27	10
# of speeches	14,804	22,187	32,421	26,372	22,298
% of speeches	13	19	27	22	19
Avg # of words per speech	210	502	590	509	407

Table 5: Basic statistics regarding age (decade of birth)

of parliament, giving roughly twice the amount of speeches than their representation is.

Interestingly, the average length of the speeches given follows roughly the distribution of the number of speakers in each age group, with the members born in the 60s holding the longest speeches, while the shortest speeches, more than half in length, are given by the oldest members. The youngest members also hold significantly shorter speeches than the three central age groups.

4.4. Political orientation

Our final analysis considers the activity and linguistic production of members given their party affiliation. The results for the six parties with the highest number of active members of parliament are given in Table 6.

The results show that all parties expect SMC give more speeches than their member number would suggest. The most active are SD and Levica, both left-wing, with SD one of the ruling parties and Levica in the opposition in this composition of the parliament. They account for twice the amount of speeches than their member count.

Regarding the length of the speeches, the speeches of the opposition parties are much longer than those by the ruling parties. The longest speeches are given by the right-wing SDS party, followed by another right-wing party, the NSI. The average length of their speeches are 4 times longer than those of DeSUS and SD who give the shortest speeches.

The top 100 keywords from the speeches given by members of the six most prominent parties are displayed in Figure 2. The biggest ruling party SMC's keywords clearly reflect their position and role in the parliament, which is to propose and pass legislation as well as take care of the procedural activities in sessions. Their keywords are very neutral and impersonal, highly procedural, administrative and legislative.

The keywords from their main opposition SDS, on the other hand, are much more discursive, critical and emotional. The most prominent topics in SDS speeches seem to be the judiciary branch, health care and migrants.

The member of the coalition, the DeSUS party, mostly dealt with the social welfare system and health care topic-wise but also made a lot of procedural comments and interacted with and referred to a lot of relevant actors by name or position.

The keywords of the third coalition party SD are almost exclusively related to procedural activities and references to other individuals.

The keywords of the opposition right-wing party NSI show big differences with their closest party SDS. They are very program-driven, mainly tackling economic issues. Interestingly, NSI is the only party with explicit references to religion.

Finally, the keywords of the left-wing opposition party Levica clearly show the core values and goals of this party, which are social rights and equality. Interestingly, the style of the keywords of Levica range from colloquial (e.g., *blazno, bajta*) to sophisticated (e.g., *nemara, ubesedovati*), thereby differing quite a lot from the rest of the parties in the parliament.

5. Conclusion

In this paper we presented the Parlameter corpus of contemporary Slovene parliamentary proceedings. We analyzed the linguistic production of the speakers according to the speaker metadata. We have shown that while male speakers take the floor much more often than their female colleagues, females make longer contributions. Female speakers mostly address the topics of social, health, family and environmental issues, while male speakers do not cover specific topics, but differentiate in using more verbs, adverbs, pronouns and particles, indicating a more discursive and debating style. In terms of education level, speakers with PhDs too deliver more but shorter speeches. Older speakers (those born in the 1950s) rarely speak but their speeches are the longest. When comparing speeches according to party lines, they are evenly distributed according to party representation in the parliament, most likely due to parliamentary bylaws. The average speech lengths of the ruling parties SMC, DeSUS and SD are the same whereas nekako tudi predlog tikati odbor praviti ter moderen član glede smc navzoč sprememba glasovati predlagan poudariti zveza unija član določen vsebina pravzaprav smer podlaga amandma omenjen zakon zakonodajno izpostaviti obravnava članica lesen omeniti postopek kar vidik prehajati center vseeno morda umetniški vsekakor odstavek obravnavati pristojnost namen okvir les sprejet evropski cilj smisel zelo prostor ureditev izpostavljen pojasniti naveden matičen nadomeščati obrazložitev raven republika določati predstavnica pripomba poslovnik vložen sklep skupina obveščati področje pomemben 2017 urejati točka sklad javnofinančen razprava tale telo poslanka faza predlagatelj podati dopolnjen praven nek uporaba malce nadalje odločanje aktivnost vezan turizem delovanje sodelovanje predložiti beps tako

biti demokratski govoriti pogledati tisti slovenski vedeti reči nič kaj ali sodnik stranka koalicija zdaj narediti noben komisija predsednik povedati potem ministrica sodišče človekov nekdo mandatno takrat soden dati vrhoven mandat napisati vaš kršiti problem danes gledati ukc kakšen niti preiskovalen stvar pač jaz ustaven ampak janez pisati tam korupcija očitno minister koliko janša iti kandidat zadeva migrant cerar dobiti zgoditi tožilec enkrat predlagati vlada prej denar ker pravnik ilegalen senat sedeti opornica nikoli pravosoden policija sodnica zakaj dejati priti samo žilen klemenčič zdajle sodstvo zločin verjetno mark največkrat kpk davkoplačevalec kjer kako volilen človek vprašati spraševati državljanka pravosodje kangler

lep izvoliti desus skupina poslanski hvala predlog stališče predstaviti pokojnina mag seveda gospod upokojenec gospa beseda zakon zdravstven pokojninski poslanka franc ter invalid dopolnitev socialen novela pripraviti predstavitev anja prehajati kultura marija delo področje matičen zdravstvo horvat naj torej prekinjati jožef matej predlagatelj irena pozdrav obravnava zavarovanje namreč tašner žan upokojen majcen miha star žnidar usklajevanje dajati uroš marko zdravje dostojen zaključevati celarc dušan blagajna žibert telo zavod kolar regres andrej dneven milojka nekateri podpredsednik spoštovan mahnič prikl javen kordiš sprememba tanko marijan bojan kulturen pristojen dimic tomaž pojbič branko zdravko janko čuš tedaj proceduralno vatovec dediščina predložiti podkrajšek bah

izvoliti demokrat beseda socialen gospod mag gospa poslanski skupina lep hvala želeti izčistiti pomemben dajati replika predlagatelj razprava rast obravnava zagotovo predstavitev tudi razpravljati zaključevati okvir hip postopkovno hainz pravzaprav anja primož pripraviti franc jožef horvat dopolnitev znotraj prehajati matej imeti izjemno kolikor predstaviti ter stališče sicer bah deti žibert vendarle bistveno položaj potrebno jože gospodarski tomaž marko janko marija branko predlog banka prekinjati zakon ključen zahteven muršič poskušati izobraževanje obdobje zvonko godec zame ugotavljati zato kočevski ferluga možnost mogoč vselej andrej potek kordiš prijava holding mlakar srečevati jan dimic tanko iva lisec veber slediti peter skozi kriza podkrajšek bizjak

nov slovenija krščanski naš digitalen evropski občina podjetnik jaz kolegica drag najbrž dober pomurski unija evro vendarle kohezijski morda država projekt davčen želeti tisoč pomurje leto denar vipavski gotovo družina mlad regija demokrat obžalovati gospodarstvo lastnik stvar reforma bančen družinski zemljišče kmetijski komunalno posloven investitor kmet politika razumeti komunalen donacija penzija zgraditi lizbonski božičnica kolega zasedanje zunanji vladen obrtnik pokojninski praktično demografija enostavno parlament kibernetski konkurenčnost poudarjati župan vesel program regresen odpadek asistenca parcela operativen blagajna agenda podonavski gozd kapica bog rodnost vplačevati vodovod strukturen ikt okrog bolezen zdravljenje zgrajen gradben fantastičen šola graditi plečnikov bolniški članica piten lanski proračunski

levica združen nek navsezadnje skratka malo kapital delavec delavski penez privatizacija nekako bistvo revščina hoteti desnica resno socialist gor pogosto odkrito reven dol rad nato bolj sporazum podjetje tuliti minimalen prečenje debata kapitalizem politika žica resen predsedujoč dobiček stoletje konoplja stanovanjski neoliberalen onkraj bogat maribor koper resoren pogovarjati brati zaposlen plača socialen prečiti cel begunec sočasno četrt nehati lobi korporacija bajta firma prekaren ameriški izhajajoč privaten žival evro deregulacija profit skoraj neenakost kot rezilen družben stanovanje prebivalec pol zgodba ampak citat čeprav ips izvršilen lekarna blazno logika namesto težiti ubesedovati nemara človek tukaj neki trenutno živeti soupravljanje minimum vračljivost niti

Figure 2: Most prominent terms in speeches given by members of six most prominent parties (SMC, SDS, DeSUS, SD, NSI, Levica)

otrok zdravstven javen ukrep zavod pravica zdravnik starš tudi socialen zdravstvo čakalen ukc program področje družina varstvo leto sprememba družinski ter človekov zdravje bolnišnica opornica izobraževanje nasilje žilen res meniti pomoč peticija ministrstvo torej sredstvo delo prav pacient višina kakovost izvajanje šola dodatek center novela doba priprava pozdravljen letošnji oskrba žival romski storitev bolnik oseba krma dejavnost varstven enak preživnina dobavitelj otroški posamezen naročilo medicinski ženska ekološki zavarovanje vendar ureditev potreba zdravilo odhodek podneben mleko mark predlagan zakon potreben živilo obvezen gensko zaposlovanje dolgotrajen cilj duševen vsekakor sicer dostopnost oziroma podati brezposeln odrasel medical izguba transfer rastlina jamstvo denaren 2016

nek imeti reči gospod tisti mag zbor hoteti gledati naprej mandatno malo jaz dneven tir red predsednik tam noben ali zgodba navzoč poslanec gospa vlada levica volilen ura resnica iti zadeva tanko seveda točka navsezadnje resen beseda glasovati državen dalje najbrž zdaj videti prehajati luka kolega kakšen biti stvar postopkoven moj kandidat franc početi priti združen matej relativno nekaj postopkovno koper obrazložitev vsaj praviti banka tak seja proti jože razumeti sds predstavitev kaj trček verjetno digitalen poslanski gor stališče minister infrastruktura misliti preprosto nekdo ime resno zaključevati opozicija uber janša prekinjen nekako minuta sklep dol promet železnica tonin ker glas

Figure 3: Most prominent terms in female and male speeches

	SMC	SDS	DeSUS	SD	NSI	Levica
# of speakers	43	21	13	9	6	6
% of speakers	44	21	13	9	6	6
# of speeches	20656	23876	17340	17367	8788	10753
% of speeches	21	24	18	18	9	11
Avg # of words per speech	366	522	151	152	462	427

Table 6: Basic statistics regarding political orientation

the opposition parties the speeches of SDS, NSI and Levica are more than twice longer.

In the future we plan to enrich the corpus with additional session records of other parliamentary seatings but also with additional metadata available through the Parlameter system, such as voting data and accepted legislation, which are also valuable for addressing a number of research questions in various research communities. In parallel, we also plan to develop comparable corpora from other parliaments, starting with Croatian and Bosnian.

Acknowledgements

The work presented in this paper was partially 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, 2017-2020).

6. References

Paul Bayley. 2004. *Cross-cultural perspectives on parliamentary discourse*, volume 10. John Benjamins Publishing.

Jennifer E Cheng. 2015. Islamophobia, muslimophobia or racism? parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. *Discourse & Society*, 26(5):562–586.

Daniel Norman Chester and Nona Bowring. 1962. *Questions in parliament*. Clarendon Press.

Darja Fišer and Jakob Lenardic. 2018. Parliamentary corpora in the CLARIN infrastructure. In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017*, number 147, pages 75–85. Linköping University Electronic Press.

Mark N Franklin and Philip Norton. 1993. *Parliamentary Questions: For the Study of Parliament Group*. Oxford University Press, USA.

Graeme Hirst, Vanessa Wei Feng, Christopher Cochrane, and Nona Naderi. 2014. Argumentation, ideology, and issue framing in parliamentary discourse. In *ArgNLP*.

Lorna M Hughes, Paul S Ell, Gareth AG Knight, and Milena Dobreva. 2013. Assessing and measuring impact of a digital collection in the humanities: An analysis of the sphere (stormont parliamentary hansards: Embedded in research and education) project. *Digital Scholarship in the Humanities*, 30(2):183–198.

Pasi Ihalainen, Cornelia Ilie, and Kari Palonen. 2016. *Parliament and Parliamentarism: A Comparative History of a European Concept*. Berghahn Books.

Cornelia Ilie. 2017. Parliamentary debates. *The Routledge Handbook of Language and Politics*.

Nikola Ljubešić, Tomaž Erjavec, Darja Fišer, Tanja Samardzic, Maja Milicevic, Filip Klubička, and Filip Petkovski. 2016. Easily accessible language technologies for Slovene, Croatian and Serbian. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, pages 120–124.

Andrej Pančur and Mojca Šorn. 2016. Smart big data: Use of slovenian parliamentary papers in digital history. *Prispevki za novejšo zgodovino/Contributions to Contemporary History*, 56(3):130–146.

Andrej Pančur, Mojca Šorn, and Tomaž Erjavec. 2017. *Slovenian parliamentary corpus SlovParl 2.0*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1167`.

Andrej Pančur. 2016. Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI (Encoding the Slovenian Parliament Session Minutes in Line with the TEI Guidelines). In *Proceedings of the Conference on Language Technologies and Digital Humanities*, pages 142–48.

Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, 11(12):e0168843.

Teun A Van Dijk. 2010. Political identities in parliamentary debates. *European Parliaments under Scrutiny. Discourse strategies and interaction practices*, pages 29–56.

KAS-term and KAS-biterm: Datasets and baselines for monolingual and bilingual terminology extraction from academic writing

Nikola Ljubešić,[*] Tomaž Erjavec,[*] Darja Fišer[†][*]

[*]Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ijs.si, tomaz.erjavec@ijs.si

[†]Department of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 6, SI-1000 Ljubljana
darja.fiser@ff.uni-lj.si

Abstract

This paper presents two datasets for supervised learning of terminology extraction. The first is focused on monolingual term extraction and is a lexicon-type dataset of Slovene term candidates labeled by four annotators. The second is focused on extracting and linking terms in different languages which are translations of each other. It contains sentences that satisfy patterns in which terms occur frequently with their translations, with manually annotated terms in English, Slovene and other languages, and links between terms and their translations. For each dataset we set up a baseline approach: for monolingual terminology extraction we train an SVM classifier, while for identifying terms in different languages we train a sequential CRF classifier. The datasets and the described baselines are made freely available.

1. Introduction

In this paper we present two new datasets for training term extraction tools developed in the scope of the Slovene national project KAS, *Slovene scientific texts: resources and description*.

KAS-term is a lexicon-type dataset containing term candidates extracted via morphosyntactic patterns from a selection of PhD theses written in Slovene. Each term candidate is annotated by multiple annotators. The dataset is meant to be used for supervised learning of ranking of term candidates extracted from Slovene texts.

KAS-biterm is a sentence-type dataset consisting of sentences that satisfy some patterns that are typical for terms and their translations into other languages such as "*ekstrakcija terminologije (angl. term extraction)*". These sentences are annotated for terms, partial terms and abbreviations in Slovene, English, or other language. Links between the Slovene terms and their terms or abbreviations in the other languages are encoded as well.

On both datasets baseline approaches are defined and evaluated: for monolingual terminology an instance-level SVM binary classifier is defined which uses various co-occurrence statistics as features, while for bilingual terminology a sequence-level CRF classifier is defined which uses context-based features and annotates each token in a candidate sentence with the respective category.

The rest of this paper is structured as follows: Section 2. gives the related work on terminology extraction and describes the KAS corpus of Slovene academic writing, from which the presented datasets are produced. Section 3. describes in detail the monolingual datasets and the implementation and evaluation of our baseline, while Section 4. does the same for the bilingual case. Finally, Section 5. gives some conclusions and directions for future research.

2. Related work

In this section we give a description of related work in monolingual and multilingual terminology extraction.

2.1. Monolingual terminology extraction

A broad overview of linguistic, statistical and hybrid approaches to automatic terminology extraction (ATE) is given in Pazienza et al. (2005).

The term recognition task is usually formulated as a two-step procedure (Nakagawa and Mori, 2003): candidate term extraction followed by term scoring and ranking. We also follow this approach for monolingual term extraction.

There is a number of ATE datasets already available. Handschuh and QasemiZadeh (2014) present ACL RD-TEC, a dataset for evaluating the extraction and classification of terms from literature in the domain of computational linguistics. The dataset is based on the ACL ARC corpus consisting of papers from the ACL anthology. From that corpus more than 83,000 term candidates are extracted via PoS-based filtering, n-gram-based techniques and noun phrase chunking. They are furthermore annotated either as non-terms, technology terms or non-technology terms. Out of the 84k terms, 22k were annotated as being valid while 62k were annotated as invalid. The authors report an observed agreement of 0.758 and Cohen's κ of 0.517, on a small double-annotated dataset of 250 terms.

A reference dataset for terminology extraction is the GENIA corpus consisting of 2,000 MEDLINE abstracts from scientific publications in biomedical literature that is accompanied by the annotations of 100,000 terms organized in a well-defined ontology (Kim et al., 2003). Another example of a bio-textmining dataset is The Colorado Richly Annotated Full Text Corpus (CRAFT), consisting of 97 articles from the PubMed Central Open Access subset annotated with biomedical concepts (Bada et al., 2012). The authors of the dataset measure weekly inter-annotator agreement (IAA), showing expected improvements through time, as well as an F1 IAA of above 90% after a few weeks / meetings for five out of six tasks. However, the tasks consisted of applying ontologies on text, and not of labeling terms as an open task.

Another reference dataset is a corpus for the evaluation of term extraction in the domain of automotive engineering (Bernier-Colborne and Drouin, 2014). The authors annotate running text, but allow for evaluation of extracted lists of term candidates.

Combining various statistical predictors in a supervised learning setting is a well known approach in natural language processing and has been also applied to the problem of automatic term extraction. Loukachevitch (2012) combines 16 features, and with their logistic regression combination improves the best single result by removing 30-50% of error, depending on the domain. Similarly, Conrado et al. (2013) show on three domain corpora of Portuguese that a combination of 19 features significantly outperform separate well known statistics for ATE.

A very similar problem to ATE is collocation extraction where Pecina and Schlesinger (2006) obtain 21.53% relative improvement when combining 82 association measures with respect to the best individual measure. They also show that feature selection can bring the number of features down to 17 without a significant loss in the evaluation metric.

2.2. Bilingual terminology extraction

Bilingual terminology extraction is typically performed on parallel data (Daille et al., 1994; Vintar, 2010). Another popular line of research is multilingual term extraction from semi-structured multilingual knowledge banks, such as Wikipedia, relying on explicitly encoded cross-lingual links (Gupta et al., 2008; Erdmann et al., 2008). However, since (extensive) parallel corpora and other types of multilingual knowledge sources are difficult to obtain for a lot of specialized domains, researchers are increasingly proposing approaches that extract terms from partially translated (Nagata et al., 2001) or comparable (Tanaka and Iwasaki, 1996) data, where they extract terms for each language separately and then perform post-hoc term pairing.

In this paper we take a different approach, identifying patterns that are used to express the Slovene term and its translation equivalent into English or another foreign language in largely monolingual scientific texts, thereby considering the task to be a classical sequence annotation task. A similar approach has been proposed by Bond (2008) who used a small set of manually defined patterns to extract bilingual term pairs from the web. Abekawa and Kageura (2009) and Abekawa and Kageura (2011) proposed an extension of this basic approach in which they first extract seed bilingual terms from the available parallel glossaries and then use the seed term pairs to identify typical patterns that are used between them, which then serve as the basis of the large-scale bilingual term extraction from the web.

2.3. The corpus

The KAS corpus (Erjavec et al., 2016) was collected via the Open Science Slovenia aggregator (Ojsteršek et al., 2014) which harvests the (meta)data of the digital libraries of Slovene universities and other research institutions. The corpus contains mainly Bachelors, Masters and Doctoral theses and comprises almost 1 billion tokens. The texts were extracted from PDF files, and, after some filtering and cleaning, were tagged with morphosyntactic descriptions (MSDs) and lemmatised with reldi-tagger[1] (Ljubešić and Erjavec, 2016) using its model for Slovene. Each text in the corpus is accompanied with extensive meta-data, containing also classificatory information, such as CERIF (Common European Research Information Format) keywords.

The current, preliminary, version of the KAS corpus contains 700 PhD theses (40 million tokens) from a large range of disciplines[2] and it is this subcorpus that was used as the textual basis for the experimental datasets presented in this paper.

3. Monolingual term extraction

3.1. The dataset

For the term extraction experiments presented here we focused on three fields: Chemistry, Computer Science, and Political Science, which we selected by matching them with their CERIF keywords, thus obtaining 48 PhD theses form Chemistry, 105 from Computer Science, and 23 from Political Science.

From these three subcorpora we sampled 5 PhD theses per area and automatically extracted term candidates, using the CollTerm tool (Pinnis et al., 2012) given a set of manually defined term-indicative MSD patterns. These patterns were initially developed for the Sketch Engine (Kilgarriff et al., 2014) terminology extraction module, and are in detail described in Fišer et al. (2016). For the present experiments we used only 31 nominal patterns, from unigrams and up to 4-grams, e.g. `Nc.*,S.*,Nc.*,Nc.*g.*` which finds sequences of *common noun, preposition, common noun, common noun in the genitive case*, such as *adheziv na osnovi topil* (*adhesive on basis (of) solvents = solvent-based adhesive*).

Each found term candidate was extracted in the form of its lemma sequence and the most frequent inflected phrase, keeping those that appear at least three times in a doctoral thesis. For manual annotation the candidates were first alphabetically sorted, in order to remove bias coming from frequency or statistical significance of co-occurrence, both types of information being provided by the CollTerm tool.

We produced a separate list of term candidates for each doctoral thesis. These lists were then annotated by four annotators. Annotators, who were graduate students of the three fields in focus, were asked to label each potential term with one of the five labels:

- *in-domain*: words and phrases that represent an in-domain term, i.e. one from the focus field;
- *out-domain*: words and phrases that represent a term from a field other than the one in focus;

[1] `https://github.com/clarinsi/reldi-tagger`

[2] The body parts of the KAS corpus and the KAS-Dr (PhD theses only) corpus are available for exploring through the concordancer at CLARIN.SI: KonText (`http://www.clarin.si/kontext/`) and noSketch Engine (`http://www.clarin.si/noske/`).

```
{
"document_id": "kas-845894",
"area": "Kemija",
"annotation_round": 1,
"lemmas": "gradient magneten polje",
"wordforms": "gradientom magnetnega polja",
"pattern": "Nc.*|A.*g.*|Nc.*g.*",
"length": 3,
"annotator_1": "t_termin",
"annotator_2": "t_termin",
"annotator_3": "n_nerelevantno",
"annotator_4": "n_nerelevantno",
"frequency": 7,
"tfidf": 0.11325,
"chisq": 0.79361,
"dice": 0.11079,
"ll": 0.25669,
"mi": 0.55263,
"tscore": 0.1324,
"cvalue": 11.09473
}
```

Figure 1: JSON encoded monolingual dataset entry

- *general*: vocabulary that is typical for academic discourse;
- *irrelevant*: words and phrases that belong to the general vocabulary, foreign-language expressions, definitions, fragments of terminology;
- *discuss*: borderline cases that needed to be discussed and resolved. These do not occur in the final dataset.

The instances of the dataset are thus term candidates annotated with the above categories, and various frequency and co-occurrence statistics. The final dataset consists of 22,950 such instances.

As illustrated in Figure 1, the fields of each instance are the thesis identifier, the scientific field, annotation round, lemma sequence, its most frequent surface form, morphosyntactic pattern, length in words and the manual annotations by annotator number 1 – 4. We also encode seven statistics calculated with the CollTerm tool during the term candidate extraction. These statistics are the frequency of the term candidate, and its tf-idf, χ^2, dice, log-likelihood point-wise mutual information and t-score values. Due to its popularity we also give the C-value (Frantzi et al., 2000), although this statistic is not based on co-occurrence, but the frequency of the term candidate and the frequency and number of other candidate terms containing that term candidate.

We distribute this dataset both in JSON and CSV formats. It is available from the CLARIN.SI repository (Erjavec et al., 2018b).

3.2. Baseline method

We set up a baseline for the task of predicting whether a candidate is a term or not given the variables available in the prepared dataset. We build the baseline as an SVM classifier with `scikit-learn` (Pedregosa et al., 2011)[3].

Given that we have four labels present in our dataset, we defined two mappings (inclusive and exclusive) of the four labels to a binary system of positive and negative classes. Both the inclusive and exclusive mappings take the irrelevant terms as instances of the negative class, but the inclusive mapping considers out-of-domain terms and academic vocabulary to be instances of the positive class, while the exclusive mapping considers them to be negative class instances. In the remainder of the paper we experiment with the more strict, exclusive mapping.

The explanatory variables we have at our disposal are the already mentioned frequency and seven co-occurrence statistics: *frequency*, *dice*, *chisq*, *ll*, *mi*, *tscore*, *tfidf*, and *cvalue*.

We consider the response variable to be the rounded average of the human responses, i.e., if three annotators claim an instance to be a term, and one annotator the opposite, the gold response for this term will be 1, i.e., the positive class. In (infrequent) cases where the average is 0.5, it is rounded up to 1.

We separate the prediction of multi-word terms (MWT) and single-word terms (SWT) as for single-word terms the only available variables are the frequency and the tf-idf statistic. For MWTs of all lengths all the seven variables are available.

We give the results on using single statistics, as well as the SVM classifier combining all the statistics in ranking multi-word term instances in a receiver-operating-characteristic (ROC) curve analysis in Figure 2. The ROC curve shows for each separate statistic to be surprisingly close to the random baseline (*baseline*), but that combining all these statistics in a supervised fashion (*all*) significantly improves the ranking of term candidates. If we quantify each ranking as an area under curve (AUC), our supervised baseline achieves a value of 0.736, while ranking by specific statistics achieves AUC scores between 0.505 (*tscore*) and 0.590 (*dice*).

For SWT ranking, where we have only two statistics at our disposal, namely *freq* and *tf-idf*, we calculated AUC scores for each separate statistic, as well as the ranking obtained through supervised learning on the two explanatory variables. The *freq* variable obtains an AUC of 0.523, *tfidf* performs much better with AUC of 0.703, while the combination of these two variables achieves an AUC of 0.613. Therefore as our baseline for SWT ranking we propose the *tfidf* statistic.

4. Bilingual term extraction

4.1. The dataset

The bilingual term extraction dataset contains complete sentences selected from all the PhD theses from the KAS corpus. We chose only sentences that have a high chance of containing the term in the original language and its translation into Slovene. The sentences were extracted using noSketch Engine via queries in its Corpus Query Language (CQL). After experimenting with various queries we

[3] The code of the baseline is published on `https://github.com/clarinsi/kas-term`

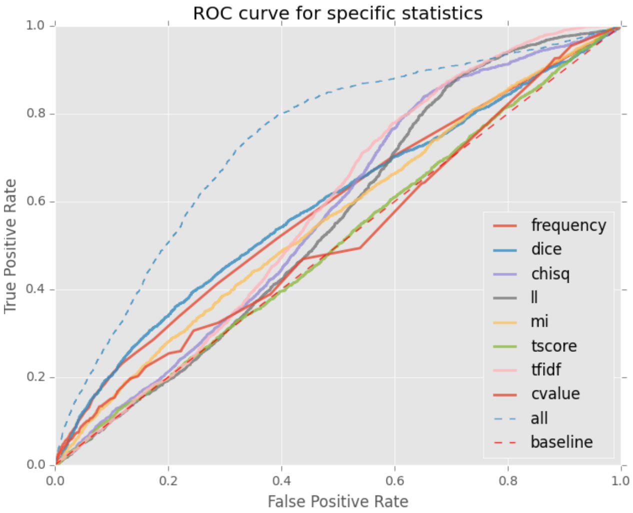


Figure 2: ROC curves for each of the variables in ranking multi-word term candidates, and for their combination *all*. The *baseline* is a random baseline.

then extracted the sentences with the following three CQL queries:[4]

1.
   ```
   "\(" ".*"?
   "an\.|ang\.|angl\.|angleš.+"
   [tag="U"]? [word!="\)"]+ "\)"
   ```

2.
   ```
   [tag!="Nj"] "\(" "."?
   [tag="Nj" & word="...+"]
   [word!="\)" & word!="[0-9,]+"]* "\)"
   ```

3.
   ```
   "[a-zA-ZščžŠČŽ]+" "ali" "[\'\"]"
   ```

The sentences retrieved by the queries were the basis for the manually annotated corpus. We first randomly sampled the results of the queries and then imported, for each query result separately, the sample into WebAnno (Yimam et al., 2013), a tool for Web-based manual annotation of corpora. Even though not all the annotations were used in the current baseline experiment, we, for the sake of completeness and possible further use, annotated the samples on the following levels:

- Type of term (*full term, partial term, abbreviation*): this distinction was made as the sample showed that the sentences often contain not only complete terms, but also terms which only partially cover its corresponding translation or original. Furthermore, the context of many terms or their translations also contains their abbreviation.

- Language of the term (*Slovene, English, Other*): even though our focus was on Slovene-English pairs, some found terms were also in other languages. We chose a middle road between ignoring these terms and marking them with their actual language, by assigning them all the *Other* language.

- Link between the term and its translation or between the term and its abbreviation (*link*): as the final goal is to automatically link terms and translations, the manual annotation of the link between the two is essential.

Each sentence was annotated by two annotators and then the differences in annotation were resolved by the curator. Table 1 gives the statistics over the dataset, by query and in total. The numbers of sentences and tokens show that the queries had a significantly different yield, while the ”Marked” column gives the number of sentences in which something was annotated, i.e. they contained either a term or abbreviation with its translation; the last query thus not only returned the least sentences, but even the ones returned were typically not marked. The next three columns give the distribution by the type of the entity marked: in all cases, complete terms predominate, with abbreviations being about one tenth as frequent, and partial terms even less. Finally, the last three columns give the distribution by language: naturally, the Slovene and English items are quite similar in size, with other languages representing a very small minority.

The dataset was exported from WebAnno and merged with the source TEI encoding of the corpus as illustrated bin Figure 3. Here, the type of term is distinguished by the name of the element (`abbr` or `term`) and, in the case of terms, its `@type` attribute (`complete` or `partial`)q, while the language is distinguished by the value of the standard `@xml:id` attribute. Furthermore, the value of the `@subtype` gives the tag as it was used in WebAnno. The linkings are made via the `@corresp` attribute, which points to the value of the `@xml:id` attribute of the relevant term(s) or abbreviation(s). It should be noted that all the pointers are two-way.

The dataset is freely available in the scope of the CLARIN.SI repository (Erjavec et al., 2018a).

4.2. Baseline

Given that in this task we have running text instances annotated per token with term information, we frame this task as a sequence labeling task. Similar as with the task of monolingual term prediction, we use the traditional method applicable given the type of data: we use CRF, in particular the CRFSuite implementation (Okazaki, 2007). The baseline is published on `https://github.com/clarinsi/kas-biterm`.

Since the first pattern is the most productive one, as well as having a much higher precision than the remaining two patterns, we run the baseline experiments only on the 1,000 sentences following that pattern. The goal of the defined baselines is, namely, not only to set the stage for future experiments, but also to produce systems that will be easily applicable to various datasets, starting with the full KAS corpus. We split the available instances 80:20 into a training and a testing set.

We experimented with various features and, given the results of our experiments, we kept the following ones:

[4] Rather than explaining each query, we give links for returning the shuffled results of the three queries, in order: `http://hdl.handle.net/11346/clarin.si-ZNAN`, `http://hdl.handle.net/11346/clarin.si-7GRN`, `http://hdl.handle.net/11346/clarin.si-WHDX`.

Query	Tokens	Sents	Marked	Complete	Partial	Abbrev	sl	en	und
q1	36,716	1,000	864	2,134	141	299	1,159	1,392	23
q2	34,773	787	427	1,324	51	169	696	707	141
q3	7,002	165	36	81	1	1	40	39	4
Σ	78,491	1,952	1,327	3,539	193	469	1,895	2,138	168

Table 1: Statistics over the BiTerm dataset

```
<abbr xml:id="patt1-001.abbr.2"
      xml:lang="en"
      corresp="#patt1-001.term.9
               #patt1-001.term.8"
      subtype="4AbbrEng">
  <w lemma="msd" ana="msd:Ncmsn">MSD</w>
</abbr>
<c> </c>
<term xml:id="patt1-001.term.8"
      xml:lang="sl"
      type="complete"
      corresp="#patt1-001.term.9
               #patt1-001.abbr.2"
      subtype="2TermSlv">
  <w lemma="oblikoskladenjski"
     ana="msd:Agpfsg">oblikoskladenjske</w>
  <c> </c>
  <w lemma="oznaka"
     ana="msd:Ncfsg">oznake</w>
</term>
<c> </c>
<pc ana="msd:Z">(</pc>
<w lemma="angl." ana="msd:Y">angl.</w>
<c> </c>
<term xml:id="patt1-001.term.9"
      xml:lang="en"
      type="complete"
      corresp="#patt1-001.abbr.2
               #patt1-001.term.8"
      subtype="1TermEng">
  <w lemma="Morpho"
     ana="msd:Npmsn">Morpho</w>
  <c> </c>
  <w lemma="Syntactic"
     ana="msd:Npmsn">Syntactic</w>
  <c> </c>
  <w lemma="Description"
     ana="msd:Npmsn">Description</w>
</term>
<pc ana="msd:Z">)</pc>
```

Figure 3: Example of a TEI bilingual term annotation for the segment *MSD oblikoskladenjske oznake (angl. Morpho Syntactic Description*)

.

- focus token: lowercased token for which features are currently extracted
- focus MSD: morphosyntactic description of the focus token
- focus PoS: part-of-speech of the focus token (first two letters of the morphosyntactic description tag)
- focus token length: number of characters in the focus token
- focus token case (lower, upper, title)
- lower cased tokens in a -3...3 window
- PoS tags in a -3...3 window

While performing baseline experiments, we calculated the informativeness of each feature set by performing ablation experiments. We ablated specific features, but also the set of features based on the focus token and the set of features based on the context window. We present the results of the ablation experiments in Table 2.

The results show that the most relevant feature sets are those of the focus token's context window, with the largest loss being when all window features are removed (8.89% relative loss), followed by the setup where all focus token features are removed (2.65% relative loss). Removing specific features generates a relative loss ranging between 1% and 0.1%.

We also experimented with other features, but they decreased our results. These are the features with their relative loss when added to the optimal feature set:

- focus token character 5-grams (best performing length), extended with a initial and ending character (loss of 0.2%)
- MSDs in a -3...3 window (loss of 0.3%)
- 100 embedding dimensions learnt from the slWaC corpus with fasttext using the skipgram model (loss of 0.3%)

The most surprising among the negative results is the loss when word embedding features are added to the sequential classifier. This result can probably be explained with the sensitivity of the CRF classifier to irrelevant features as most of the embedding dimensions do not hold any relevant information for the task at hand.

The full results of our best performing system (comparable to the system in ablation experiments with no ablated features) are presented in Table 3. As expected, the SL-ABBR class performs the worse as the number of tokens annotated with this label is by far the lowest. The class EN-TERM is better predicted as the class SL-TERM, which is also not surprising as identifying the borders of an English term in Slovene text is much easier than the borders of a Slovene term. Regarding the balance between precision and recall, there are no surprises with a good overall balance.

Ablated features	0	SL-TERM	SL-ABBR	EN-TERM	EN-ABBR	weighted
support	6177	601	10	527	66	7381
none	0.969	0.789	0.000	0.896	0.683	0.945
focus token	0.968	0.778	0.000	0.896	0.634	0.943
focus MSD	0.968	0.776	0.000	0.892	0.710	0.943
focus PoS	0.966	0.755	0.000	0.890	0.708	0.940
focus length	0.968	0.773	0.000	0.894	0.698	0.943
focus case	0.969	0.778	0.000	0.895	0.650	0.944
all focus token	0.957	0.702	0.000	0.815	0.452	0.920
tokens in window	0.964	0.733	0.000	0.894	0.625	0.936
PoS in window	0.968	0.771	0.000	0.896	0.672	0.943
all window	0.924	0.289	0.000	0.845	0.370	0.861

Table 2: Ablation experiments over the feature sets used for bilingual term extraction. The labels the results are given for are: O (other), SL-TERM (Slovene term), SL-ABBR (Slovene abbreviation), EN-TERM (English term), EN-ABBR (English abbreviation)

Metric	0	SL-TERM	SL-ABBR	EN-TERM	EN-ABBR	weighted
precision	0.965	0.839	0.000	0.872	0.737	0.945
recall	0.974	0.745	0.000	0.920	0.636	0.947
F1	0.969	0.789	0.000	0.896	0.683	0.945

Table 3: Final experiment on bilingual term extraction

5. Conclusions

We presented two newly developed manually annotated datasets for Slovene: the KAS-term dataset for learning monolingual term extraction and the KAS-biterm dataset for learning bilingual term extraction.

We set up baseline approaches with good, far from random results. However, we strongly believe that these results can further be improved and encourage other researchers and NLP practitioners to improve over these baselines and share their results.

Acknowledgements

We would like to thank the three anonymous reviewers for their helpful comments. We are also indebted to Špela Arhar Holdt and Maja Bitenc for conducting the annotation campaigns, to the the students who participated in the annotation process, and to their supervisors, Urban Bren, Marko Robnik Šikonja, and Boštjan Udovič. The research described in the paper was supported by the project ARRS J6-7094 “Slovene scientific texts: resources and description”.

6. References

Takeshi Abekawa and Kyo Kageura. 2009. Qrpotato: A system that exhaustively collects bilingual technical term pairs from the web. In *Proceedings of the 3rd International Universal Communication Symposium*, pages 115–119. ACM.

Takeshi Abekawa and Kyo Kageura. 2011. Using seed terms for crawling bilingual terminology lists on the web. *Trans. Comp.*

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner Jr., K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13:161.

Gabriel Bernier-Colborne and Patrick Drouin. 2014. Creating a test corpus for term extractors through term annotation. *Terminology*, 20:50–73.

Francis Bond. 2008. Extracting bilingual terms from mainly monolingual data. In *14th Annual Meeting of the Association for Natural Language Processing*, Tokyo.

Merley Conrado, Thiago Pardo, and Solange Rezende. 2013. A machine learning approach to automatic term extraction using a rich feature set. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 16–23.

Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th conference on Computational linguistics – Volume 1*, pages 515–521. Association for Computational Linguistics.

Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2008. An approach for extracting bilingual terminology from Wikipedia. In *International Conference on Database Systems for Advanced Applications*, pages 380–392. Springer.

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar, and Milan Ojsteršek. 2016. Slovenska znanstvena besedila: prototipni korpus in načrt analiz (Slovene Scientific Texts: Prototype Corpus and Research Plan). In *Proceedings of the Conference on Language Technologies and Digital Humanities*. Ljubljana University Press, Faculty of Arts.

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, and Maja Bitenc. 2018a. *Bilingual terminology extraction dataset KAS-biterm 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1199.

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, and Špela Arhar Holdt. 2018b. *Terminology identification dataset KAS-term 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1198.

Darja Fišer, Vit Suchomel, and Miloš Jakubiček. 2016. Terminology extraction for academic Slovene using Sketch Engine. In *RASLAN 2016: Recent Advances in Slavonic Natural Language Processing*, pages 135–141.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130, Aug.

Anand Gupta, Akhil Goyal, Aman Bindal, and Ankuj Gupta. 2008. Meliorated approach for extracting bilingual terminology from Wikipedia. In *Computer and Information Technology, 2008. ICCIT 2008. 11th International Conference on*, pages 560–565. IEEE.

Siegfried Handschuh and Behrang QasemiZadeh. 2014. The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics. In *COLING 2014: 4th International Workshop on Computational Terminology*.

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):7–36, Jul.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. In *ISMB (Supplement of Bioinformatics)*, pages 180–182.

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).

Natalia V. Loukachevitch. 2012. Automatic term recognition needs multiple evidence. In *Proceedings of the Eighth Conference on Language Resources and Evaluation, LREC 2012*, pages 2401–2407.

Masaaki Nagata, Teruka Saito, and Kenji Suzuki. 2001. Using the web as a bilingual dictionary. In *Proceedings of the workshop on Data-driven methods in machine translation-Volume 14*, pages 1–8. Association for Computational Linguistics.

Hiroshi Nakagawa and Tatsunori Mori. 2003. Automatic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–219.

Milan Ojsteršek, Mojca Kotar, Marko Ferme, Goran Hrovat, Mladen Borovič, Albin Bregant, Jan Bezget, and Janez Brezovnik. 2014. Vzpostavitev repozitorijev slovenskih univerz in nacionalnega portala odprte znanosti (The Set-Up of the Repository of Slovene Universities and the National Portal of Open Science). *Knjižnica*, 58(3).

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Maria Pazienza, Marco Pennacchiotti, and Fabio Zanzotto. 2005. Terminology extraction: an analysis of linguistic and statistical approaches. *Knowledge mining*, pages 255–279.

Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL*, pages 651–658, Stroudsburg, PA, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, and Tatiana Gornostay. 2012. Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the Terminology and Knowledge Engineering (TKE2012) Conference*.

Kumiko Tanaka and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 580–585. Association for Computational Linguistics.

Špela Vintar. 2010. Lusčenje terminologije iz angleškoslovenskih vzporednih in primerljivih korpusov (Terminology mining from English-Slovene parallel and comparable corpora). In Špela Vintar, editor, *Slovenske korpusne raziskave*, pages 37–53. Znanstvena založba Filozofske fakultete, Ljubljana.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible,web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 1–6, Stroudsburg, PA, USA, August. Association for Computational Linguistics.

Strokovno-znanstvena slovenščina: besednovrstne in oblikoskladenjske značilnosti

Nataša Logar,* Tomaž Erjavec†

* Fakulteta za družbene vede, Univerza v Ljubljani
Kardeljeva ploščad 5, 1000 Ljubljana
natasa.logar@fdv.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

V prispevku prikazujemo besednovrstne in oblikoskladenjske značilnosti slovenskih strokovno-znanstvenih besedil, do katerih smo prišli z metodo frekvenčnega profila, in sicer na podlagi primerjave štirih (pod)korpusov: celotnega uravnoteženega korpusa slovenščine Kres in njegovega leposlovnega dela ter disertacijskega in diplomskega dela korpusa akademske slovenščine KAS. Rezultati analize so med drugim pokazali, da so za slovensko strokovno-znanstveno pisanje bolj značilni samostalniki, pridevniki in okrajšave (primerjalno s Kresom pa najmanj glagoli, zaimki ter prislovi) in da med MSD-oznakami v disertacijah primerjalno s Kresom najbolj izstopajo občni samostalniki vseh treh spolov v ednini ali množini, ki so v rodilniku ali imenovalniku. Ugotovitve tako opozarjajo na slovnična mesta, ki jim bo treba pri pripravi prihodnjega opisa strokovno-znanstvene slovenščine posvetiti še več pozornosti.

Academic Slovene: Part-of Speech and Morphosyntactic Characteristics

The article presents PoS and MSD characteristics of academic Slovene. Using the frequency profiling method, we assembled the data by comparing four (sub)corpora: the balances corpus of Slovene Kres as a whole, its fiction part, the PhD part of the corpus of academic Slovene KAS and its BSc and BA thesis part. Among other findings, the results show that nouns, adjectives and abbreviations are used much more in academic genres (in contrast with verbs, pronouns and adverbs that are typical for non-specialized language). In PhD texts, the following MSDs stand out: common nouns of all three genders in singular or plural and in genitive or nominative. The findings point to grammatical issues that should not be overlooked in new descriptions of academic Slovene.

1. Uvod

Z veliko verjetnostjo je mogoče trditi, da je izmed strokovno-znanstvenih podzvrsti vseh jezikov najbolje raziskana akademska angleščina (npr. Biber in Barbieri, 2007; Gardner in Davies, 2013; Hyland, 2008). Enega njenih tehtnejših, obenem pa nejezikoslovnim uporabnikom razumljivejših opisov npr. najdemo v priročniku z naslovom *Academic Writing for Graduate Students: Essential Tasks and Skills* avtorjev J. M. Swalesa in C. B. Feak (2012). Pri pripravi te knjige sta si avtorja pomagala s korpusom študentskih izdelkov *Michigan Corpus of Upper-Level Student Papers*[1] (2004–2009; 2,6 milijona besed; Römer, 2009). To jima je omogočilo, da sta leksikalno-slovnične značilnosti akademskega pisanja ponazorila z realnimi zgledi, pri čemer sta izhajala iz pomena, namena in kohezivnosti besedil. Drugo, sicer izrazito na slovnične značilnosti akademske angleščine osredotočeno delo, ki pa ga izmed več takih prav tako velja izpostaviti, je priročnik *Advanced Grammar: For Academic Writing* avtorja R. Stevensona (2010). Stevensonovo delo sicer ni zasnovano korpusno, vendar pa avtor v njem pojave, kot so zgradba povedi, modalnost, kohezija, ton pisanja, samostalniškost, glagolski časi itn., razlaga na številnih primerih ter na bralcu prijazen način.[2]

Dela, ki bi bilo vsaj podobno prvemu ali drugemu od omenjenih priročnikov, za slovensko strokovno-znanstveno pisanje še nimamo, imamo pa obsežen Korpus akademskih besedil KAS (Erjavec et al., 2016), ki že omogoča analize, na katerih bi lahko bil osnovan tak opis. V prispevku zato na primeru besednovrstnih in oblikoskladenjskih podatkov, ki korpus KAS značilno ločijo od uravnoteženega korpusa slovenščine Kres (Logar Berginc et al., 2012), prikazujemo ugotovitve začetnih tovrstnih analiz in razmišljamo, kako nam podatki iz njih lahko pomagajo pri pripravi prvega, na obsežnem naboru besedil z različnih področij temelječega priročniškega prikaza te podzvrsti slovenskega jezika.

2. Metoda in (pod)korpusi

Podatke o besednovrstnih in oblikoskladenjskih oznakah besedil (angl. *part-of-speech*, dalje PoS; *morphosyntactic description*, dalje MSD), ki so zajeta v korpus KAS in korpus Kres, smo pridobili z metodo frekvenčnega profila (angl. *frequency profiling*).[3] Metodo sta zasnovala Rayson in Garside (2000), omogoča pa primerjavo dveh korpusov (ali podkorpusov) po ključnih besedah in slovničnih kategorijah. Rezultat primerjave so seznami, ki kažejo, kateri elementi so bolj značilni za enega oz. drugega od korpusov, če ju primerjamo med seboj. Na slovenskem gradivu je bila metoda prvič uporabljena za ugotavljanje razlik med korpusoma

[1] http://www.helsinki.fi/varieng/CoRD/corpora/MICUSP/

[2] Kratek pregled in primerjavo tujih priročnikov na temo akademskega pisanja gl. v Logar (2017: 25–40).

[3] Pri korpusu KAS smo uporabili njegovo različico v2.0 (2000–2015), https://www.clarin.si/noske/run.cgi/corp_info?corpname=kas (tudi tukajšnje Slike 1–4 in 6 so od tam), pri korpusu Kres pa različico 1.0, http://www.slovenscina.eu/korpusi/kres.

slovenščine Kres in Gigafida (Logar Berginc et al., 2012: 95–97), nato pa še za primerjavo prve različice spletnega korpusa slovenščine slWaC$_1$ z (a) njegovo nadgradnjo slWaC$_2$ ter (b) z obema že imenovanima korpusoma slovenščine, Kresom in Gigafido (Erjavec et al., 2015).

Preizkus je tudi tokrat potekal po enakem postopku: »najprej smo izdelali frekvenčni seznam lem /ter MSD oznak/ obeh korpusov, nato pa za vsako lemo /ter PoS in MSD oznako/ izračunali njeno logaritemsko verjetnost (angl. *log-likelihood*, LL). LL upošteva pogostost elementa, kot tudi velikosti obeh korpusov in večji, kot je, bolj je element značilen za enega od njiju. Elementi z najvišjimi vrednostmi razlik v LL /.../ najočitneje kažejo glavne razlike med korpusoma« (Erjavec et al., 2015: 38).

(Pod)korpusi, ki smo jih primerjali, so bili štirje:

a) Kres,
b) leposlovni del Kresa (dalje Kres-fict),
c) disertacijski del KAS-a (dalje KAS-dr) in
č) diplomski del KAS-a (dalje KAS-dipl).[4]

Iz frekvenčnih seznamov vseh štirih (pod)korpusov smo izločili enote, ki bi šumno obremenile končne sezname LL, tj. besede, zapisane v tujem jeziku, besede, zapisane s števkami, ločila, ki so bila označena kot »besede«, nize ločil ipd. Med seboj smo nato primerjali:

- KAS-dr : Kres,
- KAS-dr : Kres-fict,
- KAS-dipl : KAS-dr,
- KAS-dipl : Kres in
- KAS-dipl : Kres-fict.

Končnih seznamov je bilo skupno 15, torej po pet za vsak opazovani element (PoS, MSD in leme).[5]

KAS-dr : Kres	KAS-dr : Kres-fict	KAS-dipl : Kres	KAS-dipl : Kres-fict
1. S*	**1. S**	**1. S**	**1. S**
2. P	**2. P**	**2. P**	**2. P**
3. O	**3. O**	**3. O**	**3. O**
4. D	**4. D**	**4. D**	**4. D**
5. V**	5. V	5. V	5. V
6. N	6. M	6. K	6. M
7. K	7. L	7. M	7. L
8. L	8. R	8. L	8. R
9. R	9. Z	9. R	9. Z
10. Z	10. G	10. Z	10. G
11. G		11. G	

Tabela 1: PoS: rezultati (pod)korpusnih primerjav po metodi frekvenčnega profila.

* Kode pomenijo: S – samostalnik, P – pridevnik, O – okrajšava, D – predlog, V – veznik, M – medmet, K – števnik, L – členek, R – prislov, Z – zaimek, G – glagol.

** Črna barva – značilno za prvi (pod)korpus; siva barva (negativne vrednosti LL) – značilno za drugi (pod)korpus. Enako velja za Tabelo 2. Več o oznakah gl. v Erjavec et al. (2010) in na http://nl.ijs.si/jos/josMSD-sl.html (o označevalniku korpusa Kres gl. Grčar et al. (2012), o orodjih reldi-tokeniser[6] in reldi-tagger,[7] s katerima je bil označen korpus KAS, pa Ljubešić in Erjavec (2016)).

3. Analiza rezultatov

Podatke smo razvrstili po vrednosti LL. Zgornji del tabel tako prikazuje elemente, ki so bolj značilnosti za prvega od primerjanih (pod)korpusov, spodnji del z negativnimi vrednostmi LL pa elemente, ki so bolj značilni za drugega od primerjanih (pod)korpusov.

3.1. PoS

Ogled LL-vrednosti PoS-oznak je pokazal, da so seznami štirih izmed petih primerjav v vrhnjem delu po zaporedju povsem enaki (Tabela 1 v levem stolpcu besedila) in kažejo, da so za slovensko strokovno-znanstveno pisanje v primerjavi s Kresom ter njegovim leposlovnim delom bolj značilni samostalniki, pridevniki, okrajšave in predlogi (S – P – O – D). (O peti primerjavi, tj. primerjavi KAS-dipl : KAS-dr, gl. razdelek 3.1.1.)

Rezultati v Tabeli 1 potrjujejo izrazito samostalniškost strokovno-znanstvenih besedil, kar je skupaj s pridevnikom mogoče razumeti predvsem kot njihovo gosto terminološkost.[8] Tudi tretje mesto okrajšav ne preseneča, pri čemer je treba dodati, da prisotnost te »besedne vrste« v opazovanih besedilih poleg običajnih *oz.*, *npr.*, *angl.*, *t. i.*, *idr.*, *itd.*, *ipd.*, *tj.* in še nekaterih močno krepijo okrajšave v navedenkah ter seznamih virov in literature (zlasti okrajšave osebnih lastnih imen).

Predlogi so zadnja od štirih besednih vrst, katerih prisotnost je primerjalno večja v diplomah in doktoratih. Kot je razvidno na Sliki 1, se v večjem obsegu (s pojavitvami nad 20.000 v več kot 40-milijonskem podkorpusu doktoratov, če upoštevamo samo njihov slovenski del) pravzaprav rabi le 15 različnih predlogov, s tem, da je v izraziti prednosti predlog *v*. Vendar pa je ta slika – in skoraj enako je pri diplomah – zelo podobna Sliki 2, ki kaže pogostost predlogov v korpusu Kres. To pomeni, da se po *konkretnih* predlogih in njihovem medsebojnem razmerju doktorati ter diplome od splošne slovenščine bistveno ne razlikujejo. Odstopanje, ki ni veliko, a je dovoljšnje, da ga je v prid strokovno-znanstvenemu pisanju zaznala metoda frekvenčnega profila, gre torej pripisati večjemu skupnemu obsegu vseh predlogov (njihova pogostost na milijon pojavnic je npr. v KAS-dr 115.083, v Kresu pa 104.379).

Preostanek Tabele 1 kaže, da je po drugi strani v diplomah in doktoratih – če jih seveda še vedno primerjamo z leposlovjem in celotnim Kresom – izrazito najmanj glagolov, sledijo zaimki in prislovi, takoj za tem pa členki. Ker smo izpustili števnike (K), zapisane s števko, se ti v seznam, ki kaže razlike med diplomami in doktorati na eni strani ter Kresom in njegovim leposlovnim delom na drugi strani, sploh ne uvrstijo (gl. drugi in četrti stolpec Tabele 1) ali pa vsaj ne uvrstijo s pomembno razliko (gl. prvi in tretji stolpec).[9] Manj, vendar še vedno, so za leposlovje in Kres značilni še medmeti, primerjalno najmanj pa vezniki. Predlog in veznik sta torej besedni vrsti, pri katerih je med strokovno-znanstvenim pisanjem na eni strani ter drugimi

[4] Namesto izrazov *diplomsko delo* in *magistrsko delo drugostopenjskega bolonjskega študija* v nadaljnjem besedilu uporabljamo krajši izraz *diploma,* namesto *doktorska disertacija* pa *doktorat*.

[5] Analizo rezultatov razlik v lemah gl. v Logar in Erjavec (2017).

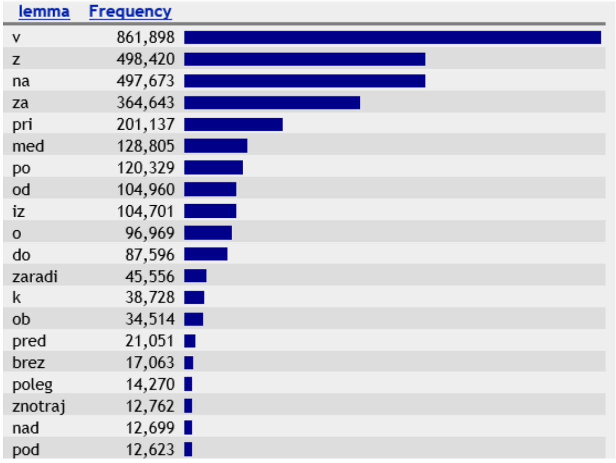
[6] https://github.com/clarinsi/reldi-tokeniser

[7] https://github.com/clarinsi/reldi-tagger

[8] Od vrhnjih 500 lem (tj. lem vseh besednih vrst) pri primerjavi KAS-dr : Kres-fict je bilo glagolnikov (samostalnikov iz glagolov s pomenom (tudi) dejanja, npr. *razumevanje*, *primerjava*, *meritev*) le 10 %, kar še dodatno potrjuje, da je značilnost besedil te zvrsti osredotočenost na statičnost in ne na delovanje oz. početje (o glagolih v zvezi s terminologijo gl. npr. Žele (2004)).

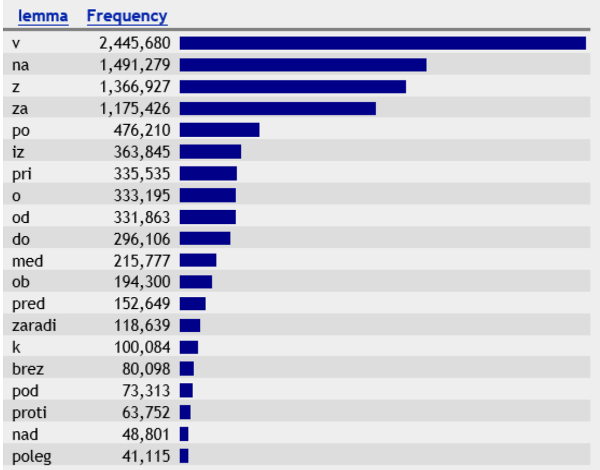
[9] Če bi števke vključili, pa bi se števnik kot besedna vrsta zelo verjetno pojavil v zgornjem, za KAS značilnem delu tabele.

zvrstmi besedil, ki so vse prisotne v Kresu, razlik najmanj (prim. tudi podobnost konkretnih veznikov po pogostosti v KAS-dr in Kresu na Slikah 3 in 4).



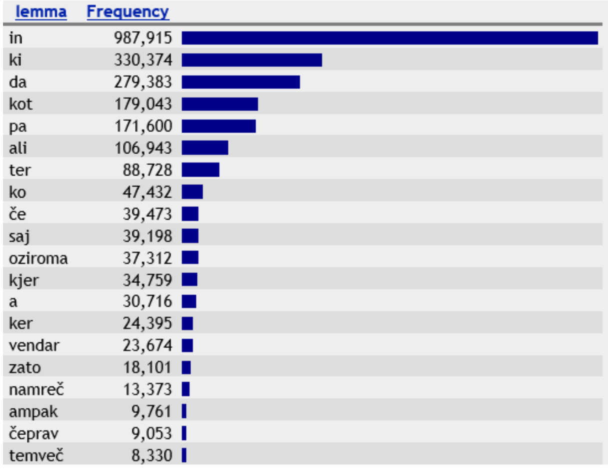
lemma	Frequency
v	861,898
z	498,420
na	497,673
za	364,643
pri	201,137
med	128,805
po	120,329
od	104,960
iz	104,701
o	96,969
do	87,596
zaradi	45,556
k	38,728
ob	34,514
pred	21,051
brez	17,063
poleg	14,270
znotraj	12,762
nad	12,699
pod	12,623

Slika 1: Predlogi po absolutni pogostosti v KAS-dr (vrhnji del).



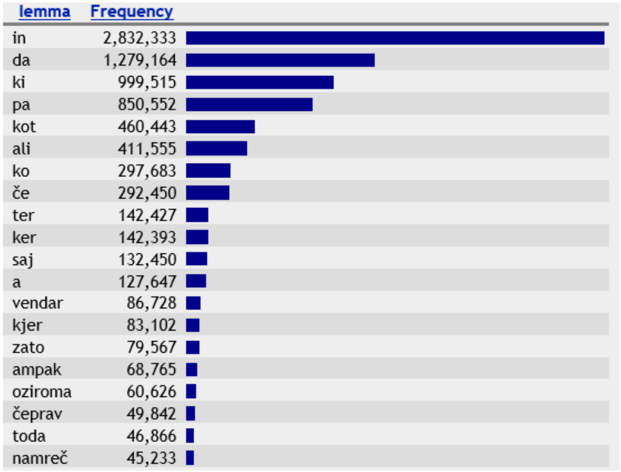
lemma	Frequency
v	2,445,680
na	1,491,279
z	1,366,927
za	1,175,426
po	476,210
iz	363,845
pri	335,535
o	333,195
od	331,863
do	296,106
med	215,777
ob	194,300
pred	152,649
zaradi	118,639
k	100,084
brez	80,098
pod	73,313
proti	63,752
nad	48,801
poleg	41,115

Slika 2: Predlogi po absolutni pogostosti v Kresu (vrhnji del).



lemma	Frequency
in	987,915
ki	330,374
da	279,383
kot	179,043
pa	171,600
ali	106,943
ter	88,728
ko	47,432
če	39,473
saj	39,198
oziroma	37,312
kjer	34,759
a	30,716
ker	24,395
vendar	23,674
zato	18,101
namreč	13,373
ampak	9,761
čeprav	9,053
temveč	8,330

Slika 3: Vezniki po absolutni pogostosti v KAS-dr (vrhnji del).



lemma	Frequency
in	2,832,333
da	1,279,164
ki	999,515
pa	850,552
kot	460,443
ali	411,555
ko	297,683
če	292,450
ter	142,427
ker	142,393
saj	132,450
a	127,647
vendar	86,728
kjer	83,102
zato	79,567
ampak	68,765
oziroma	60,626
čeprav	49,842
toda	46,866
namreč	45,233

Slika 4: Vezniki po absolutni pogostosti v Kresu (vrhnji del).

3.1.1. PoS pri KAS-dipl : KAS-dr

Kot smo že nakazali, smo pri primerjavi KAS-dipl : KAS-dr – pričakovano – dobili drugačne rezultate. Kot kaže Tabela 2, so za diplome bolj značilni glagoli, zaimki,[10] prislovi, členki in vezniki; za doktorate pa bolj samostalniki, sledijo okrajšave, pridevniki in nazadnje predlogi. Enako kot pri ostalih štirih primerjavah sta si torej obe besedilni vrsti najbolj podobni pri predlogih in veznikih.

PoS	LL	KAS-dipl (na mio pojavnic)	KAS-dr (na mio pojavnic)
1. G	44790	129.765	116.670
2. Z	39120	53.301	45.524
3. R	9186	43.705	40.249
4. L	5535	21.651	19.766
5. V	1746	87.997	85.838
6. D	–2838	111.949	115.083
7. P	–6511	149.127	154.613
8. O	–12345	8.220	10.048
9. S	–26472	384.693	402.489

Tabela 2: KAS-dipl : KAS-dr, PoS: rezultati primerjave po metodi frekvenčnega profila.

Rezultati te primerjave torej ožijo zgornjo ugotovitev: v primerjavi s splošnim jezikom je celotno strokovno-znanstveno pisanje (KAS-dr in KAS-dipl) bolj samostalniško-pridevniško (in najmanj glagolsko), obenem pa se v notranji primerjavi med doktorati in diplomami ta lastnost izkazuje za izrazitejšo pri prvih; z drugimi besedami: če diplome »zoperstavimo« doktoratom, se pokaže, da so diplome besednovrstno bližje splošnemu jeziku, kakršnega z besedili izkazujeta Kres in njegov leposlovni del.

3.2. MSD

Primerjava MSD-oznak je ob potrditvi zgornjih ugotovitev o besednovrstnih značilnostih besedil v KAS-u

[10] Pri tem je treba opozoriti, da je glagolski morfem *se* označen kot zaimek (gl. o tem tudi dalje).

in Kresu podala še uvid v razlike pri podrobnejših oblikoskladenjskih kategorijah. V nadaljevanju se bomo osredotočili predvsem na rezultate primerjave med KAS-dr in Kresom, katerih vrhnji del prikazuje Slika 5.[11]

Na Sliki 5 vidimo, da so na prvih petnajstih mestih po vrednosti LL, torej pri enotah, ki so najbolj značilne za KAS-dr v primerjavi s korpusom splošne slovenščine, predvsem samostalniške oblike (8/15), na celotnem seznamu 118 enot pa je sicer največ oblik pridevnika (64/118; samostalniških je 31/118). Rezultati nadalje pokažejo, da je na prvem mestu oznaka *Sozer*, torej 'samostalnik, občni, ženskega spola, edninski in v rodilniku' (v KAS-dr so najpogostejše tri take pojavnice: *vrednosti*, *uporabe*, *analize*); na drugem mestu je enak samostalnik, le da tokrat množinski (*storitev*, *informacij*, *sprememb*); na tretjem mestu pa 'občni samostalnik srednjega spola ednine in v rodilniku' (*leta*, *podjetja*, *dela*; več takih samostalnikov gl. na Sliki 6). Med pridevniki je na najvišjem, 7. mestu 'splošni (tj. nesvojilni in nedeležniški) pridevnik, ki ni stopnjevan, je ženskega spola, v množini in rodilniku' (daleč največji obseg ima pojavnica *človekovih*, sledijo *otrokovih*, *dušikovih* ipd.). Naslednji na seznamu ima enake lastnosti, le da ima namesto množinske edninsko obliko (*človekove*, *posameznikove*, *otrokove*) itd. Prvi glagol se pojavi na 11. mestu in ima oznako *Gp-spm-n*, ki označuje pomožnik *smo*. Zopet je visoko, tj. na 6. mestu, enorodna skupina okrajšav (*str.*, *oz.*, *npr.* ipd.).

Nadaljnje značilnosti MSD-profila doktoratov so še: pri samostalniških oblikah so najbolj značilni skloni imenovalnik, rodilnik, mestnik in orodnik, sicer pa z ženskim spolom kombinacije ednina + rodilnik, množina + rodilnik in ednina + imenovalnik; z moškim spolom ednina + imenovalnik in množina + rodilnik; s srednjim spolom pa ednina + rodilnik. Značilne pridevniške oblike so večinoma v osnovniku, v presežniku ni nobene, izmed vrst splošni/svojilni/deležniški pa ni nobenega svojilnega, deležniških je tretjina. Med zaimki sta najvišjo vrednost LL dobili oznaki *Zz-sei* (31. mesto) in *Zz-sem* (36. mesto), ki pomenita 'oziralni zaimek srednjega spola ednine v imenovalniku' oz. 'mestniku' (močno prevladujeta pojavnici *kar* in *čemer*). Od predlogov so na najvišjem, 16. mestu tisti, ki se vežejo z mestnikom (po pogostosti prevladuje predlog *v*, kar smo videli že pri analizi PoS-oznak, sledita *na*, *pri* idr.), na 29. mestu je še predlog, ki se veže z orodnikom (v veliki večini *s/z*, sledita *med* in *pred*), pri glagolih pa ob pomožniku višjo vrednost LL (25. mesto) kaže še 'nedovršnik v sedanjiku tretje osebe množine' (*vplivajo*, *predstavljajo*, *kažejo* itd.). Le nekoliko izstopajo še priredni vezniki (s skoraj milijonom pojavitev je v KAS-dr prevladujoč veznik *in*, sledi mu veznik *pa* s 170 tisoč pojavitvami, nato veznik *ali* s 100 tisoč pojavitvami idr.). Prvi števniki se uvrstijo na 70. mesto, in sicer gre za števnike, zapisane z rimsko številko.[12]

Ostale tri primerjave MSD-oznak so dale zelo podobne rezultate.

3.2.1. MSD pri KAS-dipl : KAS-dr

Primerjava MSD-oznak med KAS-dipl : KAS-dr je dala najkrajši seznam enot (115 v primerjavi z npr. KAS-dr : Kres, ki jih ima 357), kar kaže na oblikoskladenjsko podobnost teh dveh besedilnih vrst. Rezultati tu pričakovano ponavljajo večjo glagolskost diplom (med njimi pojavnic, kot so: *pride*, *začne*, *postane*; *doseči*, *zagotoviti*, *ugotoviti*; *postala*, *začela*, *pokazala*; *imeti*, *uporabljati*; *pojavijo*, *dobijo*, *postanejo*). V tem smislu izstopata tudi pomožnik za prvo osebo ednine (*sem*) in tretjo osebo ednine v prihodnjiku (*bo*), seveda pa tudi zaimek (dejansko pa glagolski morfem) *se*. Na drugi strani so doktorati, kot že rečeno, bolj samostalniško-pridevniški. Tudi tu nobena v primerjavi izstopajoča pridevniška oblika ni v presežniku. Izmed redkih za doktorate značilnih glagolskih oblik so na seznamu 'pomožni glagol v prvi osebi množine v trdilni' in 'nikalni obliki' (*smo*, *nismo*) in 'pomožni glagol v dvojini' (*sta*), dalje pa še 'dovršni glagoli v obliki deležnika, v množini in moškega spola' (npr. *uporabili*, *ugotovili*, *izvedli*), 'dovršni glagoli v sedanjiku, prvi osebi in množini' (*dobimo*, *uporabimo*, *najdemo*), 'glagoli v obliki deležnika, v množini in moškega spola' (*upoštevali*, *analizirali*, *podali*) ter še 'nedovršni glagoli v sedanjiku, tretji osebi in dvojini' (*predstavljata*, *ugotavljata*, *navajata*).

Oblikoskladenjske kategorije skupaj z njihovimi najpogostejšimi zapolnitvami torej izkazujejo tipičnost določenih lem v določenih oblikah, pri čemer gre zlasti za obče samostalnike vseh treh spolov v ednini ali množini ter v rodilniku in imenovalniku.

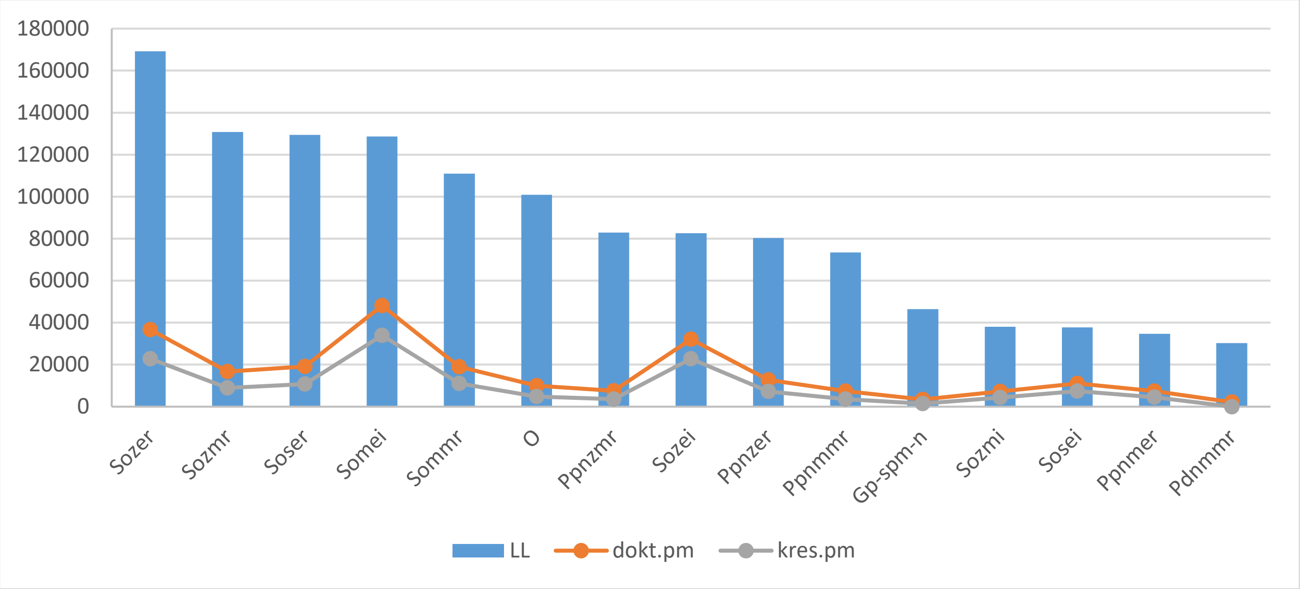
4. Iz podatkov v opis

Glavne ugotovitve naše besednovrstne analize so potrdile, da je za opis slovenskih strokovno-znanstvenih besedil še vedno pomembna njihova »samostalniškost« (Žagar Karer, 2011: 145) oz. »neglagolsko izražanje« (Toporišič, 1991: 23). A kot je pokazal frekvenčni profil lem, ki je bil prav tako pridobljen v tej raziskavi (Logar in Erjavec, 2017), se predmetnopoimenovalna zgoščenost strokovno-znanstvenih besedil, ki gre očitno »na škodo« glagolov, ne dosega le s termini, ki so večinoma pridevniško-samostalniški (Logar et al., 2013), temveč tudi s samostalniki, ki so splošnostrokovni (ter seveda njihovih tipičnim – predvsem spet pridevniškim – besedilnim okoljem). Taki splošnostrokovni samostalniki so npr. povezani z zgradbo strokovno-znanstvenih besedil (*slika*, *tabela*, *graf*, *primer*, *literatura*, *priloga*, *podpoglavje* itd.) ali predstavitvijo in interpretacijo rezultatov (npr. *podatek*, *število*, *izračun*, *ugotovitev*, *posledica*, *interpretacija*, *mnenje*).[13] Prav ustrezna raba takih izrazov je poleg poznavanja specializiranih področnih poimenovanj ključna za natančno, jasno in razumljivo pisanje besedil, katerih osrednja funkcija je informativno-spoznavna (Skubic, 1994/95).

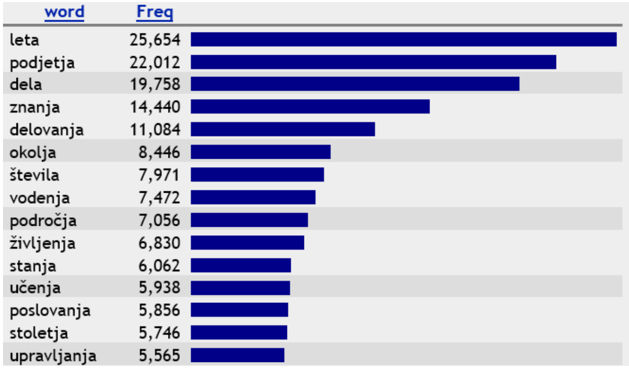
[11] Za lažjo predstavo o metodi smo tu k oznakam dodali še podatke o njihovi relativni pogostosti v obeh korpusih. Vidi se, da so razmerja med višinami stolpcev (LL) drugačna, kot so – medsebojno sicer zelo podobna – razmerja med relativnim številom pojavnic z določeno MSD-oznako (torej višino pik).

[12] A kot smo že opozorili, rezultatov pri števniku zaradi izpusta te besedne vrste v primerih, ko je šlo za zapis z arabsko števko, ne moremo upoštevati.

[13] Izmed lem (samostalniki, pridevniki, glagoli, prislovi, členki in okrajšave), ki so se pokazale kot značilne za KAS-dr, ko smo ta podkorpus primerjali s Kres-fict, smo jih kar 30 % prepoznali kot splošnostrokovnih.



Slika 5: KAS-dr : Kres, MDS: rezultati primerjave po metodi frekvenčnega profila (vrhnji del) in relativne vrednosti njihovih pojavitev (na milijon pojavnic).



word	Freq
leta	25,654
podjetja	22,012
dela	19,758
znanja	14,440
delovanja	11,084
okolja	8,446
števila	7,971
vodenja	7,472
področja	7,056
življenja	6,830
stanja	6,062
učenja	5,938
poslovanja	5,856
stoletja	5,746
upravljanja	5,565

Slika 6: KAS-dr: pojavnice z oznako *Soser* (vrhnji del).

Frekvenčni profil MSD-oznak doktoratov in diplom je še podrobneje pokazal tipičnost določenih sklonov ter števil, pri glagolih pa npr. izstopanje sedanjika nedovršnih glagolov in množinskega dovršnega deležnika moškega spola. Dalje smo pri obeh besednih vrstah, ki sicer kažeta največjo sorodnost strokovno-znanstvenega pisanja s splošnim jezikom, ugotovili, da so v doktoratih izrazito pogosti predlogi, ki se vežejo z mestnikom (zlasti *v*), in priredni vezniki (zlasti *in*). Tudi okrajšave (prim. Kompara, 2011) so primerjalno izstopale, zaradi česar bi bilo smiselno še podrobneje pogledati njihovo raznovrstnost, razlog za nastanek in položaj v besedilu.

Raba glagolov, ki so se na drugi strani v celoti izkazali kot značilni za splošni in leposlovni jezik ter bolj za diplome kot za doktorate, je v znanstvenem pisanju omejena, kar pomeni, da ne izkazuje bogate obsegovne in pomenske razpršenosti, značilne npr. za nekatere novinarske žanre ali umetnostno prozo. Prav to oženje v specifično in natančno poimenovanje dejanj (*analizirati*, *meriti*, *ugotavljati* itd.) ter v prevladujočo rabo določenih oblik (gl. npr. podatke o rabi trpnika v Logar et al., 2016), je obenem razlog, da je tudi opis glagolske rabe v strokovno-znanstvenem pisanju nujen, še zlasti če bo tak opis namenjen bralcem, ki se v tej podzvrsti šele opismenjujejo.[14]

Metoda frekvenčnega profila nam je torej dala podatke, katerih prvi izsledki že izpostavljajo besednovrstna in oblikoskladenjska mesta, ki bi bila pri pripravi prihodnjega celovitejšega opisa strokovno-znanstvene slovenščine vredna pozornosti. Na ta način naši rezultati z vidika izrazito slovničnih kategorij relevantno dopolnjujejo druge podatke, ki jih lahko prav tako pridobimo iz korpusa KAS (npr. podatke o kolokacijskem okolju za strokovno-znanstveno pisanje značilne neterminološke leksike v orodju Sketch Engine (Kilgarriff et al., 2004)). Metoda frekvenčnega profila je sicer primerna tudi za primerjavo manjših (pod)korpusov, zato bi jo bilo mogoče uporabiti tudi na drugačnih podkorpusih KAS-a, pri čemer imamo v mislih predvsem vzorčene področne podkorpuse (npr. podkorpus humanističnih in tehničnih ved, lahko pa tudi posameznih strok znotraj njih). Tak pristop bi pokazal tudi, kolikšen vpliv je imelo na naše tukajšnje rezultate dejstvo, da je KAS tako v diplomskem kot v doktorskem delu večinoma družboslovan in zelo malo humanističen (prim. Erjavec et al., 2016). Brez dodatnih analiz lahko ta hip ocenimo le, da je bil vpliv pomemben.

5. Sklep

Metodo frekvenčnega profila smo na slovenskih korpusih tokrat uporabili že tretjič, zato smo okvirno vedeli, kakšne rezultate lahko pričakujemo. Primerjave smo zastavili na štirih (pod)korpusih, kar nam je omogočilo, da smo (a) ocenili, kateri rezultati so za naš cilj najboljši, ter (b) da smo lahko prečno preverjali, koliko so rezultati prekrivni in torej relevantni.

[14] Podrobneje bi veljalo pogledati še prislove in členke (prim. Mikolič, 2005), pa tudi zaimke (prim. Gorjanc, 1998) in števnike, ki se jim tu nismo posvetili.

Rezultati primerjav PoS- in MSD-oznak v podkorpusih diplom in doktoratov med seboj – ter s Kresom in leposlovjem na drugi strani – so nam omogočili prvi tako celovit uvid v izrazito oblikoslovno-skladenjske značilnosti strokovno-znanstvenega pisanja pri nas. V nadaljevanju pa bo vsekakor treba tukajšnje dokaj kratko interpretiranje rezultatov dopolniti še z leksikalnimi zapolnitvami (kjer se izkazujejo za tipične) in vpogledom v razloge za večja odstopanja; načrtujemo pa tudi nadgradnjo z leksikalno-skladenjskimi informacijami iz izluščeni n-gramov (Dobrovoljc, 2016).

Zahvala

Avtorja se zahvaljujeta anonimnim recenzentom za koristne pripombe. Raziskavo, opisano v prispevku, je podprl projekt ARRS J6-7094 *Slovenska znanstvena besedila: viri in opis*.

6. Literatura

Douglas Biber in Federica Barbieri. 2007. Lexical Bundles in University Spoken and Written Registers. *English for Specific Purposes*, 26(3): 263–286.

Kaja Dobrovoljc. 2016. *Korpus KAS: n-grami (interno gradivo)*. Ljubljana, Trojina, zavod za uporabno slovenistiko; Filozofska fakulteta UL; Fakulteta za družbene vede UL.

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

Tomaž Erjavec, Nikola Ljubešić in Nataša Logar. 2015. The slWaC Corpus of the Slovene Web. *Informatica*, 39(1): 35–42.

Dee Gardner in Mark Davies. 2013. A New Academic Vocabulary List. *Applied Linguistics*, 35(3): 305–327.

Vojko Gorjanc. 1998. Konektorji v slovničnem opisu znanstvenega besedila. *Slavistična revija*, 46(4): 367–388.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: T. Erjavec in J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*, str. 89–94. Ljubljana, Institut »Jožef Stefan«.

Ken Hyland. 2008. As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes*, 27(1): 4–21.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz in David Tugwell. 2004. The Sketch Engine. V: *Proceedings of the 11th EURALEX international congress*, str. 105–116. Lorient, Universite de Bretagne-Sud.

Mojca Kompara. 2011. *Slovar krajšav*. Kamnik: Amebis.

Nikola Ljubešić in Tomaž Erjavec. 2016. Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. V: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA).

Nataša Logar Berginc, 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.

Nataša Logar, Špela Arhar Holdt in Tomaž Erjavec. 2016. Slovenski strokovni jezik: korpusni opis trpnika. V: E. Kržišnik in M. Hladnik (ur.): *Toporišičeva obdobja*, str. 237–245. Ljubljana, Znanstvena založba Filozofske fakultete.

Nataša Logar in Tomaž Erjavec. 2017. Slovene Academic Writing: A Corpus Approach to Lexical Analysis. V: *Interdisciplinary Knowledge-making, Challenges for LSP Research: Book of Abstracts*, str. 44. Bergen, Norwegian School of Economics. (Celotni prispevek je oddan v objavo.)

Nataša Logar, Špela Vintar in Špela Arhar Holdt. 2013. Terminologija odnosov z javnostmi: korpus – luščenje – terminološka podatkovna zbirka. *Slovenščina 2.0*, 1(2): 113–138.

Nataša Logar. 2017. *Strokovno-znanstveni jezik: študijska literatura in priročniki v Sloveniji ter kratek pregled tujih praks*. Ljubljana, Fakulteta za družbene vede, http://nl.ijs.si/kas/wp-content/uploads/2018/03/KAS-pregled-prirocnikov-navodil-in-predmetov-Logar.pdf.

Vesna Mikolič. 2005. Izrazi moči argumenta v znanstvenih besedilih. V: M. Jesenšek (ur.): *Knjižno in narečno besedoslovje slovenskega jezika*, str. 278–291. Maribor, Slavistično društvo Maribor.

Paul Rayson in Roger Garside. 2000. Comparing Corpora Using Frequency Profiling. V: *Proceedings of the Workshop on Comparing Corpora*, str. 1–6. Hong Kong, Association for Computational Linguistics.

Ute Römer. 2009. English in Academia: Does Nativeness Matter? *Anglistik: International Journal of English Studies*, 20(2): 89–100.

Andrej Skubic. 1994/95. Klasifikacija funkcijske zvrstnosti in pragmatična definicija funkcije. *Jezik in slovstvo*, 40/5: 155–168.

Richard Stevenson. 2010. *Advanced Grammar: For Academic Writing*. Morisville, Academic English Publications.

John M. Swales in Christine B. Feak. 2012 *Academic Writing for Graduate Students: Essential Tasks and Skills*. Michigan, The University of Michigan.

Jože Toporišič. 1991. *Slovenska slovnica*. Maribor, Obzorja.

Mojca Žagar Karer. 2011. *Terminologija med slovarjem in besedilom: analiza elektrotehniške terminologije.* Ljubljana, Založba ZRC, ZRC SAZU.

Andreja Žele. 2004. Stopnje terminologizacije v leksiki (na primerih glagolov). V: M. Humar: *Terminologija v času globalizacije*, str. 77–91. Ljubljana, Založba ZRC, ZRC SAZU.

Word Selection in the Slovenian Sentence Matrix Test for Speech Audiometry

Tatjana Marvin,[#] Jure Derganc,* Samo Beguš,[†] Saba Battelino‡

[#] Department of Comparative and General Linguistics, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana
tatjana.marvin@ff.uni-lj.si
* Institute of Biophysics, Faculty of Medicine, University of Ljubljana
Vrazov trg 2, SI-1000 Ljubljana
jure.derganc@mf.uni-lj.si
[†] Laboratory of Metrology and Quality, Faculty of Electrical Engineering, University of Ljubljana
Tržaška 25, SI-1000 Ljubljana
samo.begus@fe.uni-lj.si
‡ Department of Otorhinolaryngology, Faculty of Medicine, University of Ljubljana, University Medical Centre Ljubljana
Zaloška cesta 2, SI-1104 Ljubljana
saba.battelino@kclj.si

Abstract

In this paper we present the word selection process in the Slovenian matrix sentence test for speech intelligibility measurements. We focus on the phonemic distribution in the test, which should be approximated as closely as possible to the distribution in the language. We establish phonemic distribution for Slovenian by combining the orthographic distribution in the corpus ccKress and the phonetic distribution in Mihelič (2006). As a result, a phonemically balanced matrix test is proposed for Slovenian.

1. Background and Research Goals

Speech audiometry is one of the standard methods used to diagnose the type of hearing loss and to assess the communication function of the patient by determining the level of the patient's ability to understand and repeat words or sentences presented to him or her in a hearing test. For this purpose, the adaptation of the Freiburg Monosyllabic Word Test and the Freiburg Number Test are used in Slovenia. The Slovenian version was developed in Pompe (1968) and was then revised by Marvin et al. (2016).

While word tests are important diagnostic tools, sentence tests better reflect everyday communication and have proven to be highly useful and precise measurement tools in many languages. In general, two types of such tests are used; those using meaningful, everyday sentences with a variable grammatical structure (e.g. Plomp & Mimpen, 1979 and subsequent work) and sentence tests with a matrix structure, in which the syntax is fixed, but the combination of words is unpredictable (Hagerman, 1982; Wagener, 1999a, b, c; Ozimek et al., 2010; Hochmuth et al., 2012; Warzybok et al., 2015 among others). At present, there are no standard sentence tests of any kind available for Slovenian.

In this paper we present the word selection process for the sentence test with a matrix structure that we develop for Slovenian. The matrix test will be used for a more accurate assessment of hearing in people with a hearing disorder, for assessing the understanding of speech in people with central hearing disorders and comprehension disorders, for assessing cognitive abilities, for assessing the improvement of speech comprehension in patients using various removable and implanted mechanical and electronic hearing aids and in patients with disturbing tinnitus (hearing sounds in the ears or the head without any real sound inside or outside the body, Jagoda et al. 2018)). In creating the test, we follow the guidelines by International Collegium of Rehabilitative Audiology (ICRA), (Akeroyd et al., 2015). The guidelines complement the standard ISO 8253-3:2012 (Acoustics - Audiometric test methods - Part 3: Speech Audiometry) by providing the necessary steps needed to create the matrix test in any given language. The ICRA guidelines require that the matrix test should approximate the phonemic distribution of the underlying language as closely as possible. To our knowledge, the phonemic distribution of Slovenian has not been thoroughly analysed, so we derive it from the data on the letter distribution based on the corpus ccKres (see Erjavec and Logar Berginc, 2012; Logar Berginc and Krek, 2012; Logar Berginc et al., 2012 for more information on the corpus) in combination with the data on the phonetic distribution that is available in Mihelič (2006).

The paper is organized as follows. In Section 2 we describe the general construction of the test, in Section 3 we focus on the part in the test that involves word selection for the sentence construction. In Section 4 we present the word selection for the Slovenian matrix test.

2. General Guidelines for Matrix Test Construction

The matrix test was originally proposed for Swedish by Hagerman (1982). A modified version (Wagener et al., 1999a, b, c) is currently available in 14 languages (e.g. English, Dutch, German, French, Turkish and others), among them only in two Slavic languages (Polish and Russian). The test consists of five-word long sentences, each of which has the same syntax of the form Name-Verb-Numeral-Adjective-Noun, but whose semantic content is unpredictable (e.g. "Thomas wins eight red shoes"). The base matrix consists of 50 words, 10 for each of the five word positions. Each sentence is a random walk through the matrix and sentences are further grouped into test lists of ten sentences in such a way that each list contains exactly one appearance of each of the 50 words in the base matrix.

The sentences are then recorded in such a way that all combinations of two consecutive words are included (this requires the recording of at least 100 sentences).[1] The recorded sentences are cut into single words by preserving the coarticulation at the end of the cut word to the required consecutive word, but truncating the coarticulation at the word beginning. The test sentences are then resythesized by combining words with appropriate transitions and supplied with masking noise. Finally, the test protocol includes the optimization and evaluation of the recorded material. The individual's speech recognition threshold is determined by an adaptive tracking procedure using one, two or three whole test lists.

3. Slovenian Matrix Test

3.1. Sentence Structure and Word Selection

In this section we describe the process of gathering the linguistic material in the base matrix, which consists of 50 words, 10 for each of the five word positions in the sentence of the form Name-Verb-Numeral-Adjective-Noun. The selected material has to fulfil several criteria.

To begin with, the test should contain five female and five male names. Next, only highly frequent words should be chosen; we establish that by referring to the GigaFida language corpus for Slovenian (see Erjavec and Logar Berginc, 2012; Logar Berginc and Krek, 2012; Logar Berginc et al., 2012 for more information on the corpus). We make sure that words and the combinations of words that are potentially offensive are not included in this list. Also, certain repetitive combinations (e.g. *veliko velikih kamnov* "many big stones") or similar names (*Jana*, *Jasna*) are avoided.

Next, all possible sentences that can be assembled by combining the words in the matrix have to be grammatically correct and semantically unpredictable. In matrix tests for Germanic languages, the past tense forms are generally used, while in the existing matrix tests for Slavic languages the present and the future tense forms are used. Using a verb in the past or future tense in Slovenian would require a special syntactic position for the copula, which is not included in the standard matrix test. Also, in the past tense forms or the future tense forms in Slovenian, the copula *biti* "be" is marked for number and person, while the *l*-participle is marked for gender and number. The use of these two tenses would thus result into a great number of ungrammatical combinations in cases where the gender of the participle does not agree with the gender of the subject (e.g. "*Jana je *kupil* tri velike škatle"). Therefore, only verbs in the present tense can be chosen to fill the verb position in the Slovenian matrix test (e.g. "Jana *kupi* tri velike škatle").

The selection of numerals has to be adapted to the properties of the Slovenian language. Only the numerals from five on are used in the test, as these uniformly require the following adjective and noun in the genitive plural form (e.g. "Jana kupi pet/šest/sedem/osem... *velikih škatel*"). The numerals from 1-4 are replaced by quantifier expressions that require the genitive plural form of the following adjective and noun, such as *malo* "few", *nekaj* "some", etc. (e.g. *Jana kupi malo/nekaj velikih škatel* "Jana buys few/some big boxes").

The number of syllables within each word group has to be balanced; we decide to select disyllabic words and only exceptionally monosyllabic or trisyllabic words (e.g. for reasons relating to phonemic balance).

Finally, a requirement for matrix tests is that the phonemic distribution of the underlying language should be approximated as closely as possible by the matrix test. As the phonemic distribution of Slovenian has not been established, the requirement in question demands our special attention and is dealt with in detail in Sections 3.2. to 3.4.

3.2. Relations Between Phonemes, Allophones and Letters in Slovenian

Before turning to the issue of phonemic balance, we briefly explain the notions of phoneme and allophone and their relation to the letters in the Slovenian alphabet. A phoneme is standardly defined as the smallest sound unit that can be segmented from the acoustic flow of speech and which functions as a semantically distinctive unit. If a sound unit is replaced by another sound unit in a word and the two words have a different meaning, we classify the two differentiating sounds as phonemes, e.g. in the English pair *pet* – *bet*, /p/ and /b/ are phonemes. Phonemes are abstract units, each phoneme representing a class of phonetically similar sound variants that are called allophones. An allophone is standardly defined as a concretely realized variant of a phoneme and is dependent on the phonological environment. For example, in English, the phoneme /p/ has an aspirated variant [p^h] at the beginning of the syllable (as in *pet*), but a non-aspirated variant [p] elsewhere (e.g. *loop*). As a phoneme in a particular language has at least one concrete realization, the number of allophones in languages is usually higher than the number of phonemes. We use slashes for transcribing phonemes (phonemic transcription) and square brackets for transcribing allophones (phonetic transcription).

The writing systems that use letters can be organized in different ways – some of them tend to use a letter to denote a phoneme, others are closer to using a letter for an allophone. In Slovenian, the tendency is for one letter to represent one phoneme. For example, the letter "n" stands for the phoneme /n/, which has three allophones: [N] when followed by a velar consonant as in *Anglija* "England" ; [n'] (for some speakers) when followed by [j#] or [jC] as in *konj* "horse", *konjski* "horse-adj" and [n] elsewhere, e.g. *nos* "nose".[2] Despite being phonetically different, all three concrete variants are denoted by the same letter "n".

[1] The recording for the Slovenian test will be carried out in an anechoic chamber with a noise level below 15 dB(A) using the RØDE NT2000 microphone and RME Babyface Pro external soundcard at a sampling rate of 44.1 kHz.

[2] In this paper we use the machine-readable alphabet MRPA, as in Dobrišek et al. (2002). The symbol "C" is used for "consonant", while the symbol "#" marks the word boundary.

Nevertheless, in Slovenian, there are fewer letters than phonemes – 25 letters versus 29 phonemes, following Toporišič (2000). As a consequence, the distribution of phonemes cannot be established directly from the letter distribution. There are several reasons for there being more phonemes than letters. In some cases, a single letter can stand for more than one phoneme, e.g. the letter "e" can denote /e/ in *led* "ice", /E/ in *žep* "pocket" or /@/ in *pes* "dog". Similarly, the letter "o" can denote /o/ in *nos* "nose" or /O/ in *noga* "leg".[3] The phoneme /dZ/ is not expressed in writing by a single letter, but by using the two-letter combination "dž" (e.g. džip "jeep"). In addition, for the phoneme /@/ no letter is used in some instances, e.g. in many words that contain the consonant /r/ such as *vrt* "garden", *smrt* "death", etc. The same is true of the phoneme /j/, which is not expressed in writing in some combinations (e.g. *pacient* "patient" /pacijent/), but expressed in writing, though not pronounced in other cases, e.g. "nj#" is pronounced either as [n'] or [n] (depending on the speaker), the letter "j" only indicating the fact that the variant of /n/ is palatalized (with the speakers that pronounce the combination as [n']).[4]

3.3. Choosing the reference corpus for Slovenian phoneme distribution

To find a suitable reference corpus for Slovenian phoneme distribution, we refer to CLARIN.SI repository, and examine two corpora of spoken and one corpus of written Slovenian: a) the corpus of spoken Slovenian GOS (its orthographic transcription in standard Slovenian), which contains 1 million words, b) the orthographic transcription of the database SNABI Slovenian Studio Quality Speech Corpus, more precisely its subpart Lingua consisting of 910 sentences taken from different styles of text, such as books and newspapers (Kačič et al. 2002), and c) the corpus of written Slovenian ccKres, which contains 10 million words of different types of texts – from daily newspapers, magazines, books (fiction, non-fiction, textbooks), web pages – and has a balanced genre structure. We then calculate the frequencies of letters in the three corpora and compare them to the seminal work of Jakopin (1999), which analysed a number of literary works in Slovenian. The results are presented in Table 1.

[3] In this work, vowel lenght and stress are not taken into consideration.

[4] Jurgec (2011) proposes that Slovenian has nine vowels and not eight as traditionally assumed. The additional vowel is the low central tense vowel [V] (e.g. in the words *čas* "time", *brat* "brother", etc.). In this paper, we follow the traditional classification as in Toporišič (2000).

	a	b	c	č	d	e	f	g	h	i	j	k	l	m	n	o	p	r	s	š	t	u	v	z	ž
GOS	12.0	1.7	0.6	1.3	3.5	12.5	0.2	1.2	0.9	8.0	4.9	3.9	4.1	4.2	5.9	8.8	3.6	4.5	4.5	1.2	5.2	1.6	3.3	1.9	0.4
ccKres	10.4	1.8	0.9	1.4	3.5	10.2	0.2	1.5	1.1	9.0	4.3	3.7	4.6	3.1	6.9	9.3	3.5	5.3	4.8	1.0	4.6	2.0	4.1	2.2	0.6
Lingua	9.7	1.9	0.7	1.6	3.5	10.9	0.2	1.4	1.2	8.8	4.8	3.6	4.9	3.5	6.3	9.2	3.6	5.1	4.8	1.2	4.5	1.9	4.2	2.2	0.6
Jakopin	10.5	1.9	0.7	1.5	3.4	10.7	0.1	1.6	1.1	9.0	4.7	3.7	5.3	3.3	6.3	9.1	3.4	5.0	5.1	1.0	4.3	1.9	3.8	2.1	0.7

Table 1: Letter frequencies (in %) in three reference corpora and Jakopin (1999)

We find that the letter frequencies in the three corpora agree within approximately 10% (except for the less frequent letters, where the variations are larger), yet GOS has a relatively high proportion of letters “a”, “e” and “m”, possibly because they are used as fillers in spoken Slovenian. Based on the corpus size and the letter distribution we adopt ccKres as the basis for establishing the phonemic balance.

3.4. Establishing Phoneme distribution

To our knowledge, the distribution of Slovenian phonemes has not been thoroughly analysed, which is understandable given the fact that a phoneme is an abstract unit that appears more often in (theoretical) linguistic research, while in the work on corpus linguistics or applied phonetics we usually find analyses based on orthographic or phonetic transcription. To obtain the phonemic distribution, we take the orthographic data, i.e. the letter distribution, as our basis and supplement it with the distribution of particular phonemes in cases where these are not directly evident from the letters (for phonemes /e/, /E/, /@/, /o/, /O/, /dZ/, /j/, see Section 3.2.).[5] This is done by adding the phonemes that are missing in the orthographic transcription (/@/, /j//), subtracting the number of phonemes that are not pronounced (/j/) and by referring to the ratios of the phonemes in the corpus that contains a phonetic transcription (/e/, /E/, /@/, /o/, /O/). For the latter, we refer to the distribution as established in Mihelič (2006), where 300.000 phonetically transcribed sentences are analysed in terms of allophone distribution.[6]

The procedure is described in detail in the following points:

1) All letters in the corpus ccKres are transformed into lower case characters. Next, the standard Slovenian diacritic marks on the letters “a”, “e” and “o” are discarded ("á"→"a", "à"→"a", "é"→"e", "ê"→"e" ,"è"→"e" ,"ô"→"o" ,"ó"→"o"). Finally, all the characters that are not in the standard Slovenian alphabet (except for “đ”) are discarded from the corpus.
2) The number of phonemes /dZ/ is determined by counting the total occurrences of “dž” and “đ”.
3) The number of phonemes /j/ is adjusted by adding the occurrences that are pronounced, but not expressed in writing between the two vowels in the following combinations: "ia", "ie", "io", "ea", "oi". The number of phonemes /j/ is reduced in the instances where the latter is found is spelling, but is not pronounced: nj#, njC, lj#, ljC.
4) The number of phonemes /o/ and /O/ is determined by dividing the number of letters “o” according to the distribution in Mihelič (2006): /o/ (21 % of letter "o" occurrences), /O/ (79 % of letter "o" occurrences).
5) The number of phonemes /e/, /E/ and /@/ is determined by first summing the number of letters “e” plus the number of occurrences of /@/ that are not expressed in writing. According to Toporišič (2000), the phoneme /@/ can be found in combinations with “CrC”, “Cr#”, “#rC”, “vn#”, “jn#”, “ln#”, “lm#”, “jm#”, “lmN”, “jmN”, “jnN”, “lnN”, “vnN”, where “C” stands for any consonant, “N” for any obstruent, and “#” for a word boundary. The total count of these occurrences is divided into the phoneme counts according to the distribution of these three phonemes in Mihelič (2006): /e/ (25 %), /E/ (66 %) and /@/ (9 %).

The proposed phonemic distribution, on which the matrix test for Slovenian is based, is presented in Section 4 in Table 3 and Figure 1 (phonemes and their percentage of occurrence).

4. Results: Matrix Test for Slovenian

The proposal for the Slovenian matrix test, based on the criteria from Section 3, is presented in Table 2.

[5] In principle it would be possible to arrive at the phonemic transcription on the basis on the phonetic transcription. However, in several cases, the allophones of different phonemes overlap and thus make it impossible to obtain a precise phonemic transcription without referring to orthography. For example, the phonemes /l/ and /v/ in the final position are pronounced in the same way. The words *pil* "drink-participle" and *piv* "beer-plural.genitive" share the phonetic transcription /piU/, but are different in terms of phonemic transcription, /pil/ and /piv/, respectively.

[6] Lingua also contains a phonetic transcription that could serve as a basis for establishing the ratios concerning the vowels in question. It is, however, a much smaller corpus in comparison to Mihelič's database (910 versus 300.000 sentences). The ratio concerning "o" is very similar to the one in Mihelič (2006): 75% of letters "o" correspond to the phoneme /O/ and 25% to /o/ (79% vs. 21% in Mihelič (2006)). The ratios for the letter "e" are /E/ (51%), /e/ (45%) and /@/ (4%) and differ considerably to the ones in Mihelič (2006) (66% vs. 25% vs. 9%). We believe that one of the causes for the differences lies in the fact that the pronunciation in Mihelič's corpus relies on the standard, while Lingua contains the transcription of the colloquial speech, mostly the variant from the Štajerska region.

Name	Verb	Numeral	Adjective	Noun
Gregor	kupi (buys)	pet (five)	velikih (big)	stolov (chairs)
Tone	dobi (gets)	šest (six)	lepih (beautiful)	copat (slippers)
Jure	najde (finds)	sedem (seven)	novih (new)	škatel (boxes)
Urban	skrije (hides)	osem (eight)	čudnih (strange)	avtov (cars)
Sašo	vzame (takes)	enajst (eleven)	starih (old)	zvezkov (notebooks)
Branka	ima (has)	sto (hundred)	dobrih (good)	koles (bicycles)
Jana	pelje (conveys)	tristo (three hundred)	dragih (expensive)	kamnov (stones)
Nada	nese (carries)	tisoč (thousand)	modrih (blue)	majic (T-shirts)
Lara	proda (sells)	nekaj (some)	rumenih (yellow)	loncev (pots)
Petra	išče (looks for)	malo (few)	zelenih (green)	nožev (knives)

Table 2: The proposed fifty-word matrix for the Slovenian Matrix Test.

The phonemic distribution in ccKres (as established in the previous section) in comparison to the phonemic distribution in the Slovenian fifty-word matrix is presented in Table 3 and Figure 1.

phoneme	test	ccKres
a	9,88	10.46
i	7,51	9.04
O	7,11	7.37
E	6,72	7.08
n	6,32	6.95
r	5,93	5.36
s	5,14	4.79
t	5,53	4.64
l	3,95	4.63
v	4,35	4.17
k	3,56	3.68
d	3,56	3.51
p	2,77	3.48
j	3,16	3.28
m	3,56	3.12
e	3,56	2.68
z	1,58	2.20
u	1,98	2.04
o	1,58	1.96
b	1,58	1.84
g	1,19	1.50
tS	1,19	1.40
x	3,95	1.11
S	1,58	0.97
@	1,19	0.97
c	1,19	0.90
Z	0,40	0.61
f	0,00	0.24
dZ	0,00	0.01

Table 3: Phoneme frequencies (in %) in the proposed Slovenian Matrix Test and in the corpus ccKres.

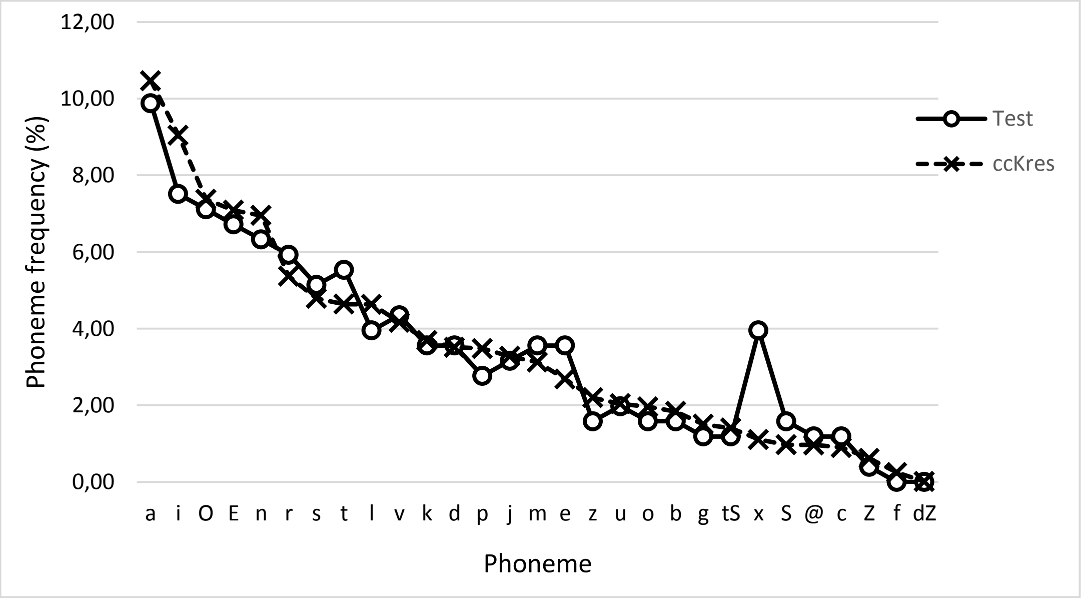


Figure 1: Phoneme frequencies (in %) in the proposed Slovenian Matrix Test and in the corpus ccKres.

The phonemic balance achieved in the Slovenian matrix test is comparable to the one in Polish and Russian, see Ozimek et al. (2010) and Warzybok et al. (2015) for a comparison. It can be seen from the figures that the phoneme /x/ is overrepresented, as the number of occurrences in the test is approximately 3.5 times higher than the number of occurrences in the language (similarly in the Polish and Russian tests). This can be explained by the inherent nature of the sentence structure: Adjectives that follow quantifier expressions and the numerals from five on must appear in their genitive plural form, which ends in the phoneme /x/ with all adjectives. The phoneme /e/ is slightly over-represented because it appears in several numerals.

5. Acknowledgements

We would like to thank Nejc Robida and Hotimir Tivadar for useful discussions on some of the issues regarding phonemic transcription, Anna Warzybok for commenting on the matrix test, Darinka Verdonik and Zdravko Kačič for providing the information on SNABI, and the anonymous reviewers for useful comments on the abstract.

6. References

Akeroyd, Michael A., Stig Arlinger, Ruth A. Bentler, Arthur Boothroyd, Nobert Dillier, Wouter A. Dreschler, Jean-Piere Gagne, Mark Lutman, Jan Wouters, Lena Wong and Birger Kollmeier. 2015. International Collegium of Rehabilitative Audiology (ICRA) Recommendations for the Construction of Multilingual

Speech Tests. *International Journal of Audiology*, Early Online:1–6.

Dobrišek, Simon, Kačič, Zdravko, Weiss, Peter, Zemljak Jontes, Melita, Žganec Gros, Jerneja. 2002. Računalniški simbolni fonetični zapis slovenskega govora. *Slavistična revija*, 50(2):159–169.

Erjavec, Tomaž and Nataša Logar Berginc. 2012. Referenčni korpusi slovenskega jezika (cc)Gigafida in (cc)KRES. In T. Erjavec/ J. Žganec Gros (eds), *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Jožef Stefan Institute.

Gigafida – WRITTEN CORPUS, corpus of written Slovene: http://www.gigafida.net/

Hagerman Björn. 1982. Sentences for testing speech intelligibility in noise. *Scandinavian Audiology*, 11:79–87.

Jagoda, Laura, Natalie Giroud, Patrick Neff, Andrea Kegel, Tobias Kleinjung and Martin Meyer. 2018. Speech perception in tinnitus is related to individual distress level – A neurophysiological study. *Hearing Research*, 367:48–58.

Jakopin, Primož. 1999. Zgornja meja entropije pri besedilih v slovenskem jeziku. Doctoral dissertation, University of Ljubljana.

Jurgec, Peter. 2011. Slovenščina ima 9 samoglasnikov. *Slavistična revija*, 59(3):243–268.

Kačič, Zdravko, Bogomir Horvat, Aleksandra Markuš Zögling, Robert Veronik, Matej Rojc, Andrej Žgank, Mirjam Sepesy Maučec and Tomaž Rotovnik. 2002. SNABI Database for Continuous Speech Recognition 1.2. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1051.

Logar Berginc, Nataša and Simon Krek. 2012. New Slovene corpora within the communication in Slovene project. *Prace Filologiczne*, 63:197–207.

Logar Berginc, Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt and Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, FDV.

Marvin, Tatjana, Jure Derganc and Saba Battelino. 2017. Adapting the Freiburg Monosyllabic Word Test for Slovenian. *Linguistica*, 57(1):197–210.

Mihelič, Aleš. 2006. *Sistem za umetno tvorjenje slovenskega govora, ki temelji na izbiri in združevanju nizov osnovnih govornih enot*. Doktorska disertacija, Univerza v Ljubljani.

Ozimek Edward, Warzybok Anna and Kutzner Dariusz. 2010. Polish sentence matrix test for speech intelligibility measurement in noise. *International Journal of Audiology*, 49:444–454.

Hochmuth Sabine, Brand Thomas, Zokoll Melanie A., Zenker Castro Franz Jozef, Wardenga Nina et al. 2012. A Spanish matrix sentence test for assessing speech reception thresholds in noise. *International Journal of Audiology*, 51:536–544.

Plomp Reiner and Mimpen A.M. 1979. Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, 18:43–53.

Pompe, Janko. 1968. *Razvoj avdiometrije na ORL kliniki v Ljubljani*. [Development of audiometry at ORL Clinic in Ljubljana]. Unpublished manuscript, University Medical Center Ljubljana, Ljubljana, Slovenia.

Toporišič, Jože. 2000. *Slovenska slovnica*. Založba Obzorja, Maribor.

Wagener Kirsten, Brand Thomas and Kollmeier Birger. 1999a. Entwicklung und Evaluation eines Satztests in deutscher Sprache Teil II: Optimierung des Oldenburger Satztests (in German). (Development and evaluation of a German sentence test, Part II: Optimization of the Oldenburg sentence tests). *Zeitschrift für Audiologie*, 38:44–56.

Wagener Kirsten, Brand Thomas and Kollmeier Birger. 1999b. Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III: Evaluation des Oldenburger Satztests (in German). (Development and evaluation of a German sentence test – Part III: Evaluation of the Oldenburg sentence test). *Zeitschrift für Audiologie*, 38:86–95.

Wagener Kirsten, Kühnel Volker and Kollmeier Birger 1999c. Entwicklung und Evaluation eines Satztests in deutscher Sprache I: Design des Oldenburger Satztests (in German). (Development and evaluation of a German sentence test – Part I: Design of the Oldenburg sentence test). *Zeitschrift für Audiologie*, 38:4–15.

Warzybok, Anna, Melanie Zokoll, Nina Wardenga, Edward Ozimek, Maria Boboshko and Birger Kollmeier. 2015. Development of the Russian Matrix Sentence Test. *International Journal of Audiology*, 54: 35–43.

Korpusna analiza nestandardne stave vejice po uvajalnih prislovnih zvezah

Eneja Osrajnik,† Darja Fišer,‡* Vojko Gorjanc‡

† Ulica Pohorskega bataljona 43, 1000 Ljubljana
eneja.osrajnik@gmail.com
‡ Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
* Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
darja.fiser@ff.uni-lj.si

Povzetek

V pričujočem prispevku obravnavamo razširjenost nestandardne stave vejice po uvajalnih prislovnih zvezah (UPZ) v slovenskem akademskem diskurzu, natančneje v diplomskih in doktorskih delih, ter diskurzu uporabniško generiranih spletnih vsebin, in sicer v zasebnih in poslovnih blogih. Proučujemo tudi vpliv atraktorjev, tj. besed v predložni funkciji, ki uvajajo UPZ, in vpliv dolžine UPZ na stavo tovrstne vejice. Na podlagi vzorca 6000 zadetkov korpusa spletne nestandardne slovenščine Janes in korpusa akademske slovenščine KAS smo ugotovili, da je tovrstna stava vejice glede na žanr občutno bolj razširjena v diskurzu uporabniško generiranih spletnih vsebin, znotraj akademskega diskurza pa se pogosteje pojavlja v delih študentov dodiplomskega kot podiplomskega študija. Poleg tega se tovrstna stava vejice najpogosteje pojavlja po kratkih UPZ, njena razširjenost pada z večanjem razdalje med atraktorjem in vejico, najpogostejši atraktorji v obeh vrstah diskurza pa so predlogi *zaradi, za, po, na, kljub, poleg* in *ob.*

Corpus analysis of non-standard comma usage after introductory adverbial phrases

This paper addresses the prevalence of non-standard comma usage after introductory adverbial phrases (AIP) in Slovene academic discourse (in undergraduate and PhD theses) and user-generated content (UGC; in private and corporate blogs). We also analyse how the so-called attractors, i.e. words with a prepositional function at the beginning of IAPs, and the length of AIPs impact this type of non-standard comma usage. Based on the comparison of a sample of 6,000 hits in the Janes corpus of non-standard UGC and the KAS corpus of academic texts, it was determined that non-standard comma is considerably more widespread in UGC than in academic discourse, where it is more frequent in undergraduate than postgraduate theses. This type of comma most often occurs after short IAPs, its prevalence declines with the increasing distance between the attractor and the comma, and among the most widespread attractors, regardless of the discourse type, are Slovene prepositions *zaradi (because), za (for), po (after), na (on), kljub (despite), poleg (by),* and *ob (by).*

1. Uvod in namen raziskave

Nestandardna stava vejice, tj. stava v nasprotju s trenutno kodifikacijo, je vedno aktualna tema, saj vejica velja za eno zahtevnejših prvin slovenskega standardnega jezika, »o stavi katere so si bili kritiki najbolj in najpogosteje navzkriž« (Dobrovoljc, 2004: 188). Napake pri stavi vejice se pojavljajo celo v besedilih jezikovnih uporabnikov s formalno jezikoslovno izobrazbo, kar je pokazala raziskava stave vejice v lektoriranih avtorskih besedilih in prevodih v korpusu Lektor (Popič, 2014), kjer je vstavljena vejica najpogostejši lektorski popravek. Poleg tega sta Fišer in Popič (2015) analizirala stavo vejic v uravnoteženem korpusu Kres (Logar et al., 2012) s standardnimi in lektoriranimi besedili ter v korpusu nestandardne spletne slovenščine Janes (Fišer et al., 2014). S primerjavo stave vejic po »predložnih zvezah«,[1] uvedenih zlasti s predlogi *kljub*, *zaradi* in *glede*, sta ugotovila, da je vejica za tovrstnimi zvezami pogosteje stavljena v Janesu kot v Kresu.

To raziskavo je dopolnila pilotna študija o nestandardni stavi vejice v slovenskih tvitih različnih stopenj standardnosti (Popič et al., 2016). Avtorji študije so želeli analizirati, na kakšen način raba vejice v uporabniških spletnih vsebinah odstopa od norme in kako nanjo vpliva formalnost komunikacije. Rezultati analize so pokazali, da odvečna vejica predstavlja zelo majhen delež primerov nestandardne stave vejice in se največkrat pojavlja v besednih zvezah (npr. v sestavljenih veznikih), v vezalnem ali ločnem priredju, za stavčnim členom in med enakovrednima odvisnikoma.

Odločili smo se, da bomo dozdajšnje raziskave nadgradili s primerjavo stave nestandardne »odvečne« vejice po uvajalnih prislovnih zvezah v akademskem diskurzu, in sicer v diplomskih in doktorskih delih, ter diskurzu uporabniško generiranih spletnih vsebin, pri čemer se bomo osredotočili na poslovne in zasebne bloge. Proučevanje »odvečne« vejice je zelo zanimivo, saj tovrstna vejica ne glede na standardnost konteksta opozarja na pomanjkljivo pravopisno znanje piscev besedil – iz psihološkega vidika namreč stava ločil zahteva več napora kot njihovo opuščanje (ibid.: 152).

Cilj pričujočega prispevka je ugotoviti, kakšna je razširjenost nestandardne stave vejice po uvajalnih prislovnih zvezah, ter kako t. i. atraktorji in dolžina uvajalnih prislovnih zvez vplivata na nestandardno stavo vejice v akademskem diskurzu in diskurzu uporabniško generiranih spletnih vsebin.

1.1. Vejica po UPZ v slovenski kodifikaciji

Uvajalne prislovne zveze oz. UPZ (Marko in Osrajnik, 2015: 494) so samostalniške zveze na začetku povedi, ki določajo krajevne, časovne in vzročnostne okoliščine,

[1] V pričujočem prispevku jih imenujemo uvajalne prislovne zveze.

uvedene pa so z besedami v predložni funkciji, ki jih Lengar Verovnik (2003) in Žibert (2006) imenujeta atraktorji: »Opazujoč jezikovno prakso, se zdi, da se vejica piscem še posebej rada prikrade ob stavčnih členih, uvedenih s t. i. lažnimi atraktorji, kot so predlogi kljub, zaradi in glede« (ibid.: 52) (primer [1]).

[1] *Razen ožjih družinskih članov, nihče ne prihaja v otrokovo bližino* (korpus Janes v0.4, podkorpus blogov s komentarji, objave podjetij na platformi rtvslo.si).

Kot pravi Lengar Verovnik (2003: 51), je razlog za tovrstne »odvečne« vejice najbrž možna pretvorba v odvisnik, ki že po definiciji vsebuje povedek in tako zahteva tudi rabo vejice, delno pa uporabnike najbrž zavede tudi stava vejic glede na stavčno fonetiko. Po koncu uvajalnih prislovnih zvez – zlasti daljših – namreč z glasom premolknemo, jezikovni uporabniki pa tako »začutijo potrebo« po stavi vejice. To potrjujejo izsledki pilotne študije iz leta 2015 (Marko in Osrajnik), ki kažejo, da jezikovni uporabniki zaradi kombinacije nepoznavanja slovničnih pravil in stavčnofonetičnih kriterijev pogosto stavljajo nestandardno vejico po UPZ. Kljub temu pa je tovrstna stava vejice v Slovenskem pravopisu 2001 le površno omenjena pri določilu o stavi vejic pri polstavčnih konstrukcijah: »Samostalniških zvez, nastalih iz odvisnika, z vejico ne ločimo od okolja« (SP, 2001: § 333) (primer [2]).

[2] *Ob slovesni podelitvi bralnih značk osnovnošolcem je govoril tudi šolski ravnatelj* (SP, 2001: § 333).

1.2. Akademski diskurz in diskurz uporabniško generiranih spletnih vsebin

V pričujoči raziskavi smo primerjali nestandardno stavo vejice po UPZ v akademskem diskurzu, in sicer v podkorpusih diplomskih in doktorskih del korpusa akademske slovenščine KAS (Erjavec et al., 2016), ter diskurzu uporabniško generiranih spletnih vsebin, in sicer v podkorpusu blogov korpusa spletne nestandardne slovenščine Janes (Fišer et al., 2014). Za to primerjavo smo se odločili, ker nas zanimajo razlike med besedilnimi zvrstmi, ki se pojavljajo v precej različnih kontekstih. Akademski diskurz je relativno homogen, obravnava ozko zastavljeno, pogosto strokovno temo, in nastaja z namenom informiranja ožjega kroga bralcev, pri čemer si avtorji prizadevajo, da bi bila njihova dela sprejeta v določenem krogu bralcev – pogosto gre za strokovnjake na določenem področju (Suomela-Salmi in Dervin 2009: 120). Akademski diskurz mora upoštevati veljavne standarde in zahteve glede jezikovne standardnosti, obsega pa na primer diplomska in magistrska dela, doktorske disertacije, znanstvene članke itd. Po drugi strani diskurz uporabniško generiranih spletnih vsebin zajema vse vrste vsebin, ki jih proizvedejo neplačani uporabniki na spletu, na primer tvite, bloge, forume, komentarje videoposnetkov itd. Jezik uporabniško generiranih spletnih vsebin (zlasti na družbenih omrežjih) pogosto ne upošteva slovničnih pravil, vključuje pogovorne in regionalne izraze, velikokrat pa prihaja tudi do zatipkanih in pravopisnih napak. V nasprotju z akademskim diskurzom tovrstna besedila niso zavezana strogo določenim in natančno predpisanim formalnim standardom za oblikovanje besedil in načeloma ne obravnavajo akademskih tematik.

1.2.1. Korpus Kas

Korpus akademske slovenščine KAS zajema 50.793 besedil oz. 1.189.100.198 pojavnic. Obsega diplomska (81 %) in magistrska dela (13 %), doktorske disertacije (1,4 %), znanstvena dela (1,5 %), kot so prispevki na konferencah in izvirni znanstveni članki, ostala besedila (1,4 %), na primer predgovore, spremna besedila in učna gradiva, specialistična dela (1,1 %) in strokovna dela (0,8 %), kot so strokovni članki in monografije.

Pri naši raziskavi smo uporabili različico korpusa KAS-proto in se osredotočili na podkorpus diplomskih nalog (81 % korpusa), ki zajema 42.212 besedil oz. 850.937.549 pojavnic, in podkorpus doktorskih disertacij (1,4 % korpusa), ki vsebuje 700 besedil oz. 52.874.876 pojavnic. Izpustili smo podkorpus magistrskih nalog, saj se nam je zdelo zanimivo primerjati dve skrajnosti – pravopisno kompetenco študentov po koncu prve stopnje univerzitetnega študija in na koncu njihove univerzitetne izobrazbe.

1.2.2. Korpus Janes

Korpus spletne slovenščine Janes 0.4[2] obsega skupno 9.055.251 besedil oz. 208.261.725 pojavnic in vključuje tvite (83 %), forumska sporočila (9 %), blogovske zapise (4 %), komentarje na spletne novice (3 %) in vsebine pogovornih strani na Wikipediji (1 %). Da bi omejili vrstno raznolikost, smo v raziskavo zajeli le podkorpus blogov (4 % korpusa), ki vključuje izvirne objave in komentarje zasebnih uporabnikov (private) in podjetij (corporate) z domen publishwall.si in rtvslo.si. Pri tem smo v raziskavo zajeli le izvirne objave, saj obsegajo v povprečju skoraj desetkrat več besed na besedilo kot pa komentarji. Izvirne objave v podkorpusu blogov obsegajo štiri kategorije in vključujejo 42.030 besedil oz. 18.256.749 pojavnic. Največji delež so prispevali zasebni uporabniki z domene rtvslo.si (50,3 %), sledijo pa jim podjetja z domene publishwall.si (24,4 %), zasebni uporabniki z domene publishwall.si (19,7 %) in nazadnje podjetja z domene rtvslo.si (5,6 %).

Na blogovska besedila smo se osredotočili tudi zaradi njihove dolžine – obsegajo namreč daljše število besed na besedilo (v povprečju 71,3) kot drugi zajeti žanri, zato so najbolj primerljiva z besedili v korpusu Kas. Poleg tega so se nam blogi zdeli zanimivi, saj jih Crystal (2011: 20) umešča med dve skrajnosti spletnega oz. internetnega jezika, ki predstavlja kombinacijo govorjenega in zapisanega jezika. Crystal v eno skrajnost umešča vrste internetnega jezika, ki se ne razlikujejo bistveno od tradicionalnih besedil. Pri znanstvenih besedilih, kamor prištevamo diplomske naloge in doktorske disertacije, sta digitalna in tiskana oblika celo identični. V drugo skrajnost pa uvršča spletne klepetalnice in podobne platforme za neposredno sporočanje (na primer Facebook ali Twitter), kjer je besedilo sicer zapisano, vendar vsebuje določene ključne lastnosti govorjenega jezika (ibid.), kot so pričakovanje takojšnjega odziva na besedilo, nestalnost besedil (saj se lahko izbrišejo) in sproščeno vzdušje, značilno za pogovor. Obenem glede na jezikovne značilnosti spletnih besedil loči besedila z

[2] Pred objavo tega prispevka je bila objavljena posodobljena različica korpusa 1.0, ker pa so bile vse ročne analize že zaključene, smo v okviru te raziskave ostali pri različici 0.4.

enako zgradbo in grafično oblikovanostjo kot tradicionalni tisk (na primer doktorske disertacije) ter sporočila z omejenim številom znakov in posledično preprostejšo stavčno strukturo (na primer tviti). Med ti skrajnosti umešča bloge, kjer je uporaba različno zapletene zgradbe besedila odvisna od posameznega jezikovnega uporabnika.

2. Zasnova raziskave

Za celostno proučevanje pojava nestandardne vejice po UPZ je treba poznati število vseh povedi z UPZ (z nestandardno stavo vejice oz. z »odvečno« vejico in s standardno stavo brez vejice v skladu s trenutno kodifikacijo) v posameznem podkorpusu. Tovrstne povedi smo v posameznih podkorpusih izluščili z iskalnim ukazom CQL (Corpus Query Language) v konkordančniku Sketch Engine (Kilgariff et al., 2004), nato pa pregledali izluščene zadetke v obliki besednih nizov oz. konkordanc.

2.1. Iskanje zadetkov v podkorpusih

Ker zaradi dolžine in zapletenosti ni bilo mogoče oblikovati enotnega iskanega ukaza, s katerim bi hkrati izluščili zadetke z vejico in brez nje, ob tem pa ne bi vseboval tudi velikega števila nerelevantnih zadetkov, smo skupno število povedi z UPZ in nestandardno stavo vejice poiskali z dvema ločenima iskalnima ukazoma. S prvim iskalnim ukazom smo zajeli vse povedi z nestandardno stavo vejice po UPZ z dolžino do vključno 7 besed po atraktorju, z drugim iskanjem pa povedi z nestandardno stavo vejice po UPZ z dolžino 8 besed in več po atraktorju. Nato smo sešteli število tovrstnih povedi iz obeh iskanj in tako dobili skupno število vseh zadetkov z nestandardno stavljeno vejico ne glede na dolžino UPZ v posameznem podkorpusu. Pri tem smo drugi iskalni ukaz zastavili tako, da smo z njim obenem zajeli tudi povedi z UPZ in standardno stavo vejice.

2.1.1. Prvo iskanje

S prvim iskalnim ukazom smo torej v podkorpusih izluščili vse zadetke z UPZ dolžine od 2 do vključno 7 besed po atraktorju, ki jim sledi nestandardno stavljena vejica (v nadaljevanju »zadetki z vejico znotraj 2–7 besed po atraktorju«). V ukazu smo določili, da najkrajša izluščena UPZ vsebuje vsaj dve besedi po atraktorju, saj smo v praksi opazili, da tovrstnim zvezam z le eno besedo po atraktorju večinoma sledi odvisnik (npr. *Zaradi tega, ker*), takšni primeri pa niso relevantni za našo raziskavo. Omejitev na sedem besed pa je poljubna, saj smo UPZ z osmimi besedami in več zajeli v drugem iskanju.

```
[word="[[:upper:]].+" & tag="D.*"] [word!="," &
tag!="G.*" & word=".*[[:lower:]].*"] [...] {2,7}
[word=","] [tag!="V.*" & word!="zakaj | naj | kar | pa |
ki | kjer "] within <s/>
```

Slika 1: Iskalni izraz za prvo iskanje

Primer [3] zadetka z UPZ s petimi besedami po atraktorju, ki ji sledi nestandardno stavljena vejica:

[3] *Za iskrico dvoma o medsebojnem sporazumevanju, ni ostalo več prostora* (korpus Janes v0.4, podkorpus blogov s komentarji, objave zasebnih uporabnikov na platformi publishwall.si).

2.1.2. Drugo iskanje

Z drugim iskalnim ukazom smo poiskali vse relevantne povedi, v katerih se morebitna vejica pojavlja šele po 8 besedah in več po atraktorju (v nadaljevanju »zadetki z UPZ brez vejice znotraj 2–7 besed po atraktorju«).

Predvidevali smo, da bomo s tem iskalnim ukazom izluščili dve kategoriji zadetkov. Prva obsega vse zadetke z nestandardno stavo vejice po UPZ, ki jih nismo izluščili v prvem iskanju – torej zadetke z UPZ dolžine 8 besed in več, ki jim sledi nestandardno stavljena vejica. Število tovrstnih zajetih zadetkov smo prišteli k relevantnim zadetkom, zajetim s prvim iskanjem, in tako dobili skupno število vseh zadetkov z nestandardno stavljeno vejico ne glede na dolžino UPZ v posameznem podkorpusu.

```
[word="[[:upper:]].+" & tag="D.*"] [word!="," &
tag!="G.*" & word=".*[[:lower:]].*"] [...] {2,7}
[word!="," ] [tag!="V.*" & word!="zakaj | naj | kar | pa
| ki | kjer"] within <s/>
```

Slika 2: Iskalni izraz za drugo iskanje

Primer [4] zadetka z UPZ z osmimi besedami, ki ji sledi nestandardno stavljena vejica:

[4] *Ob navideznem izčrpanju vseh zalog novih konstruktivnih političnih idej, ljudstvo pesimistično razpoloženje, nezaupanje v prihodnost in demoralizacijo preganja tako, da si svoje frustracije in ponižanje nacionalnega ponosa zdravi z uspehi športnikov, ki jih sili /.../* (korpus Janes v0.4, podkorpus blogov s komentarji, objave zasebnih uporabnikov na platformi publishwall.si).

Ker ta iskalni ukaz prepoveduje prisotnost vejice znotraj 2–7 besed po atraktorju, smo zajeli še eno kategorijo zadetkov – povedi s standardno stavo vejice po UPZ. Gre za povedi, skladne s trenutno slovensko kodifikacijo, zato smo pričakovali, da bodo predstavljale precejšen delež zadetkov tega iskanja.

Primer [5] zadetka z UPZ s standardno stavo vejice:

[5] *Po mnenju za migracije pristojnega grškega ministra Ioanisa Muzalasa se meja z Makedonijo za te migrante ne bo več odprla, poroča nemška tiskovna agencija dpa* (korpus Janes v0.4, podkorpus blogov s komentarji, objave zasebnih uporabnikov na platformi publishwall.si).

2.2. Vzorci za analizo

Ker so izbrani podkorpusi preveliki, da bi pregledali vse dobljene zadetke, smo pri vsakem iskanju pregledali vzorec 250 naključnih dedupliciranih zadetkov v vsaki izmed štirih kategorij izvirnih objav Janesovega podkorpusa blogov (objave zasebnih uporabnikov in objave podjetij z domene rtvslo.si ter objave zasebnih uporabnikov in objave podjetij z domene publishwall.si – skupaj 1000 zadetkov), 1000 naključnih zadetkov v Kasovem podkorpusu diplomskih nalog in 1000 naključnih zadetkov v Kasovem podkorpusu doktorskih disertacij. V skupnem seštevku obeh iskanj smo torej ročno pregledali 6000 zadetkov.

Omeniti je treba, da smo med objavami podjetij z domene rtvslo.si pri prvem iskanju analizirali le 240, pri drugem pa 187 zadetkov, saj smo jih toliko izluščili z ukazom CQL. Ker je vzorec pregledanih povedi v drugem iskanju manjši, je tudi število identificiranih povedi z nestandardno vejico v drugem iskanju sorazmerno manjše.

3. Analiza rezultatov in razprava

Z ročnim pregledom smo v obeh iskanjih izluščili zadetke z UPZ (s standardno in nestandardno stavo vejice). Čeprav smo si prizadevali oblikovali kar najbolj učinkovita iskalna ukaza, smo med zadetke zajeli številne nerelevantne zadetke brez UPZ, na primer naslove poglavij in navajanje virov (primer [6]), ter zadetke s standardno stavljeno vejico po UPZ, pri čemer prislovni zvezi sledi odvisnik, vrinjen stavek ali polstavčna struktura (primer [7]). Pojavljali so se tudi primeri povedi, ki se začnejo z naštevanjem, zato so posamezne naštevalne enote ločene z vejico, kot to narekuje trenutna kodifikacija (primer [8]).

[6] *Iz knjige nasvetov, 09.11.15* (korpus Janes v0.4, podkorpus blogov s komentarji, objave zasebnih uporabnikov na platformi rtvslo.si).

[7] *Na podlagi rezultatov, dobljenih pri izvajanju encimske esterifikacije D,L-MK v SC CO2 pri 7,5 MPa in 35 °C, smo v nadaljevanju izvedli encimsko esterifikacijo z razmerjem substratov 1:3,6, brez molekularnih sit in s hitrostjo mešanja 700 obr/min* (korpus KAS-proto, podkorpus doktorskih disertacij).

[8] *Ob današnji strukturi, strokovni usposobljenosti in številu gradbenih inšpektorjev, finančni podhranjenosti gradbene inšpekcije in ob upoštevanju prepovedi zaposlovanja upravičeno dvomimo, da bodo inšpekcijske službe kos tej pomembni nalogi* (korpus Janes v0.4, podkorpus blogov s komentarji, objave podjetij na platformi rtvslo.si).

Tovrstne zadetke smo zanemarili, saj niso relevantni za našo raziskavo. V nadaljevanju smo podrobneje analizirali identificirane relevantne primere povedi z nestandardno in standardno stavo vejice po UPZ, pri čemer smo analizo razdelili na tri dele: (1) določanje razširjenosti nestandardne stave vejice po UPZ v posameznih podkorpusih, (2) analiza razdalje (tj. števila besed) med atraktorjem in nestandardno vejico ter (3) analiza atraktorjev, po katerih uporabniki stavljajo nestandardno vejico.

3.1. Razširjenost nestandardne vejice po UPZ

Prvi del analize smo razdelili na dva koraka. V prvem smo z enostavnim sklepnim računom ocenili število vseh zadetkov s standardno in nestandardno stavo vejice po UPZ v posameznem podkorpusu. To pomeni, da smo v pričujočem prispevku celotno število tovrstnih zadetkov v podkorpusu ocenili po naslednji enačbi:

$$n_{(ne)standardnih} = n_{zadetkov} \cdot \frac{n_{identificiranih}}{n_{pregledanih}}$$

Slika 3: Sklepni račun za ocenitev števila vseh zadetkov s standardno in nestandardno stavo vejice po UPZ v posameznem podkorpusu.

Ob tem:

– $n_{(ne)standardnih}$ označuje ocenjeno število primerov s standardno (*Ocenjeno št. s standardno stavo vejice* v tabeli 2) oz. nestandardno stavo vejice (*Ocenjeno št. z nestandardno stavo vejice* v tabelah 1 in 2) po UPZ v posameznem podkorpusu,

– $n_{zadetkov}$ označuje število vseh izluščenih zadetkov v posameznem podkorpusu (npr. 139.576 v Kasovem podkorpusu diplomskih nalog; gl. tabelo 1),

– $n_{identificiranih}$ označuje število identificiranih primerov z nestandardno stavo vejice po UPZ v vzorcu $n_{pregledanih}$ (npr. 400 v Kasovem podkorpusu diplomskih nalog; gl. tabelo 1),

– $n_{pregledanih}$ označuje število pregledanih zadetkov v posameznem podkorpusu (npr. 1000 v Kasovem podkorpusu diplomskih nalog).

Rezultate izračuna za ocenitev števila identificiranih primerov s standardno in nestandardno vejico po UPZ v podkorpusih predstavljamo v tabelah 1 in 2, pri čemer:

– »KAS Dipl« označuje Kasov podkorpus diplomskih nalog,

– »KAS Dr« označuje Kasov podkorpus doktorskih disertacij,

– »RTV Z« označuje objave zasebnih uporabnikov bloga rtvslo.si v korpusu Janes,

– »RTV P« označuje objave podjetij z bloga rtvslo.si v korpusu Janes,

– »PUB Z« označuje objave zasebnih uporabnikov bloga publishwall.si v korpusu Janes in

– »PUB P« označuje objave podjetij z bloga publishwall.si v korpusu Janes.

Podkorpus	Št. vseh zadetkov	Št. pregledanih zadetkov	Št. identificiranih primerov z nestandardno stavo vejice (delež glede na št. pregledanih zadetkov)	Ocenjeno št. z nestandardno stavo vejice
KAS Dipl	139.576	1.000	400 (40,0 %)	55.830
KAS Dr	5.423	1.000	254 (25,4 %)	1.377
RTV Z	1.759	250	86 (34,4 %)	605
RTV P	240	240	102 (42,5 %)	102[3]
PUB Z	844	250	89 (35,6 %)	300
PUB P	753	250	80 (32,0 %)	241

Tabela 1: Rezultati iskanja povedi z vejico znotraj 2–7 besed po atraktorju.

[3] V tem primeru ne gre za ocenitev, temveč dejansko število primerov, saj smo pregledali vse zadetke.

Podkorpus	Št. vseh zadetkov	Št. pregledanih zadetkov	Št. identificiranih primerov z nestandardno stavo vejice (delež glede na št. pregledanih zadetkov)	Ocenjeno št. z nestandardno stavo vejice	Št. ident. primerov s stand. stavo vejice (delež glede na št. pregledanih zadetkov)	Ocenjeno št. s standardno stavo vejice
KAS Dipl	199.194	1.000	67 (6,7 %)	13.346	778 (77,8 %)	154.972,9
KAS Dr	91.523	1.000	42 (4,2 %)	3.844	836 (83,6 %)	76.513,2
RTV Z	883	250	23 (9,2 %)	81	142 (56,8 %)	501,5
RTV P	187	187	21 (11,2 %)	21	118 (63,1 %)	118,0
PUB Z	440	250	25 (10,0 %)	44	139 (55,6 %)	244,6
PUB P	832	250	18 (7,2 %)	60	169 (67,6 %)	562,4

Tabela 2: Rezultati iskanja povedi brez vejice znotraj 2–7 besed po atraktorju.

V drugem koraku tega dela analize smo na podlagi dobljenih rezultatov ocenitve izračunali še relativni delež nestandardne stave vejice v posameznih podkorpusih.

$$razširjenost\ nestand. vejice = \frac{n_{nestandardnih}}{n_{nestandardnih} + n_{standardnih}}$$

Slika 4: Enačba relativnega deleža za izračun razširjenosti nestandardne stave vejice v podkorpusih.

Tako smo dobili odstotek povedi z nestandardno stavo vejice po UPZ znotraj vseh izluščenih zadetkov v posameznem podkorpusu. Rezultati izračuna razširjenosti nestandardne stave vejice po UPZ so predstavljeni v tabeli 3:

	Št. primerov s stand. vejico	Št. primerov z nest. vejico	Razširjenost nest. vejice po UPZ
KAS Dipl	154.972,9	69.176,4	30,9 %
KAS Dr	76.513,2	5.221,4	6,4 %
RTV Z	501,5	686,3	57,8 %
RTV P	118,0	123,0	51,0 %
PUB Z	244,6	344,5	58,5 %
PUB P	562,4	300,9	34,9 %

Tabela 3: Razširjenost nestandardne stave vejice po UPZ v posameznem podkorpusu.

Kot je razvidno iz rezultatov, se nestandardna vejica najpogosteje pojavlja v objavah zasebnih uporabnikov blogovskih besedil (58,5 % na blogu publishwall.si in skoraj 58 % na blogu rtvslo.si), nekoliko manj je razširjena v objavah podjetij bloga rtvslo.si (51 %), zanimiv pa je precej nižji delež razširjenosti v objavah podjetij na publishwall.si (skoraj 35 %).

Po pričakovanjih se nestandardna stava vejice najredkeje pojavlja v podkorpusu doktorskih disertacij (nekaj več kot 6 %), presenetljivi pa so rezultati za podkorpus diplomskih nalog, v katerih njena razširjenost znaša skoraj 31 %. Na podlagi teh rezultatov lahko torej sklepamo, da študenti dodiplomskega študija nestandardno vejico stavijo precej pogosteje kot študenti podiplomskega študija. Oboji so sicer svoje izobraževanje na področju slovnice in pravopisa končali že v srednji šoli (razen študentov jezikoslovnih smeri), vendar imajo študenti po koncu študija več izkušenj pri pisanju formalnih besedil in morda zato vejico stavijo redkeje, zelo verjetno pa je tudi, da primere nestandardne stave vejice v njihovih besedilih prestrežejo mentorji ali lektorji.

Poleg tega lahko sklepamo, da je nestandardna stava vejice po UPZ občutno bolj razširjena v besedilih korpusa Janes kot korpusa Kas. To smo tudi pričakovali, saj morajo pisci besedil v okviru akademskega diskurza strogo upoštevati pravila knjižne slovenščine, medtem ko pisci uporabniško generiranih spletnih vsebin jezikovni standardnosti posvečajo manj pozornosti. V slednjih je pogosta prisotnost odvečne vejice vendarle zanimiva, saj je znano, da uporabniki v uporabniško generiranih besedilih pogosto opuščajo črke in ločila ter uporabljajo okrajšave oz. simbole (Fišer et al., 2018: 125), da prihranijo čas in prostor, zato bi morda prej pričakovali opuščanje sicer obveznih vejic.

3.2. Razdalja med atraktorjem in vejico

V drugem delu analize smo proučevali razdaljo (število besed) med atraktorjem in nestandardno vejico v posameznih podkorpusih. Za pridobitev uravnoteženih rezultatov analize obeh iskanj smo v drugem delu analize (v prvem delu to ni bilo potrebno, saj analiza temelji na deležih) število povedi z nestandardno vejico v večjem vzorcu sorazmerno zmanjšali, in sicer smo ga pomnožili z razmerjem med vzorcema (187/240). Tako smo izračunali število povedi z nestandardno vejico, kot če bi tudi v prvem iskanju pregledali vzorec 187 povedi.

V tabeli 4 predstavljamo rezultate analize stave nestandardne vejice glede na razdaljo med atraktorjem in vejico v posameznih podkorpusih:

Razdalja med atraktorjem in vejico	Kas Dipl	Kas Dr	RTV P	RTV Z	PUB Z	PUB P	Skupaj (delež)
2	86	46	23,4	32	22	26	235,4 (19,9 %)
3	89	57	22,6	18	21	18	225,6 (19,0 %)
4	97	50	11,7	19	16	14	207,7 (17,5 %)
5	50	40	10,9	10	11	18	139,9 (11,8 %)
6	45	36	4,7	3	6	9	103,7 (8,8 %)
7	33	25	6,2	4	4	4	76,2 (6,4 %)
8	15	13	10	7	5	9	59 (5,0 %)

9	18	10	3	8	3	7	49	(4,1 %)
10	9	3	3	3	4	1	23	(1,9 %)
11	5	6	0	3	2	3	19	(1,6 %)
12	7	4	1	0	2	1	15	(1,3 %)
13	4	2	1	1	0	2	10	(0,8 %)
14	4	0	2	1	1	0	8	(0,7 %)
15	0	2	1	0	0	0	3	(0,3 %)
16	2	0	0	0	0	0	2	(0,2 %)
17	2	1	0	0	0	0	3	(0,3 %)
18	0	0	0	0	0	0	0	(0,0 %)
19	0	0	0	0	0	2	2	(0,2 %)
20	0	1	0	0	0	0	1	(0,1 %)
21	0	0	0	0	1	0	1	(0,1 %)
Skupaj:	467	296	100,5	109	98	114	1.184,5	(0,1%)

Tabela 4: Število primerov glede na razdaljo in podkorpus ter delež vsote primerov posameznih razdalj glede na število vseh primerov.

Na podlagi izsledkov dozdajšnjih raziskav (Marko in Osrajnik, 2015) glede pogostejše nestandardne stave vejice po daljših UPZ zaradi stavčnofonetičnih kriterijev smo pričakovali, da bo nestandardna stava vejice naraščala z večanjem števila besed med atraktorjem in vejico, po določeni razdalji pa začela upadati. Ugotovili pa smo, da v splošnem njena stava z večanjem razdalje precej enakomerno pada. Ob tem smo več kot polovico (skoraj 60 %) primerov identificirali v UPZ z 2–4 besedami med atraktorjem in vejico. Najkrajše UPZ (z 2 besedama med atraktorjem in vejico) predstavljajo kar 20 % vseh identificiranih primerov nestandardne stave vejice, temu tesno sledi stava vejice po 3 (19,2 %) in 4 besedah (17,5 %), šele nato se pojavi nekoliko večji preskok do deleža stavljenih vejic po 5 besedah (11,8 %). Po UPZ z 10 besedami med atraktorjem in vejico pa se v podkorpusih pojavljajo le še posamezni primeri povedi z nestandardno stavo vejice.

Stava nestandardne vejice enakomerno pada tudi po posameznih podkorpusih. Manjše izjeme so Kasov podkorpus diplomskih nalog, kjer se največ primerov z vejico pojavi po UPZ s 4 besedami, podkorpus doktorskih disertacij, kjer je vejic po UPZ s 3 in 4 besedami nekoliko več kot vejic po UPZ z 2 besedama, in zasebne objave v Janesovem podkorpusu blogov publishwall.si, kjer je število primerov z vejico po UPZ s petimi besedami nekoliko večje od števila nestandardnih vejic po UPZ s štirimi besedami.

Preskoki med posameznimi kategorijami Janesovega podkorpusa blogov so nekoliko večji, kar je razumljivo, saj smo raziskavo opravili na manjših vzorcih. Zanimivo je, da se kar v polovici podkorpusov pojavi poved z UPZ, ki vsebuje vsaj 20 besed med atraktorjem in vejico. Primera najdajše UPZ (dolžine 21 besed) z nestandardno stavo vejice smo identificirali v Kasovem podkorpusu diplomskih nalog (primer [9]) in v objavah podjetij Janesovega podkorpusa blogov publishwall.si (primer [10]):

[9] *Ob tej razdelitvi zahodne Evrope na Evropsko gospodarsko skupnost (EGS) in Evropsko združenje za prosto trgovino (EFTA) konec petdesetih let prejšnjega stoletja, se je Danska skupaj z Združenim kraljestvom, Švedsko, Norveško, Švico, Avstrijo in Portugalsko pridružila zvezi EFTA* (korpus KAS-proto, podkorpus diplomskih nalog).

[10] *Ob napovedih načrtovanja dveh plinskih terminalov v italijanskem delu Tržaškega zaliva in po začetku predvidenih postopkov za izdajo potrebnih soglasij italijanske vlade, je vlada Republike Slovenije, na pobudo zainteresirane /.../* (korpus Janes v0.4, podkorpus blogov s komentarji, objave podjetij na platformi publishwall.si).

4. Atraktorji

V tretjem delu analize smo podrobneje analizirali atraktorje v zadetkih z nestandardno stavo vejice in primerjali atraktorje v povedih z nestandardno in standardno stavo vejice v korpusih KAS (torej skupno v vseh štirih podkorpusih blogov) in Janes (torej skupno v podkorpusih diplomskih in doktorskih nalog).

4.1. Atraktorji v povedih z nestandardno stavo vejice

Pri podrobni analizi atraktorjev v povedih z nestandardno vejico (iz obeh korpusov) smo ugotovili, da vsi identificirani atraktorji spadajo med predloge, edina izjema je oziralni zaimek *kakršnihkoli*. Identificirane atraktorje smo glede na skladenjsko oz. pomensko razmerje, ki ga vzpostavijo v UPZ, razdelili v 16 kategorij: vzročnost (*zaradi, ob, od, vsled*), rezultat (*do*), izbor (*od, med, izmed*), odnosnost (*pod*), predmetnost/opredelitev (*kakršnihkoli, za, pri, ob, med, glede, pod, znotraj, nad*), odsotnost/izvzemanje (*brez, razen*), premikanje/sprememba (*iz, onkraj*), časovnost (*po, za, na, pri, ob, od, pred, med, skozi, do, okrog*), prostorskost/položaj (*na, pred, znotraj, sredi, izza*), stališče/vir (*po, na, iz, skozi*), primerjava/nasprotje/zamenjava (*po, za, na, namesto*), nasprotovanje (*zoper, proti*), namen (*za*), način/sredstvo (*na, med, skozi, preko*), dopustnost (*kljub, navkljub*) in dodajanje/naštevanje/stopnjevanje (*poleg, razen, zraven*).

4.2. Atraktorji v korpusih KAS in Janes

Glede na razširjenost posameznih predlogov v obeh korpusih smo ugotavljali, kateri najpogosteje uvajajo povedi s standardno in nestandardno stavo vejice. Tako smo preverili, ali se v povedih s standardno oz. nestandardno stavo vejice tipično pojavlja določen nabor besed v predložni funkciji – tako bi namreč lahko potrdili, da gre za močne »atraktorje«, ki govorce »zavajajo« k stavi nestandardne vejice.

Rezultati primerjave so predstavljeni v tabeli 5, zaradi prostorske omejitve pa navajamo le 10 najbolj razširjenih besed v predložni funkciji, popoln seznam pa je objavljen v Osrajnik (2018).

KAS				Janes			
Nestandardna stava vejice (delež)		Standardna stava vejice (delež)		Nestandardna stava vejice (delež)		Standardna stava vejice (delež)	
zaradi	16,9 %	za	19,6 %	po	20,9 %	po	22,2 %
za	13,4 %	pri	16,3 %	zaradi	11,5 %	na	18,1 %
pri	11,7 %	na	16,0 %	na	10,4 %	za	15,1 %
po	10,9 %	po	10,1 %	za	9,9 %	zaradi	7,7 %
na	9,7 %	zaradi	10,0 %	ob	7,4 %	ob	5,8 %
kljub	9,4 %	ob	4,8 %	kljub	5,0 %	pri	4,0 %
poleg	9,3 %	kljub	4,5 %	poleg	4,7 %	kljub	3,7 %
ob	5,5 %	med	4,3 %	od	4,1 %	med	3,5 %
med	1,8 %	poleg	3,7 %	pri	3,8 %	od	3,3 %
pred	1,8 %	iz	3,4 %	namesto	3,6 %	poleg	2,8 %

Tabela 5: 10 najbolj razširjenih predlogov, ki uvajajo UPZ s standardno in nestandardno stavo vejice, in njihov delež v izbranih podkorpusih korpusov KAS in Janes.

V povedih z UPZ in nestandardno stavo vejice obeh korpusov smo identificirali skupno 34 besed v predložni funkciji, pri čemer se 22 predlogov pojavi v obeh korpusih, 3 se pojavijo le v KAS-u (*znotraj, izmed, vsled*), 9 pa le v Janesu (*zraven, zaradi, nad, proti, onkraj, sredi, kakršnihkoli, okrog, izza, zoper*). V povedih z UPZ in standardno stavo vejice pa se pojavi skupno 31 različnih besed v predložni funkciji, in sicer se jih 21 pojavlja v obeh korpusih, 5 smo jih našli le v KAS-u (*znotraj, izmed, konec, vzdolž, zraven*), 5 pa le v Janesu (*čez, okoli, kraj, onstran, mimo*).

Po primerjavi najpogostejših predlogov v zadetkih s standardno in nestandardno stavo vejice – podrobneje smo analizirali predloge z vsaj 5-odstotno razširjenostjo v korpusu – smo ugotovili, da se v KAS-u tako v povedih s standardno kot z nestandardno stavo vejice najpogosteje pojavljajo atraktorji *za, pri, na, po, zaradi,* v povedih z nestandardno stavo vejice pa tudi predlogi *kljub, poleg* in *ob*. V Janesu pa so tako v povedi s standardno kot z nestandardno stavo vejice med najbolj razširjenimi atraktorji predlogi *po, na, za, zaradi* in *ob*, v povedih z nestandardno stavo vejice pa je precej pogosti tudi predlog *kljub*.

Zavedati se je treba, da so omenjeni predlogi nasploh pogosto rabljeni v slovenščini, saj z njimi izražamo najrazličnejša razmerja in pomene, kot so vzročnost (npr. s predlogom *zaradi*), predmetnost (npr. s predlogom *za*), časovnost (npr. s predlogoma *pri* in *ob*), primerjavo (npr. s predlogom *na*), stališče (npr. s predlogom *po*), dopustnost (s predlogom *kljub*) in naštevanje (npr. s predlogom *poleg*). Ker so že nasploh pogosto rabljeni, je težko trditi, da imajo določeni predlogi vlogo atraktorja, ki uporabnike »zavaja« k stavi nestandardne vejice po UPZ, zlasti kadar so primerljivo razširjeni tako v povedih s standardno kot z nestandardno stavo vejice oz. so v nekaterih primerih celo pogostejši v povedih s standardno kot z nestandardno stavo vejice (npr. predlogi *za, pri, na* in *po*). Lahko pa na osnovi občutno večje razširjenosti določenih predlogov v povedih z nestandardno stavo vejice sklepamo, da je tovrstna stava vejice verjetnejša v povedih z določenimi predlogi, kot so *zaradi* (skoraj 7 % večja razširjenost v povedih z nestandardno kot s standardno vejico v KAS-u in skoraj 4 % večja razširjenost v povedih z nestandardno kot s standardno vejico v Janesu), *kljub* (skoraj 5 % večja razširjenost v povedih z nestandardno kot s standardno vejico v KAS-u in skoraj 1,5 % večja razširjenost v povedih z nestandardno kot s standardno vejico v Janesu), *poleg* (5,6 % večja razširjenost v povedih z nestandardno kot s standardno vejico v KAS-u in skoraj 2 % večja razširjenost v povedih z nestandardno kot s standardno vejico v Janesu) in *ob* (nekaj manj kot 1 % večja razširjenost v povedih z nestandardno kot s standardno vejico v KAS-u in nekaj več kot 1,5 % večja razširjenost v povedih z nestandardno kot s standardno vejico v Janesu).

5. Zaključek

Analiza v pričujočem prispevku je pokazala, da je nestandardna stava vejice po UPZ pogostejša v diskurzu uporabniško generiranih spletnih vsebin kot v akademskem diskurzu. Najbolj razširjena je v blogovskih objavah zasebnih uporabnikov, ki uporabljajo najbolj sproščen jezik, najredkeje se pojavlja v doktorskih disertacijah, opazna pa je tudi precejšnja razlika med razširjenostjo vejice v doktorskih in diplomskih delih. Čeprav je stava vejice v slednjih relativno pogosta, je kljub temu redkejša kot v besedilih korpusa Janes.

Zaradi stavčnofonetičnih kriterijev bi sicer pričakovali, da je nestandardna stava vejice pogostejša po daljših UPZ, vendar smo ugotovili, da njena stava z večanjem razdalje med atraktorjem in nestandardno vejico enakomerno pada tako v splošnem kot tudi po posameznih podkorpusih. Več kot polovica primerov nestandardne stave vejice se pojavi po UPZ z razdaljo 2–4 besed med atraktorjem in vejico, kar 20 % vseh identificiranih primerov pa predstavljajo najkrajše UPZ (z 2 besedama med atraktorjem in vejico). Ugotovili smo tudi, da so najpogostejši atraktorji v obeh proučevanih korpusih predlogi *za, pri, na, po, zaradi, kljub, poleg* in *ob*, pri čemer so atraktorji *zaradi, kljub, poleg* in *ob* bolj razširjeni v povedih z nestandardno stavo vejice.

Rezultati te raziskave so potrdili ugotovitve dosedanjih študij, da imajo uporabniki slovenskega jezika težave pri stavi vejice po UPZ tako v akademskem diskurzu kot v diskurzu uporabniško generiranih spletnih vsebin. S tem smo še dodatno opozorili na potrebo po podrobnejši pojasnitvi tovrstne stave vejice v kodifikacijskih priročnikih, pa tudi pri poučevanju slovnice in pripravi učnih gradiv.

Poleg tega smo v pričujočem prispevku ugotovili, da določene besede v predložni funkciji zares igrajo vlogo atraktorjev, ki jezikovne uporabnike »zavajajo« k nestandardni stavi vejice po UPZ. Za dokončno potrditev obstoja tovrstnih atraktorjev bi bilo treba raziskavo v prihodnosti izvesti na večjem vzorcu povedi, zanimivo pa bi bilo raziskavo razširiti še na magistrske naloge in tako dopolniti proučevanje tovrstne stave vejice v zaključnih pisnih delih vseh treh univerzitetnih študijskih stopenj. Poleg tega bi bilo dragoceno analizirati še druge sorodne besedilne žanre, kot so različna specialistična (tj. strokovna in tehnična) besedila. Glede na rezultate analize, da se nestandardna vejica najpogosteje stavi po krajših UPZ, pa bi v raziskavo lahko zajeli tudi krajša uporabniško generirana besedila, kot so tviti.

Zahvala

Avtorji se zahvaljujejo anonimnim recenzentom za koristne pripombe. Raziskavo, opisano v prispevku, je podprl projekt ARRS J6-7094 »Slovenska znanstvena besedila: viri in opis«.

6. Literatura

Andrea L. Berez in Stefan Th. Gries. 2009. In defense of corpus-based method: a behavioral profile analysis of polysemous *get* in English. Moran, Steven, Darren Tanner in Michael Scanlon (ur.): *Proceedings of the 24th Northwest Linguistics Conference.* University of Washington Working Papers in Linguistics vol. 27. Seattle, WA, Department of Linguistics, str. 157–166.

David Crystal. 2011. *Internet Linguistics: A student guide*. London, New York, Routledge.

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

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar in Milan Ojsteršek. 2016. Slovenska akademska besedila: prototipni korpus in načrt analiz. V: T. Erjavec (ur.) in D. Fišer (ur.). *Zbornik konference Jezikovne tehnologije in digitalna humanistika*. Ljubljana, , Filozofska fakulteta, Univerza v Ljubljani, Slovenija, str. 58–64.

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez in Nikola Ljubešić. 2014. Janes se predstavi: metode, orodja in viri za nestandardno slovenščino. V: T. Erjavec in J. Žganec Gros (ur.). *Language technologies: proceedings of the 17th International Multiconference Information Society – IS 2014*. Ljubljana, Slovenija. Institut »Jožef Stefan«, str. 56–61.

Darja Fišer in Damjan Popič. 2015. Vejica je mrtva, živela vejica. *Simpozij Obdobja 34*. Ljubljana, Znanstvena založba Filozofske fakultete Univerz v Ljubljani, str. 609–618.

Darja Fišer, Maja Miličević Petrović in Nikola Ljubešić. 2018. Zapisovalne prakse v spletni slovenščini. Fišer, Darja (ur.): *Viri orodja in metode za analizo spletne slovenščine.* Ljubljana, Znanstvena založba Filozofske fakultete Univerze v Ljubljani, str. 124–139.

Tina Lengar Verovnik. 2003. Vejica premalo, vejica preveč (2). *Pravna praksa* 22(21): 51.

Igor Locatelli. 2011. Randomizacija in velikost vzorca. Predstavljeno 9. 12. 2011 na Fakulteti za farmacijo Univerze v Ljubljani.

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.

Dafne Marko in Eneja Osrajnik. 2015. Slovenska vejica in stavčna fonetika – zakaj ne? *Simpozij Obdobja 34.* Ljubljana, Znanstvena založba Filozofske fakultete Univerze v Ljubljani, str. 493–501.

Eneja Osrajnik. 2018. *Nestandardna stava vejice po uvajalnih prislovnih zvezah v slovenščini*. Magistrsko delo, Ljubljana, Filozofska fakulteta (v tisku).

Damjan Popič. 2014. *Korpusnojezikoslovna analiza vplivov na slovenska prevodna besedila* (doktorska disertacija). Ljubljana, Filozofska fakulteta.

Damjan Popič, Darja Fišer, Katja Zupan in Polona Logar. 2016. Raba vejice v uporabniških spletnih vsebinah. *Proceedings of the Conference on Language Technologies & Digital Humanities, September 29th – October 1st, 2016 Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*. Ljubljana, Slovenija. Filozofska fakulteta Univerze v Ljubljani, str. 149–153.

SP 2001 = *Slovenski pravopis* (Jože Toporišič et al.). Ljubljana, Založba ZRC, ZRC SAZU.

Eija Suomela-Salmi in Fred Dervi. 2009. *Cross-linguistic and crosscultural perspectives on academic discourse.* Amsterdam: John Benjamins Publishing Company.

Živa Žibert. 2006. *Slovenska vejica: balast ali skladenjska nujnost slovenskega knjižnega jezika?* Diplomsko delo, Ljubljana, Fakulteta za družbene vede.

Trajnost digitalnih izdaj
Uporaba statičnih spletnih strani na portalu Zgodovina Slovenije - SIstory

Andrej Pančur*

* Inštitut za novejšo zgodovino
Kongresni trg 1, 1000 Ljubljana
andrej.pancur@inz.si

Povzetek

Prispevek izhaja iz stališča, da je pri digitalnih izdajah potrebno poskrbeti za čim bolj celovito digitalno trajnost tako podatkov kot prezentacij, funkcionalnosti in programske kode. To je velik izziv predvsem za manjše digitalno humanistične projekte z omejenim financiranjem, ki ne omogoča dolgoročnega vzdrževanja tehnično zahtevnih digitalnih izdaj. Kot alternativno rešitev so v prispevku predstavljene rešitve, ki jih v zadnjih letih ponuja hiter razvoj statičnih spletnih strani. Digitalne izdaje, ki temeljijo na TEI, so s pomočjo osnovnih XML (XSLT) in spletnih tehnologij (HTML, CSS, JavaScript) kot statične spletne strani uspešno vključene v repozitorij portala SIstory. Vse statične spletne strani imajo tudi možnost dinamičnega prikazovanja vsebine.

Sustainability of digital editions
Use of static web pages at the History of Slovenia – SIstory portal

The contribution is based on the position that, with regard to digital editions, the highest possible degree of digital sustainability of data, presentations, functionalities, and programme code should be ensured. This represents a significant challenge, especially in case of smaller digital humanities projects with limited financing, which does not allow for the long-term maintenance of technically-demanding digital editions. The alternative solutions facilitated by the swift development of static web pages in the recent years are presented in the contribution. Digital editions based on the TEI have been successfully included in the SIstory portal repository as static web pages, employing basic XML (XSLT) and web technologies (HTML, CSS, JavaScript). All the static web pages also have the possibility of displaying dynamic content.

1. Uvod

V digitalni humanistiki je že dolgo časa prisotno zavedanje o pomembnosti digitalne trajnosti in trajnega ohranjanja digitalnih virov (Schaffner in Erway, 2014: 7). Raziskovalni podatki posameznega projekta ponavadi obstajajo dlje kot sam projekt, v okviru katerega so bili ti podatki zbrani, urejeni in objavljeni. Zato je zelo pomembno, da tudi po zaključku projekta poskrbimo za kvalitetno in trajnostno hrambo digitalnih podatkov.

V zadnjih letih se je intenzivno razpravljalo o tehničnih vidikih upravljanja z raziskovalnimi podatki in njihovega dolgoročnega arhiviranja: metapodatki, arhivski formati in mediji za hrambo, dokumentacija. Toda šele v zadnjem času smo se začeli zavedati, da je za kvalitetno upravljanje in ponovno uporabo teh podatkov skoraj še bolj pomembno ohranjevanje podatkov v skladu s specifičnimi potrebami različnih znanstvenih disciplin. (Moeller et al., 2018) Medtem, ko v naravoslovnih in v družboslovnih vedah v glavnem uporabljajo podatke iz meritev in vprašalnikov, je v humanističnih vedah v prvem planu uporaba kulturnih objektov kot so rokopisi, besedila, slike in posnetki. V humanističnih raziskavah nato digitalne kulturne objekte pogosto še dodatno obdelajo, vizualizirajo, označijo, povežejo in interpretirajo (DHd-AG Datenzentren, 2017: 7).

Takšen način obdelave podatkov je še zlasti pomemben pri digitalnih izdajah, ki so ključen del digitalne humanistike (Andorfer et al., 2016). Digitalne znanstvene izdaje so seveda v prvi vrsti predvsem raziskave, v okviru katerih nastajajo različni prepisi, označbe, analize, razlage ipd. Zato bi morali biti predvsem ti raziskovalni podatki dolgoročno in pod odprtimi pogoji dostopni raziskovalni skupnosti (Robinson, 2016). Tako je pri digitalnih izdajah kodirano besedilo najpomembnejši dolgoročni rezultat projekta. Zelo pomemben je tudi sam prikaz, ki v okviru določene aplikacije predstavlja pogled projektne skupine na te podatke. Toda vsak tak pogled še zdaleč ni unikaten ali celo edini možen, temveč se lahko te podatke prikazuje na različne možne načine. (Turska et al., 2016) Z vsako novo interpretacijo se število možnih načinov prikazovanja še dodatno povečuje. Vsaka takšna prezentacija pa je nov raziskovalni rezultat, ki si prav tako zasluži dolgoročno hrambo.

Zato v humanističnih vedah rezultati raziskav niso samo raziskovalni podatki, temveč tudi predstavitveno okolje in aplikacije, ki omogočajo interpretacijo podatkov, iskanje, filtriranje in brskanje po podatkih in njihovo povezovanje (DHd-AG Datenzentren, 2017: 7). Če bi torej hranili samo raziskovalne podatke, bi bila prvotna prezentacija za vedno izgubljena, čeprav je tudi prezentacija sestavni del digitalne izdaje (Fechner, 2018). Obenem ne smemo pozabiti, da je tudi programska koda, ki smo jo uporabil za izdelavo digitalnih izdaj, prav tako kot digitalna izdaja tudi sestavni del znanstvene argumentacije (Andrews in Zundert, 2016).

Zato trajnostna hramba digitalnih izdaj predstavlja še toliko večji izziv. Pri tem se lahko digitalne izdaje glede na vsebino, izgled in funkcionalnosti med seboj zelo razlikujejo. Večinoma so rezultat specifičnega raziskovalnega projekta, ki ima na razpolago relativno omejene finančne in človeške vire. Ker člani projektnih skupin, ki izhajajo iz humanističnih ved, po navadi nimajo ustreznih tehničnih znanj, se morajo pri tehničnem razvoju večinoma zanašati na zunanje izvajalce. Pri tem so digitalne izdaje odvisne od zelo hitrega razvoja spletnih tehnologij in standardov (Andorfer et al., 2016).

Ti izzivi glede trajnostne hrambe digitalnih izdaj se bodo s hitrim naraščanjem števila digitalnih izdaj v bodoče samo še dodatno okrepili (Fechner, 2018). Velik izziv je in bo predvsem za manjše digitalno humanistične projekte z omejenim financiranjem, ki ne omogočajo dolgoročnega vzdrževanja tehnično zahtevnih digitalnih izdaj. Kot alternativo rešitev bom v nadaljevanju predstavil rešitve, ki jih ponuja hiter razvoj statičnih spletnih strani. V zadnjih nekaj letih so statične spletne strani postale eden glavnih trendov spletnega razvoja. Vse kaže, da se bo ta trend nadaljeval tudi v prihodnje.[1] V prispevku bom predstavil izkušnje, ki sem jih z generiranjem statičnih spletnih strani za digitalne izdaje pridobil v okviru dejavnosti Raziskovalne infrastrukture slovenskega zgodovinopisja, ki med drugim upravlja tudi portal Zgodovina Slovenije – SIstory.[2] Pri tem se bom omejil samo na statične spletne strani, ki jih generiramo iz XML datotek, kodirane v skladu s Smernicami Text Encoding Initiative (TEI) (TEI Consortium, 2018). Smernice TEI so namreč v digitalni humanistiki *de facto* standard za kodiranje besedil, ki ga uporablja veliko različnih humanističnih projektov in raziskav (Romary et al., 2017: 5).

V drugem poglavju bom najprej predstavil glavne prednosti in slabosti modernih statičnih spletnih strani. V našem primeru smo se odločili za nadgradnjo osnovnih pretvorb XSLT konzorcija TEI. V tretjem poglavju bom predstavil svojo generično nadgradnjo pretvorb XSLT konzorcija TEI in v četrtem poglavju projektno specifične možnosti njegove nadgradnje. V obeh teh poglavjih bom predstavil še različne možnosti dodajanja dinamične vsebine statičnim spletnim stranem. V zaključku bom omenil, kako te statične spletne strani vključujem v digitalno repozitorij portala SIstory in še naše načrte za bodoči razvoj.

2. Moderne statične spletne strani

Sprva so bile vse spletne strani statične, zaradi česar so bile seveda tudi vse digitalne izdaje s področja digitalne humanistike narejene kot statične HTML spletne strani. To je veljalo tudi za slovenske elektronske znanstvenokritične izdaje (Ogrin in Erjavec, 2009),[3] ki so uvajale paradigmo digitalnih izdaj tudi v Sloveniji (Ogrin, 2005). Že kmalu so se ustvarjalci teh digitalnih izdaj srečali z nekaterimi pomanjkljivostmi takšnih statičnih spletnih strani. Predvsem so pogrešali možnost strukturiranega iskanja po besedilu, prilagodljivih parametrov prikaza in dinamičnega povezovanja vsebin. Zato so se pri novih digitalnih izdajah raje odločili uporabiti platformo Fedora Commons (Erjavec et al., 2011).

V tem času so na spletu že dolgo časa kraljevale dinamične spletne strani, ki so uspešno zamenjale zastarele statične spletne strani, pri katerih je bilo možno vsebino spreminjati le tako, da so razvijalci neposredno posegali v HTML kodo. Šele dinamične spletne strani so s pomočjo sistemov za upravljanje vsebin (npr. zelo popularni WordPress, Drupal in Joomla) naposled omogočile tudi tehnično nepodkovanim uporabnikom, da so lahko začeli na spletu objavljati želeno vsebino.

Dinamične spletne strani imajo vsebino shranjeno v bazah podatkov. Strežnik vsebino zgradi šele takrat, ko odjemalec zahteva spletno stran in je kot takšna prilagojena zahtevam uporabnika. Za komuniciranje s strežnikom se uporablja ustrezen programski jezik. Največji problem takšnih dinamičnih spletnih strani je ta, da so njene tehnične rešitve pogosto bolj zapletene od dejanskih potreb uporabnikov.

Moderne statične spletne strani so nastale kot odgovor na težave, ki pestijo dinamične spletne strani. V nasprotju z njimi statične spletne strani nimajo baz podatkov in strežniškega programskega jezika, temveč so zgolj skupek datotek HTML, CSS in JavaScript. Statične spletne strani imajo zato v primerjavi s dinamičnimi številne prednosti (Rinaldi, 2015):

- Zmogljivost: ker nimajo podatkovnih baz in procesiranja s strani strežnika, ni nevarnosti, da bi se takšne strani lahko upočasnile.
- Gostovanje: ker nimajo strežniškega programskega jezika, je gostovanje enostavno in poceni. Obstajajo celo zastonjske možnosti kot so GitHub strani.
- Varnost: nimajo baz podatkov in strežniških programskih jezikov, ki bi jih nekdo lahko izkoristil za računalniške vdore. Zato so takšne strani varne, dokler so datoteke teh strani varno shranjene.
- Vzdrževanje: ker nimajo baz podatkov, strežniškega programskega jezika in sistema za upravljanje vsebin, je njihovo vzdrževanje zelo preprosto.
- Kontrola verzij: ker je celotna statična spletna stran sestavljena samo iz tekstovnih datotek, je vse njene verzije možno dokaj enostavno hraniti v sistemih za kontrolo verzij kot je npr. Git.

Ti razlogi so zlasti pomembni zaradi trajnosti digitalnih izdaj. Uporaba standardnih formatov kot so TIFF in JPEG za digitalne slike, HTML in XML za besedila ipd., ustvarjalcem digitalnih izdaj zagotavlja, da bodo njihove izdaje berljive in uporabne še dolgo časa (Turco, 2016). Zato so to paradigmo začeli poudarjati tudi v sorodnih projektih s področja digitalne humanistike (Viglianti, 2017; Daengeli in Zumsteg, 2017).

Ti razlogi pa so manj prepričljivi, če glede digitalnih izdaj pričakujemo, da bodo vsebovale tudi uporabniško generirano vsebino. Zato statične spletne strani niso primerne za vse digitalne izdaje s področja digitalne humanistike. V mnogih primerih njihove rešitve ne bodo mogle zadovoljiti potreb ustvarjalcev in uporabnikov. Po drugi strani pa je zelo veliko digitalnih projektov, kjer vsebina in njen prikaz nista tako zelo zahtevni. V teh primerih bi bile obstoječe rešitve, ki jih prinašajo statične spletne strani, več kot zadovoljive, predvsem zaradi tega, ker moderne statične strani niso povsem brez možnosti dodajanja dinamičnih vsebin. V resnici so statične spletne strani doživele svojo renesanso šele takrat, ko so se pojavile različne storitve in programske rešitve, ki so tem stranem omogočile dodajanje dinamičnih vsebin.

Modernih statičnih spletnih strani ne kodiramo več ročno, temveč jih generiramo s pomočjo generatorjev statičnih spletnih strani. Izbira teh generatorjev je danes zelo široka. Med najbolj uporabljenimi je Jekyll,[4] ki se ga uporablja tudi pri izdelavi GitHub strani. Zato se je njegova uporaba razširila tudi na humanistiko (Visconti, 2016). Generatorji statičnih spletnih strani domnevajo, da bo uporabnik za pisanje vsebine uporabil besedilne

[1] Web Development Trends in 2018, https://clockwise.software/blog/web-development-trends-in-2018/.

[2] Raziskovalna infrastruktura slovenskega zgodovinopisja, http://www.sistory.si/publikacije/?menuBottom=2.

[3] eZISS, http://nl.ijs.si/e-zrc/.

[4] https://jekyllrb.com/.

formate kot je npr. med razvijalci zelo popularen Markdown.[5] Te formate z generatorjem pretvorimo v HTML spletne strani in jih nato objavimo na spletu. Toda Markdown sintaksa je zelo pomanjkljiva in omogoča le osnovno objavljanje vsebine, zaradi česar je neprimerna za označevanje kompleksnih humanističnih besedil. Zato besedila v humanistiki večinoma kodiramo z XML označevalnim jezikom. XSLT (Extensible Stylesheet Language for Transformation) pa uporabljamo kot orodje za pretvorbo XML. Skupaj tvorita ključni tehnologiji za digitalno humanistiko. (Flanders et al., 2016) Ker je uporaba pretvorb XSLT pogosto zelo podobna pretvorbam v generatorjih statičnih spletnih strani, lahko XSLT kljub njegovi dolgi tradiciji tudi označimo za »moderen, zmogljiv generator statičnih spletnih strani« (Kraetke in Imsieke, 2016).

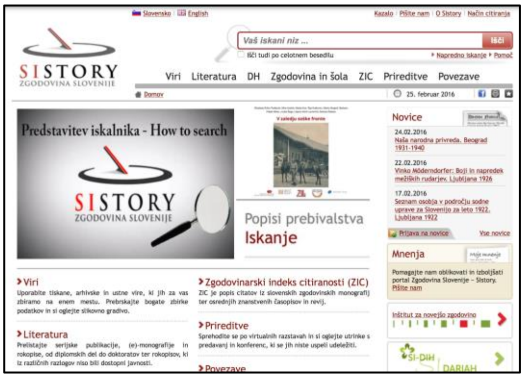
3. SIstory TEI profil

Konzorcij TEI že dolga leta redno vzdržuje in posodablja pretvorne programe XSLT, s katerimi je iz dokumentov TEI mogoče generirati ne le (X)HTML spletne strani, temveč tudi številne ostale formate, mdr. LaTeX, XSL FO, ePub, DOCX in ODT. Ti pretvorni programi so odprto dostopni v GitHub repozitoriju in jih skupaj z novimi verzijami Smernic TEI redno posodabljajo z novimi verzijami.[6] Nimajo samo dobre pisne dokumentacije,[7] temveč je tudi programska koda vzorno komentirana. S pomočjo the pretvornih programov tako med drugim generirajo tudi statične spletne strani vsakokratnih Smernic TEI.[8]

Predvsem pa pretvorbe XSLT konzorcija TEI preko profilov omogočajo zelo fleksibilno prilagoditev glede na potrebe posameznega projekta. Pretvorni programi TEI so bili v resnici napisani z namenom, da se jih lahko čim bolj prilagaja. Obstaja kopica parametrov, ki jih lahko konfiguriramo v skladu s svojimi željami. Stili vsebujejo številne XSLT spremenljivke in predloge, ki jih lahko predelamo v skladu s svojimi potrebami. Avtorji kode so celo pomislili na prazne (hook) predloge, katerim lahko dodamo svojo vsebino in programsko kodo XSLT. Vse te možnosti sem izkoristil pri pisanju SIstory profila za pretvorbe XSLT konzorcija TEI.[9]

Sprva sem pri pisanju teh profilov izhajal iz potreb Raziskovalne infrastrukture slovenskega zgodovinopisja po fleksibilnem in čim bolj sprotnem objavljanju naše tehnične dokumentacije na spletu. V okviru raziskovalne infrastrukture s sodelavci med drugim upravljamo portal Zgodovina Slovenije – SIstory, ki vsebuje tudi repozitorij in digitalno knjižnico. Zato smo se odločili, da te digitalne izdaje po možnosti čim bolj intenzivno vključimo v obstoječo infrastrukturo. Do leta 2016 smo tako statične spletne strani digitalnih izdaj hranili na dodatnem www2 strežniku portala SIstory,[10] v sami digitalni knjižnici smo hranili samo metapodatke o digitalnih izdajah in povezave na te statične spletne strani. Z nadgradnjo portala SIstory v letu 2016 smo HTML in vse druge datoteke teh digitalnih izdaj lahko začeli hraniti neposredno v repozitoriju in digitalni knjižnici.

Zaradi želje po čim večji integraciji digitalnih izdaj v portal SIstory sem tudi zunanji izgled digitalnih izdaj poskusil čim bolj približati uporabniškemu vmesniku portala. Na sliki 1 je kot primer prikazan posnetek vstopne strani portala med leti 2012 in 2016 ter na sliki 2 uporabniški vmesnik digitalne izdaje iz leta 2014.


Slovensko | English
Kazalo | Pišite nam | O SIstory | Način citiranja
Vaš iskani niz ...
Išči
Išči tudi po celotnem besedilu
Napredno iskanje
Pomoč
SISTORY
ZGODOVINA SLOVENIJE
Viri
Literatura
DH
Zgodovina in šola
ZIC
Prireditve
Povezave
Domov
25. februar 2016
Predstavitev iskalnika - How to search
V zaledju soške fronte
Popisi prebivalstva
Iskanje
Novice
24.02.2016
Naša narodna privreda. Beograd 1931-1940
22.02.2016
Vinko Möderndorfer: Boji in napredek mežiških rudarjev. Ljubljana 1926
17.02.2016
Seznam osobja v področju sodne uprave za Slovenijo za leto 1922. Ljubljana 1922
Prijava na novice
Vse novice
Mnenja
Pišite nam
Inštitut za novejšo zgodovino
SI-DIH
DARIAH

Viri

Uporabite tiskane, arhivske in ustne vire, ki jih za vas zbiramo na enem mestu. Prebrskajte bogate zbirke podatkov in si oglejte slikovno gradivo.

Literatura

Prelistajte serijske publikacije, (e)-monografije in rokopise, od diplomskih del do doktoratov ter rokopisov, ki iz različnih razlogov niso bili dostopni javnosti.

Zgodovinarski indeks citiranosti (ZIC)

ZIC je popis citatov iz slovenskih zgodovinskih monografij ter osrednjih znanstvenih časopisov in revij.

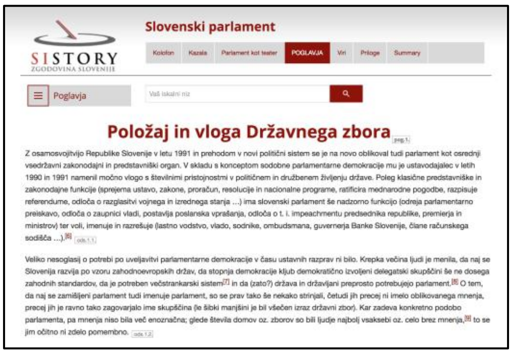
Prireditve

Sprehodite se po virtualnih razstavah in si oglejte utrinke s predavanj in konferenc, ki se jih niste uspeli udeležiti.

Povezave

Pomagajte nam oblikovati in izboljšati portal Zgodovina Slovenije – Sistory.

Slika 1: Vhodna spletna stran portala SIstory iz leta 2016



Položaj in vloga Državnega zbora pog.1.

Z osamosvojitvijo Republike Slovenije v letu 1991 in prehodom v novi politični sistem se je na novo oblikoval tudi parlament kot osrednji vsedržavni zakonodajni in predstavniški organ. V skladu s konceptom sodobne parlamentarne demokracije mu je ustavodajalec v letih 1990 in 1991 namenil močno vlogo s številnimi pristojnostmi v političnem in družbenem življenju države. Poleg klasične predstavniške in zakonodajne funkcije (sprejema ustavo, zakone, proračun, resolucije in nacionalne programe, ratificira mednarodne pogodbe, razpisuje referendume, odloča o razglasitvi vojnega in izrednega stanja ...) ima slovenski parlament še nadzorno funkcijo (odreja parlamentarno preiskavo, odloča o zaupnici vladi, postavlja poslanska vprašanja, odloča o t. i. impeachmentu predsednika republike, premierja in ministrov) ter voli, imenuje in razrešuje (lastno vodstvo, vlado, sodnike, ombudsmana, guvernerja Banke Slovenije, člane računskega sodišča ...).[6] ods.1.1.

Veliko nesoglasij o potrebi po uveljavitvi parlamentarne demokracije v času ustavnih razprav ni bilo. Krepka večina ljudi je menila, da naj se Slovenija razvija po vzoru zahodnoevropskih držav, da stopnja demokracije kljub demokratično izvoljeni delegatski skupščini še ne dosega zahodnih standardov, da je potreben večstrankarski sistem[7] in da (zato?) država in državljani preprosto potrebujejo parlament.[8] O tem, da naj se zamišljeni parlament tudi imenuje parlament, so se prav tako še nekako strinjali, četudi jih precej ni imelo oblikovanega mnenja, precej jih je ravno tako zagovarjalo ime skupščina (le šibki manjšini je bil všečen izraz državni zbor). Kar zadeva konkretno podobo parlamenta, pa mnenja niso bila več enoznačna; glede števila domov oz. zborov so bili ljudje najbolj vsaksebi oz. celo brez mnenja,[9] to se jim očitno ni zdelo pomembno. ods.1.2.

Slika 2: Uporabniški vmesnik digitalne izdaje iz leta 2014

Čeprav so barvni toni povsem isti in se tudi postavitve logotipa, iskalne vrstice, glavne zgornje navigacije in vsebine zelo dobro zgledujejo po portalu SIstory, uporabniška vmesnika le nista povsem ista. Uporabniški vmesnik takratnega portala je bil namreč še vedno zgrajen na osnovi stare HTML 4 tehnologije, medtem ko sem pri digitalnih izdajah že začel uporabljati odziven dizajn spletnih strani in HTML 5. Pri tem sem se odločil za uporabo ogrodja za odzivne uporabniške vmesnike ZURB Foundation.[11] Moje prilagoditve in dodatke CSS in JS hranim v GitHub repozitoriju.[12] Ker se je uporaba tega ogrodja izkazala za zelo koristno, smo ga leta 2016 vključili tudi v prenovljen portal SIstory. Novemu izgledu portala sem nato prilagodil še izgled digitalnih izdaj (primerjaj sliki 3 in 4).

[5] http://www.daringfireball.net/projects/markdown/.

[6] TEI XSL Stylesheets, https://github.com/TEIC/Stylesheets.

[7] XSL stylesheets for TEI XML, http://www.tei-c.org/release/doc/tei-xsl/.

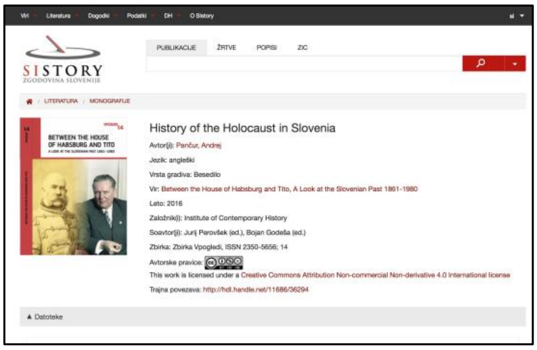
[8] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

[9] SIstory TEI XSL Stylesheets, https://github.com/SIstory/Stylesheets.

[10] http://www2.sistory.si/.

[11] Foundation, https://foundation.zurb.com/.

[12] SIstory themes, https://github.com/SIstory/themes.



History of the Holocaust in Slovenia

Avtor(ji): Pančur, Andrej

Jezik: angleški

Vrsta gradiva: Besedilo

Vir: Between the House of Habsburg and Tito, A Look at the Slovenian Past 1861-1980

Leto: 2016

Založnik(i): Institute of Contemporary History

Soavtor(ji): Jurij Perovšek (ed.), Bojan Godeša (ed.)

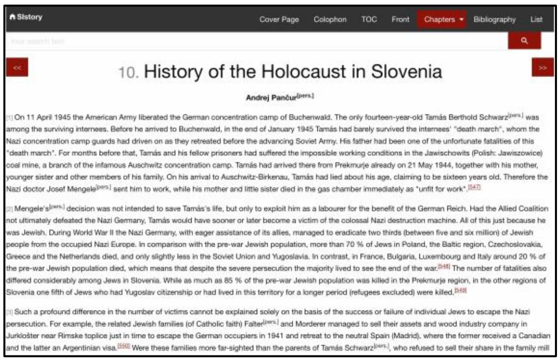
Zbirka: Zbirka Vpogledi, ISSN 2350-5656; 14

Avtorske pravice:

This work is licensed under a Creative Commons Attribution Non-commercial Non-derivative 4.0 International license

Trajna povezava: http://hdl.handle.net/11686/36294

Slika 3: Zgornja navigacija, iskalna vrstica in metapodatkovna stran portala SIstory



10. History of the Holocaust in Slovenia

Andrej Pančur[pers.]

[1] On 11 April 1945 the American Army liberated the German concentration camp of Buchenwald. The only fourteen-year-old Tamás Berthold Schwarz[pers.] was among the surviving internees. Before he arrived to Buchenwald, in the end of January 1945 Tamás had barely survived the internees' "death march", whom the Nazi concentration camp guards had driven on as they retreated before the advancing Soviet Army. His father had been one of the unfortunate fatalities of this "death march". For months before that, Tamás and his fellow prisoners had suffered the impossible working conditions in the Jawischowits (Polish: Jawiszowice) coal mine, a branch of the infamous Auschwitz concentration camp. Tamás had arrived there from Prekmurje already on 21 May 1944, together with his mother, younger sister and other members of his family. On his arrival to Auschwitz-Birkenau, Tamás had lied about his age, claiming to be sixteen years old. Therefore the Nazi doctor Josef Mengele[pers.] sent him to work, while his mother and little sister died in the gas chamber immediately as "unfit for work".[547]

[2] Mengele's[pers.] decision was not intended to save Tamás's life, but only to exploit him as a labourer for the benefit of the German Reich. Had the Allied Coalition not ultimately defeated the Nazi Germany, Tamás would have sooner or later become a victim of the colossal Nazi destruction machine. All of this just because he was Jewish. During World War II the Nazi Germany, with eager assistance of its allies, managed to eradicate two thirds (between five and six million) of Jewish people from the occupied Nazi Europe. In comparison with the pre-war Jewish population, more than 70 % of Jews in Poland, the Baltic region, Czechoslovakia, Greece and the Netherlands died, and only slightly less in the Soviet Union and Yugoslavia. In contrast, in France, Bulgaria, Luxembourg and Italy around 20 % of the pre-war Jewish population died, which means that despite the severe persecution the majority lived to see the end of the war.[548] The number of fatalities also differed considerably among Jews in Slovenia. While as much as 85 % of the pre-war Jewish population was killed in the Prekmurje region, in the other regions of Slovenia one fifth of Jews who had Yugoslav citizenship or had lived in this territory for a longer period (refugees excluded) were killed.[549]

[3] Such a profound difference in the number of victims cannot be explained solely on the basis of the success or failure of individual Jews to escape the Nazi persecution. For example, the related Jewish families (of Catholic faith) Falter[pers.] and Morderer managed to sell their assets and wood industry company in Jurklošter near Rimske toplice just in time to escape the German occupiers in 1941 and retreat to the neutral Spain (Madrid), where the former received a Canadian and the latter an Argentinian visa.[550] Were these families more far-sighted than the parents of Tamás Schwarz[pers.], who refused to sell their share in the family mill

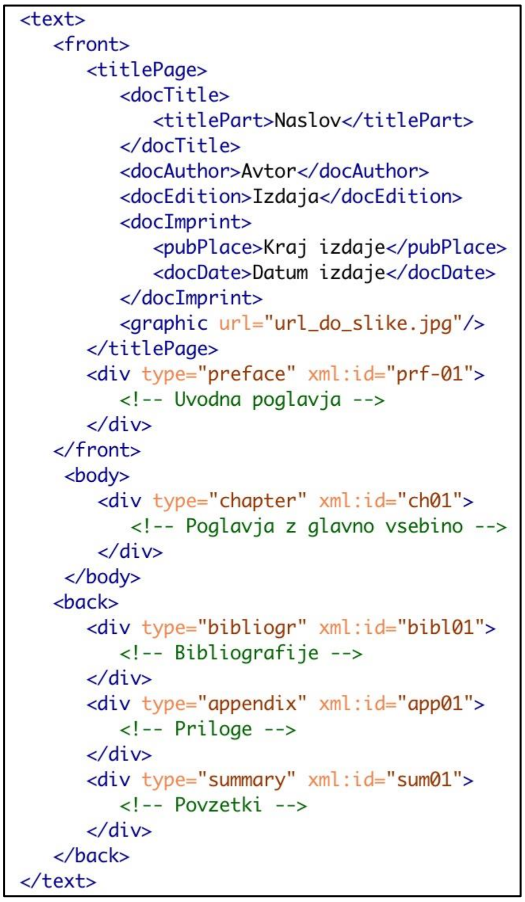
Slika 4: Uporabniški vmesnik digitalne izdaje iz leta 2016

Poleg prvotno načrtovane tehnične dokumentacije smo na spletu v formatu HTML kmalu začeli objavljati še druge vrste publikacij, predvsem monografije, zbornike in revije. Zato sem SIstory TEI profil konfiguriral v skladu s potrebami po objavljanju takšnih vrst digitalnih izdaj. Profil omogoča pretvorbo:

- posameznega TEI dokumenta;
- več TEI dokumentov iz skupnega TEI korpusa. V tem primeru je potrebno vsak TEI dokument pretvoriti posebej. Posebej je potrebno pretvoriti še sam TEI korpus in njegov <teiHeader>, saj na ta način generiramo skupno naslovnico, kolofon in kazala vsebine.

Glavna menijska navigacija po digitalni izdaji se nahaja povsem na vrhu spletne strani kot horizontalna navigacija s spustnim menijem. Struktura te navigacije odraža strukturo ter sklope in razdelke posameznih dokumentov TEI. V nadaljevanju bom na kratko predstavil možne vsebinske sklope tako navigacije kot dokumenta TEI. V praksi seveda noben dokument nima prav vseh teh razdelkov, temveč si jih avtor dokumentov TEI oblikuje povsem v skladu s svojimi potrebami.

Osrednji del vsebine je vedno znotraj <body> elementa. Glavna vsebina mora biti nujno v enem ali več <div> elementih, ki morajo obvezno imeti atribut @xml:id. Vsak tak <div> predstavlja svoj razdelek vsebine oziroma poglavja. Zato so v navigaciji vsi <div> znotraj <body> prikazani v enem spustnem meniju. Pred in za tem spustnim menijem je lahko dostopna še različna ostala vsebina, ki je v TEI dokumentu kodirana znotraj <front> in <back> elementov. V sliki 5 so tako prikazani vsi ti glavni vsebinski sklopi.



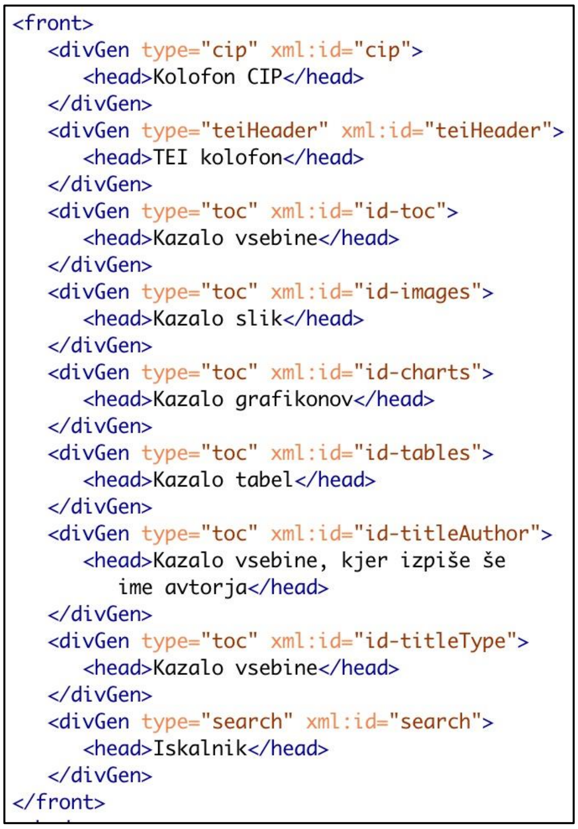
```
<text>
   <front>
      <titlePage>
         <docTitle>
            <titlePart>Naslov</titlePart>
         </docTitle>
         <docAuthor>Avtor</docAuthor>
         <docEdition>Izdaja</docEdition>
         <docImprint>
            <pubPlace>Kraj izdaje</pubPlace>
            <docDate>Datum izdaje</docDate>
         </docImprint>
         <graphic url="url_do_slike.jpg"/>
      </titlePage>
      <div type="preface" xml:id="prf-01">
         <!-- Uvodna poglavja -->
      </div>
   </front>
    <body>
       <div type="chapter" xml:id="ch01">
          <!-- Poglavja z glavno vsebino -->
       </div>
    </body>
   <back>
      <div type="bibliogr" xml:id="bibl01">
         <!-- Bibliografije -->
      </div>
      <div type="appendix" xml:id="app01">
         <!-- Priloge -->
      </div>
      <div type="summary" xml:id="sum01">
         <!-- Povzetki -->
      </div>
   </back>
</text>
```

Slika 5: Glavni vsebinski sklopi dokumenta TEI

Med njimi je obvezen le <titlePage>, ki je izhodiščni index.html element in je kot takšen v navigaciji dostopen na prvem mestu kot Naslovnica. Znotraj <front> se lahko nahajajo eden ali več <div> elementov, ki v navigaciji predstavljajo sklop uvodnih poglavij. Znotraj <back> elementa imamo tri možne vsebinske sklope (bibliografije, priloge, povzetki), zaradi česar je nujno, da imajo vedno ustrezen atribut @type. Vsak ta sklop ima lahko enega ali več poglavij. V večini primerov poteka pretvorba vsebine teh razdelkov na podlagi standardnih pretvorb XSLT konzorcija TEI, ki sem jih sem za potrebe naših digitalnih izdaj le delno prilagodil. Povsem na novo sem napisal pretvorbe za generirane razdelke <divGen>. Vsi so vključeni v SIstory TEI profil. Ti generirani razdelki so lahko vključeni v <front> (Slika 6) ali <back> (Slika 7). Vsak <divGen> mora vsebovati <head> s poljubnim naslovom razdelka. Ti naslovi so nato vključeni v navigacijo spletne izdaje.

Za razliko od zgoraj omenjenimi razdelkov <div>, kjer so identifikatorji @xml:id samo priporočljivi (HTML datoteke teh razdelkov dobijo imena po teh identifikatorjih), so pri generiranih razdelkih nujni in imajo tudi semantičen pomen, ki je ključen za njihovo pretvorbo. Atribut @type opredeljuje glavno kategorijo, ki je v horizontalni navigaciji posebej izpostavljena. Atribut @xml:id bolj natančno opredeljuje podkategorijo, ki je v navigaciji prikazana v spustnem meniju. Najobsežnejša kategorija je skupina kazal vsebine (Table of Contents TOC), ki poleg različnih kazal vsebine poglavij in podpoglavij vsebuje še kazalo tabel, slik in grafikonov. Kazalo grafikonov je v resnici posebna skupina kazal slik

(<figure>), ki zajema slike z atributom @type in vrednostjo chart.



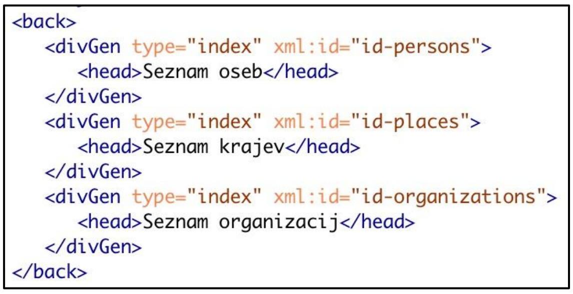
```
<front>
   <divGen type="cip" xml:id="cip">
      <head>Kolofon CIP</head>
   </divGen>
   <divGen type="teiHeader" xml:id="teiHeader">
      <head>TEI kolofon</head>
   </divGen>
   <divGen type="toc" xml:id="id-toc">
      <head>Kazalo vsebine</head>
   </divGen>
   <divGen type="toc" xml:id="id-images">
      <head>Kazalo slik</head>
   </divGen>
   <divGen type="toc" xml:id="id-charts">
      <head>Kazalo grafikonov</head>
   </divGen>
   <divGen type="toc" xml:id="id-tables">
      <head>Kazalo tabel</head>
   </divGen>
   <divGen type="toc" xml:id="id-titleAuthor">
      <head>Kazalo vsebine, kjer izpiše še
         ime avtorja</head>
   </divGen>
   <divGen type="toc" xml:id="id-titleType">
      <head>Kazalo vsebine</head>
   </divGen>
   <divGen type="search" xml:id="search">
      <head>Iskalnik</head>
   </divGen>
</front>
```

Slika 6: Seznam vseh možnih generiranih razdelkov <divGen> v <front>

V okviru elementa <back> se nahaja samo ena kategorija generiranih razdelkov, ki zajema različne sezname oziroma indekse oseb, krajev in organizacij. Generirani razdelki zajemajo vse osebe iz dokumenta TEI kodirane z elementov <persName> ali vse kraje <placeName> ali vse organizacije <orgName>. Vse tako kodirane imenske entitete morajo imeti atribut @ref, preko katerega se sklicujejo na ustrezni kanonični element v seznamu entitet (<listPerson> za osebe, <listOrg> za organizacije in <listPlace> za kraje) v glavi dokumenta TEI (<teiHeader>). Atribut @ref elementa <placeName> lahko vsebuje tudi sklic na GeoNames[13] ali DBpedia[14] URI, kjer SIstory profil poišče geografske koordinate, ki jih nato prikaže v seznamu krajev.

Ker je s SIstory profilom mogoče pretvarjati tudi dokumente TEI iz TEI korpusa, <divGen> iz različnih dokumentov TEI ne morejo imeti istih @xml:id identifikatorjev. Zato se podkategorije generiranih razdelkov določi tako, da je identifikator podkategorije zapisan za zadnjim pomišljajem vrednosti tega identifikatorja (glej sliko 6 in 7, kjer vrednost id pred pomišljajem v @xml:id opredeljuje poljubni identifikator, za pomišljajem pa podkategorijo).

SIstory profil omogoča tudi prikaz dinamične vsebine. Kot osnovna funkcionalnost je vključen iskalnik Tipue Search.[15] Vključimo ga z generiranim razdelkom tipa search v <front>. Tipue Search je odprtokodni jQuery vtičnik, ki ga je mogoče relativno enostavno vključiti tudi v statične spletne strani. Na grafičnem vmesniku je iskalna vrstica postavljena takoj pod spodnjo navigacijo, <divGen> iskalnika pa generira search.html spletno stran, ki vključuje dinamičen prikaz rezultatov iskanja. Vsebina dokumenta TEI je kot JavaScript objekt (JSON) indeksirana v datoteki tipuesearch_content.js, ki mora biti v istem direktoriju kot datoteka search.html. Indeksacija vsebine poteka na ravni odstavkov <p>, seznamov <list>, tabel <table>, slik <figure> in vseh drugih možnih elementov TEI, ki so neposredni child elementi razdelka <div>. Zato morajo vsi ti elementi imeti identifikator @xml:id. Edina izjema so seznami. Če le-ti nimajo atributa @xml:id, imajo pa jih njihovi child elementi, potem so indeksirani slednji.



```
<back>
   <divGen type="index" xml:id="id-persons">
      <head>Seznam oseb</head>
   </divGen>
   <divGen type="index" xml:id="id-places">
      <head>Seznam krajev</head>
   </divGen>
   <divGen type="index" xml:id="id-organizations">
      <head>Seznam organizacij</head>
   </divGen>
</back>
```

Slika 7: Seznam vseh možnih generiranih razdelkov <divGen> v <back>

4. Konfiguracija in nadgradnja SIstory profila

Tako kot glavni pretvorni programi XSLT konzorcija TEI je tudi SIstory profil narejen z namenom, da ga je mogoče prilagoditi potrebam posameznega projekta. V ta namen vključuje nekaj izvornih parametrov pretvorb XSLT konzorcija TEI, ki sem jim dodal še nekaj novih parametrov. Vse te parametre je sicer mogoče na novo nastavljati ob vsaki pretvorbi, vendar je bolj priporočljivo, da za vsak projekt naredimo nov projektni profil. Običajno pretvorba poteka tako, da projektni XSLT profil vključi SIstory XSLT profil, ta pa vključi TEI pretvorbe XSLT.

Privzeti SIstory profil tako npr. predvideva, da bo pri pretvorbi vsako poglavje oziroma prvi razdelek <div> samostojna HTML spletna stran. V tem primeru se na spletnih straneh avtomatično doda navigacija z gumbi naprej in nazaj. Za razliko od prvotnih pretvorb TEI ta navigacija vključuje tudi generirane razdelke <divGen>. Toda s spremembo parametra splitLevel (izvorno parameter pretvorb TEI) lahko določimo, da so tudi podpoglavja ločene HTML spletne strani. Temu primerno je sedaj prilagojena tudi navigacija naprej/nazaj, navzgor/navzdol med spletnimi stranmi. Trenutno SIstory profil podpira le globino treh razdelkov.

S parametrom documentationLanguage je trenutno mogoče nastaviti slovensko, angleško in srbsko navigacijo (latinica ali cirilica). Z dodajanjem novih prevodov v dokument myi18n.xml je mogoče to lokalizacijo še razširiti. Ustrezno je poskrbljeno tudi za lokalizacijo iskalnika Tipue Search.

SIstory profil omogoča še vzporedno objavo različnih jezikovnih verzij besedila. V tem primeru morajo imeti vsi glavni razdelki <div> in generirani razdelki <divGen> atributa @xml:lang z ustrezno jezikovno kodo in @corresp s kazalko na vse drugojezične verzije tega

[13] GeoNames, http://www.geonames.org/.

[14] DBpedia, http://wiki.dbpedia.org/.

[15] Tipue Search, http://www.tipue.com/search/.

besedila. Obenem je potrebno še nastaviti parameter languages-locale na vrednost true in s parametrom languages-locale-primary določiti jezikovno kodo izhodiščne datoteke index.html.

Podobno prilagodljiv je tudi prikaz vseh metapodatkov dokumenta TEI iz <teiHeader>. To pretvorbo najprej določimo s vključitvijo generičnega razdelka tipa teiHeader (glej sliko 6). Celotna vsebina <teiHeader> je pretvorjena v HTML elemente seznam definicij <dl>, pri katerem oznaka definicije <dt> opredeljuje ime TEI elementa ter imena in vrednosti atributov (element [atribut = vrednost | atribut = vrednost]), definicija elementa <dd> pa vsebino besedila elementa. Definicije so seveda temu primerno tudi gnezdene. S dodanimi parametri lahko to pretvorbo konfiguriramo tako, da namesto imen elementov in atributov izpiše njihova opisna imena v angleščini ali slovenščini.

Poleg te preproste konfiguracije SIstory profila lahko pri projektni pretvorbi vključimo še kakršno koli dodatno pretvorbo XSLT, ki jo tako povsem prilagodimo potrebam digitalne izdaje. Obenem lahko z vključitvijo različnih JavaScript knjižnic in vtičnikov ter spletnih aplikacij omogočimo še dodatno dinamično prikazovanje vsebine. V primeru digitalnih izdaj na portalu SIstory sem npr. za prikazovanje večje količine tabelarnih podatkov uspešno uporabil DataTables,[16] za grafikone Highcharts,[17] za zemljevide Google Maps, za slike ImageViewer.[18] In to so samo nekateri primeri, ki imajo tudi različne alternative. Vsako leto pa se jim pridružijo še številne nove možnosti.

Obenem so se leta 2017 z objavo Saxon-JS[19] še dodatno izboljšale možnosti dinamičnega prikazovanja vsebine XML dokumentov v statičnih spletnih straneh. To možnost sem že uspešno uporabil pri spletnih seznamih, kjer sem prikaz celotne vsebine seznama filtriral glede na želeni parameter (identifikator).

5. Zaključek

Privzeta pretvorba SIstory profila generira vse HTML, JS in morebitne ostale datoteke v isti direktorij. Ker je tako generirana digitalna izdaja sestavljena zgolj iz statičnih spletnih strani, jo lahko uporabljamo tudi na osebnem računalniku. Na ta način je možno digitalno izdajo tudi učinkovito testirati še pred objavo na spletu, kjer jo lahko hitro in enostavno objavimo na poljubnem dostopnem strežniku. Kot zastonjska možnost obstajajo tudi spletne strani GitHub repozitorija. Na ta način zagotovimo še učinkovito kontrolo verzij.

Toda glavni namen SIstory profilov je vključitev digitalnih izdaj neposredno v repozitorij portala SIstory in njegovo digitalno knjižnico. Na ta način lahko zagotovimo ne le učinkovito hrambo vseh datotek digitalne izdaje, skupaj z dodajanjem unikatnih handle identifikatorjev in kotrolne vsote (checksum) za vse datoteke, temveč tudi fleksibilno razvrščanje digitalne izdaje v enega ali več digitalnih objektov, z enim ali več intelektualnih entitet. Vsaka intelektualna entiteta ima svoj Handle identifikator in svoje metapodatke. Vključuje lahko nič ali več datotek. Datoteke posamezne intelektualne entitete se nahajajo v istem direktoriju. Pot do direktorija vsebuje tudi predpono (suffix) Handle identifikatorja, ki je v primeru portala SIstory vedno številčna vrednost (npr. za predpono 555 je relativna pot /cdn/publikacije/1-1000/555/datoteka). SIstory XSLT profil mora zato nujno že vnaprej natančno vedeti, kakšne so vrednosti teh identifikatorjev. Tako lahko tudi vnaprej natančno določimo, ali bo celotna vsebina digitalne izdaje samo v eni intelektualni entiteti portala SIstory ali pa bodo različne datoteke digitalne izdaje vključene v različne intelektualne entitete. Te identifikatorje zapišemo med ostale metapodatke v <teiHeader> in sicer v okviru <publicationStmt> kot vrednost enega ali več elementov <idno>. Ta element mora imeti vrednost atributa @type sistory, atribut @corresp pa mora imeti kazalke na vse ustrezne razdelke <div> in <divGen>, katerih vsebina bo vključena v intelektualno entiteto s tem identifikatorjem.

SIstory XSLT profil je pod odprtimi pogoji dostopen v GitHub repozitoriju.[20] V drugem GitHub repozitoriju so dostopne še vse digitalne izdaje, ki se nahajajo na portalu SIstory. Za vsako teh izdaj so dostopne še projektne nadgradnje SIstory XSLT profila.[21] Profil redno dopolnjujem in seveda tudi vzdržujem v skladu s spremembami pretvornih programov XSLT konzorcija TEI. V letošnjem letu načrtujem še obsežnejšo nadgradnjo. Trenutni portal SIstory bomo namreč v naslednjem letu zamenjali z novim repozitorijem, ki ga pravkar intenzivno razvijamo. Novi repozitorij bo v okviru METS aplikacijskega profila omogočil fleksibilno dodajanje različnih vrst metapodatkov, med drugim tudi teiHeader. S tem bomo občutno izboljšali metapodatkovno opremljenost digitalnih izdaj na portalu SIstory.

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.

7. Literatura

Peter Andorfer, Matej Ďurčo, Thomas Stäcker, Christian Thomas, Vera Hildenbrandt, Hubert Stigler, Sibylle Söring in Lukas Rosenthaler. 2016. Nachhaltigkeit technischer Lösungen für digitale Editionen: Eine kritische Evaluation bestehender Frameworks und Workflows von und für Praktiker_innen. V: *DHd 2016: Modellierung – Vernetzung – Visualisierun: Die Digital Humanities als fächerübergreifendes Forschungsparadigma: Konferenzabstracts*, str. 36-39. Universität Leipzig. http://www.dhd2016.de/.

Tara Andrews in Joris van Zundert. 2016. What Are You Trying to Say? The Interface as an Integral Element of Argument. V: *Digital Scholarly Editions as Interfaces*, International Symposium at the University of Graz, Austria, str. 31-32. Centre for Information Modelling – Austrian Centre for Digital Humanities. https://static.uni-graz.at/fileadmin/gewi-zentren/Informationsmodellierung/PDF/dse-interfaces_BoA21092016.pdf.

[16] DataTables: Table plug-in for jQuery, https://datatables.net/.

[17] Highcharts, https://www.highcharts.com/products/highcharts/.

[18] ImageViewer, http://ignitersworld.com/lab/imageViewer.html.

[19] Saxon-JS, http://www.saxonica.com/saxon-js/index.xml.

[20] SIstory TEI XSL Stylesheets, https://github.com/SIstory/Stylesheets.

[21] https://github.com/SIstory/publications.

Peter Daengeli in Simon Zumsteg. 2017. Hermann Burgers Lokalbericht: Hybrid-Edition mit digitalem Schwerpunkt. V: *DHd 2017: Digitale Nachhaltigkeit: Konferenzabstracts*. Universität Bern. str. 151-155. http://www.dhd2017.ch/.

DHd-AG Datenzentren. 2017. *Geisteswissenschaftliche Datenzentren im deutschsprachigen Raum*, Grundsatzpapier zur Sicherung der langfristigen Verfügbarkeit von Forschungsdaten. Hamburg. DOI: 10.5281/zenodo.1134760

Tomaž Erjavec, Jan Jona Javoršek, Matija Ogrin in Petra Vide Ogrin. 2011. Od biografskega leksikona do znanstvenokritične izdaje: vprašanje trajnosti elektronskih besedil. *Knjižnica*, 55(1): 103-114. https://knjiznica.zbds-zveza.si/knjiznica/article/view/6004.

Martin Fechner. 2018. Eine nachhaltige Präsentationsschicht für digitale Editionen. V: Georg Vogeler (ur.), *DHd 2018: Kritik der digitalen Vernunft: Konferenzabstracts*, str. 203-207. Universität zu Köln. http://dhd2018.uni-koeln.de/.

Julia Flanders, Syd Bauman in Sarah Connell. 2016. XSLT: Transforming our XML data. V: C. Crompton, R. J. Lane in R. Siemens (ur.), *Doing Digital Humanities: Practice, Training, Research*, str. 255-272. Oxon in New York, Routledge.

Martin Kraetke in Gerrit Imsieke. 2016. XSLT as a Modern, Powerful Static Website Generator: Publishing Hogrefe's Clinical Handbook of Psychotropic Drugs as a Web App. V: *Proceedings of XML in, Web Out: International Symposium on sub rosa XML*, Balisage Series on Markup Technologies, vol. 18. https://doi.org/10.4242/BalisageVol18.Kraetke02.

Katrin Moeller, Matej Ďurčo, Barbara Ebert, Marina Lemaire, Lukas Rosenthaler, Patrick Sahle, Urlike Wuttke in Jörg Wettlaufer. 2018. Die Summe geisteswissenschaftlicher Methoden? Fachspezifisches Datenmanagement als Voraussetzung zukunftsorientierten Forschens. V: Georg Vogeler (ur.), *DHd 2018: Kritik der digitalen Vernunft: Konferenzabstracts*, str. 89-93. Universität zu Köln. http://dhd2018.uni-koeln.de/.

Matija Ogrin in Tomaž Erjavec. 2009. Ekdotika in tehnologija: Elektronske znanstvenokritične izdaje slovenskega slovstva. *Jezik in slovstvo*, 54(6): 57-72. http://www.dlib.si/?URN=URN:NBN:SI:doc-BOC8BANS.

Matija Ogrin (ur.). 2005. *Znanstvene razprave in elektronski mediji: razprave*. Ljubljana: Založba ZRC, ZRC SAZU. http://nl.ijs.si/e-zrc/bib/eziss-knjiga.pdf.

Brian Rinaldi. 2015. *Static Site Generators: Modern Tools for Static Website Development*. Sebastopol, CA: O'Reilly Media.

Peter Robinson. 2016. Why Interfaces Do Not and Should Not Matter for Scholarly Digital Editions. V: *Digital Scholarly Editions as Interfaces*, International Symposium at the University of Graz, Austria, str. 29-30. Centre for Information Modelling – Austrian Centre for Digital Humanities. https://static.uni-graz.at/fileadmin/gewi-zentren/Informationsmodellierung/PDF/dse-interfaces_BoA21092016.pdf.

Laurent Romary, Piotr Banski, Jack Bowers, Emiliano Degl'innocenti, Matej Ďurčo, Roberta Giacomi, Klaus Illmayer, Adeline Joffres, Fahad Khan, Mohamed Khemakhem, et al. 2017. *Report on Standardization (draft).* [Technical report] 4.2 Inria. https://hal.inria.fr/hal-01560563.

Jennifer Schaffner in Ricky Erway. 2014. *Does Every Research Library Need a Digital Humanities Center?* Dublin, Ohio: OCLC Research. https://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-digital-humanities-center-2014.pdf.

TEI Consortium, ur. 2018 TEI P5: *Guidelines for Electronic Text Encoding and Interchange 3.3.0.* TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Roberto Rosselli Del Turco. 2016. The Battle We Forgot to Fight: Should We Make a Case for Digital Editions? V: Matthew James Driscoll in Elena Pierazzo (ur.), *Digital Scholarly Editing: Theories and Practices*, str. 219-238. Cambridge, UK: Open Book Publishers. http://dx.doi.org/10.11647/OBP.0095.

Magdalena Turska, James Cummings in Sebastian Rahtz. 2016. Challenging the Myth of Presenation in Digital Editions. *Journal of the Text Encoding Initiative,* (9). DOI: 10.4000/jtei.1453

Raffaele Viglianti. 2017. Your own Shelley-Godwin Archive: An off-line strategy for an onile publication (poster). V: *TEI 2017 Victoria*. . https://hcmc.uvic.ca/tei2017/abstracts/t_126_viglianti_shelleygodwin.html.

Amanda Visconti. 2016. Building a static website with Jekyll and GitHub Pages. *The Programming Historian,* 5. https://programminghistorian.org/lessons/building-static-sites-with-jekyll-github-pages.

Spregledana kulturna dediščina in uporaba digitalne raziskovalne infrastrukture za humanistiko v raziskavi Odlivanje smrti

Andrej Pančur,* Alenka Pirman,† Maruša Kocjančič‡

* Inštitut za novejšo zgodovino
Kongresni trg 1, 1000 Ljubljana
andrej.pancur@inz.si
† Društvo za domače raziskave
Šarhova 34, 1000 Ljubljana
alenka.pirman@gmail.com
‡ Društvo za domače raziskave
Šarhova 34, 1000 Ljubljana
marusa.koc@gmail.com

Povzetek

V prispevku je predstavljeno sodelovanje med raziskovalci projekta Odlivanje smrti (TRACES) in Raziskovalne infrastrukture slovenskega zgodovinopisja pri uporabi kulturne dediščine iz različnih ustanov za varstvo kulturne dediščine (GLAM – galerije, knjižnice, arhivi, muzeji) v raziskovalne namene. Sodelovanje je potekalo v skladu z življenjskim ciklom raziskovalnih podatkov, ki je bil povsem prilagojen potrebam raziskave. Največji izziv je predstavljala standardizacija. Pri vključitvi zbirke digitalnih objektov (posmrtne maske) v portal Zgodovina Slovenije – SIstory se je uporabil Dublin Core aplikacijski profil. Pri izdelavi digitalne izdaje smo uporabili TEI in LIDO. S pomočjo javne predstavitve vmesnih rezultatov projekta smo uspešno začeli zbirati še dodatne objekte kulturne dediščine.

Overlooked cultural heritage and the use of digital research infrastructure for humanities in the research action Casting of Death

The contribution presents the cooperation between the researchers of the Casting of Death project (TRACES) and the Research Infrastructure of Slovenian Historiography with regard to the research use of cultural heritage from various GLAM cultural heritage institutions (galleries, libraries, archives, and museums). The cooperation was conducted in accordance with the research data life cycle, which was completely adapted to the requirements of the research. Standardisation represented the greatest challenge. The Dublin Core application profile was used for the inclusion of the collection of digital objects (death masks) in the History of Slovenia – SIstory portal. The TEI and LIDO were used to make the digital edition. The public presentation of the interim project results has allowed us to start collecting additional cultural heritage objects.

1. Uvod

Raziskave v humanistiki in umetnosti večinoma temeljijo na analizah različnih sledi človekovega delovanja, ki jih hranijo ustanove s področja varstva kulturne dediščine kot so galerije, knjižnice, muzeji in arhivi (Seillier et al., 2017). V tem oziru predstavlja dostop do kulturne dediščine velik izziv bodočega uspešnega razvoja digitalne humanistike. Kakovostni podatki in metapodatki o kulturni dediščini so nujni predpogoj za izvajanje zanesljivih, uspešnih in preverljivih raziskav na številnih področjih humanistike in umetnosti (Baillot et al., 2017).

Zato je Digitalna raziskovalna infrastruktura za umetnost in humanistiko DARIAH[1] leta 2016 sprožila pobudo (Baillot et al., 2016) za razvoj listine o ponovni uporabi podatkov kulturne dediščine (Cultural Heritage Data Reuse Charter), ki so se ji kmalu pridružile še ostale evropske organizacije (APEF,[2] CLARIN,[3] Europeana,[4] E-RIHS[5]) in projekti (Iperion-CH,[6] PARTHENOS[7]).

Ta pobuda namerava vzpostaviti načela in mehanizme za uporabo in ponovno uporabo podatkov o kulturni dediščini v raziskovalne namene. Pri tem priporoča, da tako raziskovalci kot ustanove, ki hranijo kulturno dediščino, upoštevajo sledeča splošna načela:[8]

- recipročnost: obe strani dajeta druga drugi na razpolago svoje podatke in raziskovalne rezultate;
- interoperabilnost: to vsebino dajeta na razpolago v skladu z mednarodnimi standardi in interoperabilnimi protokoli;

[1] Digital Research Infrastructure for the Arts and Humanities, https://www.dariah.eu/.

[2] Archives Portal Europe Foundation, http://www.archivesportaleuropefoundation.eu/.

[3] European Research Infrastructure for Language Resources and Technology, https://www.clarin.eu/.

[4] https://www.europeana.eu.

[5] European Research Infrastructure for Heritage Science, http://www.erihs.fr/.

[6] Integrated Platform for the European Research Infrastructure ON Cultural Heritage, http://www.iperionch.eu/.

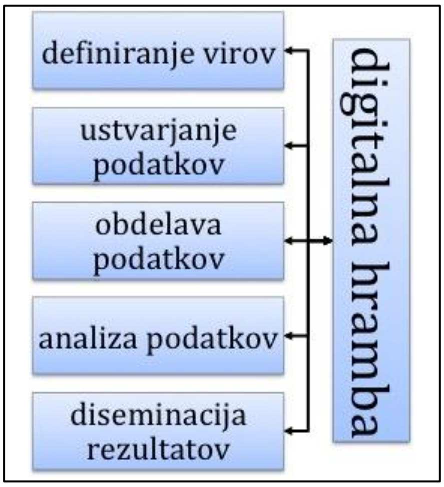
[7] Pooling Activities, Resources and Tools for Heritage E-research Networking Optimization and Synergies, http://www.parthenos-project.eu/.

[8] Cultural Heritage Data Reuse Charter: Mission Statement, https://sondages.inria.fr/index.php/593568/lang-en.

- odprtost: po možnosti naj bodo dostopni pod odprtimi pogoji;
- skrbništvo: poskrbi naj se za dolgoročno hrambo in dostop do vseh verzij podatkov in rezultatov;
- zanesljivost: jasno naj bo razviden njihov izvor, dokumentacija, tehnologija, procedure, protokoli, celovitost;
- citiranost: poskrbeti je potrebno za njihovo citiranost.

Dokler ta ambiciozna načela ne bodo splošno sprejeta, bodo raziskovalci pri pridobivanju želenih virov in podatkov o kulturni dediščini še vedno potrebovali ogromno časa, volje in potrpežljivosti. Kljub temu pisci tega prispevka menimo, da raziskovalcem ni potrebno zgolj čakati, kdaj bodo raziskovalne in kulturne ustanove začele tudi izvajati ta načela in podpisovati ustrezne listine, temveč lahko s svojo raziskovalno dejavnostjo v skladu z zgornjimi načeli že sami pomembno prispevajo k ustvarjanju primerov dobrih praks ter na ta način krepijo zaupanje med raziskovalci in ustanovami, ki hranijo kulturno dediščino. Pri tem se lahko raziskovalci zanašajo na aktivno pomoč digitalnih raziskovalnih infrastruktur za humanistiko.

V nadaljevanju bomo kot primer takšnega sodelovanja predstavili raziskavo Odlivanje smrti, ki jo Društvo za domače raziskave skupaj s sodelavci iz različnih kulturnih ustanov opravlja v okviru evropskega projekta TRACES.[9] Podatke o kulturni dediščini, ki so zajeti v to raziskavo, raziskovalci nato v skladu z digitalno humanističnimi metodami obdelujejo v sodelovanju z Raziskovalno infrastrukturo slovenskega zgodovinopisja iz Inštituta za novejšo zgodovino.[10] Ker ima življenjski cikel raziskovalnih podatkov ključen pomen v digitalni humanistiki (Collins et al., 2015: 14), smo v skladu s tem ciklom strukturirali tudi podajanje vsebine v tem članku. Zaradi velike količine različnih definicij življenjskega cikla raziskovalnih podatkov smo se odločili v rahlo prilagojeni obliki (Slika 1) prevzeti tisto, ki najbolj ustrezno odraža delovanje raziskovalne skupine in raziskovalne infrastrukture v tukaj opisanem projektu (Puhl et al., 2015).



Slika 1: Življenjski cikel raziskovalnih podatkov

V drugem poglavju bomo predstavili vsebinsko zasnovo raziskave in virov, ki smo jih pri tem uporabili. V tretjem poglavju bomo opisali pridobivanje in ustvarjanje podatkov o kulturni dediščini. V četrtem poglavju o obdelavi podatkov bo predstavljen podatkovni model digitalnih objektov, uporabljeni metapodatkovni standardi in način dostopa. V nadaljevanju bo šestemu poglavju z analizo podatkov sledilo poglavje o diseminaciji v obliki razstave. V zaključku bomo orisali še naše načrte za ponovno uporabo teh raziskovalnih podatkov, skupaj z njihovo nadaljnjo obdelavo, analizo in diseminacijo. Potrebno se je namreč zavedati, da tukaj predstavljeni življenjski cikel raziskovalnih podatkov ni enosmerna pot od definiranja virov do diseminacije raziskovalnih rezultatov, ki se konča s hrambo raziskovalnih podatkov in rezultatov v ustreznih digitalnih repozitorijih, temveč so vsi ti procesi med seboj povezani v interaktivnem odnosu.

2. Vsebinska zasnova raziskave

Raziskava Odlivanje smrti poteka v okviru triletnega evropskega projekta TRACES[11], ki ga financira Evropska komisija (Obzorje 2020)[12]. Društvo za domače raziskave v njem sodeluje kot partner, koordinira pa ga Univerza v Celovcu. Projekt želi preseči uveljavljeno prakso umetniških intervencij in posebno pozornost posveča razvoju metodologij sodelovanja. Jedro raziskovalnega projekta je pet interdisciplinarnih umetniško-raziskovalnih delovnih skupin, ki smo jih poimenovali “ustvarjalne koprodukcije”, v njih pa enakopravno sodelujejo umetniki, znanstveniki in upravljavci kulturne dediščine. Ustvarjalna koprodukcija s sedežem v Ljubljani se na primeru posmrtnih mask ukvarja z vlogo umetnika pri posredovanju sporne kulturne dediščine, kar je tudi krovna tema celotnega projekta.

Odlivanje posmrtne maske je ena najstarejših portretnih kiparskih tehnik (Didi-Huberman, 2013). V 19. stoletju je postala še posebej priljubljena, saj je sovpadla z družbenim uveljavljanjem meščanskega razreda, pri čemer so ključno vlogo odigrali tudi muzeji (Mattl-Wurm, 1998). Izhajali smo iz teze, da posmrtne maske za skupnost pomembnih osebnosti delujejo kot eksploatacijski medij, vpet v natančno strukturirane politične in družbene projekte (nacionalizem, razredni boj, sekularizacija), in iz opažanja, da je javno življenje posmrtnih mask v zatonu in da se ob prenovah muzejskih postavitev umikajo v depoje. Zanimalo nas je, koliko posmrtnih mask sploh hranijo javne zbirke po Sloveniji in kdo pravzaprav so ti ljudje, katerih obličja so bila odlita za javni namen.

Izkazalo se je, da tovrstna raziskava še ni bila opravljena in da podatki javnosti niso na voljo. V prvi fazi smo opravili sondaže v izbranih muzejih (Muzeji in galerije mesta Ljubljane, Muzej novejše zgodovine Slovenije, Moderna galerija in Narodna in univerzitetna knjižnica). Na podlagi ogleda gradiva v depojih, pogovorov s kustosi posameznih zbirk ter primerjave z obstoječimi kataložnimi zapisi smo že identificirali vrsto problemov, vezanih na podatke o posmrtnih maskah: identifikacija upodobljencev (neznani ali različno atribuirani odlitki), določanje avtorstva (različne

[9] Odlivanje smrti, http://ddr.si/sl/category/projekti/odlivanje-smrti/.

[10] Raziskovalna infrastruktura slovenskega zgodovinopisja, http://www.sistory.si/publikacije/?menuBottom=2.

[11] Transmitting Contentious Cultural Heritages with the Arts: From Intervention to Co-Production, http://www.traces.polimi.it/.

[12] https://cordis.europa.eu/project/rcn/200849_en.html.

atribucije; nejasni podatki za tehnološke različice), nejasna provenienca gradiva, ločevanje »originala« od kopij ipd. Večino navedenih težav pripisujemo dejstvu, da slovenske dediščinske institucije posmrtnih mask niso zbirale načrtno, pač pa so te v zbirke zašle po različnih poteh, iz reakcije skrbnikov teh zbirk pa utemeljeno sklepamo, da so bile posmrtne maske doslej dejansko spregledana kulturna dediščina.

Za drugo fazo raziskave smo sestavili nabor 114 kulturnih in znanstvenih organizacij (muzeji, arhivi, knjižnice, galerije, gledališča, inštituti), ki bi utegnile hraniti posmrtne maske, ter z njimi sistematično navezali stike. Obenem smo zaradi kompleksnosti naloge in v želji po zagotovitvi trajne hrambe podatkov po zaključku evropskega projekta vzpostavili sodelovanje med Društvom za domače raziskave in Inštitutom za novejšo zgodovino, v okviru katerega deluje Raziskovalna infrastruktura slovenskega zgodovinopisja. Sodelovanje je omogočilo koncipiranje zbiranja podatkov o izbranem gradivu kot metodološkega pripomočka, ki je že med raziskavo samo vplival na njen potek.

3. Ustvarjanje raziskovalnih podatkov

Sodelovanje med digitalno humanistiko in ustanovami s področja varstva kulturne dediščine temelji na konceptu upravljanja s podatki, kjer glavno vlogo igrajo digitalni nadomestki. To so informacijske strukture, ki identificirajo, dokumentirajo ali predstavljajo primarne vire, ki se uporabljajo v raziskovalnem delu. (Romary, 2014) Digitalni nadomestki torej niso le digitalne fotografije originalnega analognega gradiva, ki ga večinoma hranijo knjižnice, muzeji, arhivi in galerije, temveč tudi metapodatkovni zapisi, prepisi besedil, označevanje strukture in vsebine besedil, digitalne anotacije, oziroma kakršno koli pridobivanje novih podatkov ali pretvorba obstoječih podatkov.

Zbiranje podatkov in fotografij je dolgotrajen in zahteven proces. Izbor javnih institucij, ki so bile povabljene k sodelovanju, teži k temu, da je številčno čim bolj obsežen, hkrati pa upošteva tudi različnost tipov (kulturnih) ustanov. Ta dva kriterija sta pri koncipiranju seznama institucij ključna predvsem zaradi dejstva, da ni vzorca, po katerem bi lahko predvideli, kje bo koncentracija teh predmetov največja. Odločitev, da bo glavni poudarek predvsem na kulturnih ustanovah, izhaja iz razumevanja temeljne funkcije posmrtnih mask, tj. ohranjanje spomina na pokojnika, procesi zgodovinjenja pa se v prvi vrsti odvijajo prav v muzejih, spominskih sobah, domoznanskih oddelkih splošnih knjižnic ipd.

Začetni nabor institucij se je tekom raziskave postopoma spreminjal: nekatere ustanove so bile izločene iz prvotnega seznama, druge spet dodane. Na podlagi naključno pridobljenih informacij o lokacijah posmrtnih mask (pričevanja obiskovalcev razstave, zaposlenih v raznih ustanovah, pisni viri) so se na seznam uvrstile nekatere nove institucije.

Ker so posmrtne maske predmeti, ki so v našem prostoru relativno slabo strokovno obdelani, poleg tega pa so pogosto javnosti tudi nedostopni (v večini primerov hranjeni v depojih), so podatki, ki jih posedujejo njihovi lastniki oz. skrbniki, ključnega pomena za nadaljnje raziskovanje in interpretiranje tega fenomena. Izkazalo se je, da se je skozi čas precejšen del podatkov o posmrtnih maskah izgubil. Ustanove namreč pogosto posredujejo le tiste podatke, ki so jih o maskah vzpostavile same (inventarna števila, tehnika, nahajališče, dimenzije, stanje), podatki o provenienci so v večini primerov skopi oziroma jih sploh ni, prav tako pa je zelo malo znanega o (širšem) kontekstu nastanka posamezne maske. Precej dvoumnosti se pojavlja tudi na področju identificiranja upodobljencev in avtorjev posmrtnih mask.

metapodatek	ustanove	
	št.	v %
upodobljenec	31	100
avtor	24	77,4
tehnika	18	58,1
datacija	14	45,2
inventarna številka	17	54,8
nahajališče	23	74,2
provenienca	20	64,5
dimenzije	13	41,9
stanje	7	22,6
ohranjenost	1	3,2
število kopij	2	6,5
viri	1	3,2

Tabela 1: Metapodatki, ki so jih ustanove lahko posedovale (ne nujno za vsako posmrtno masko)

Po letu in pol intenzivnega poizvedovanja je podatke o številu posmrtnih mask v svojih zbirkah posredovalo dobrih 50% vseh vprašanih ustanov, kar pomeni: 32 muzejev in galerij, 19 knjižnic, 3 gledališča, 4 spominske sobe/hiše ter 6 različnih kulturnih institucij, ki ne sodijo v nobeno od prej naštetih kategorij (SAZU, Cankarjev dom, AGRFT idr.). Fotografije teh posmrtnih mask je imelo 19 ustanov, izmed katerih so nekatere prav zaradi naše raziskave maske šele prvič tudi fotografirale. V primeru osmih ustanov so maske fotografirali šele raziskovalci. V nekaterih primerih mask ni bilo mogoče fotografirati, mdr. tudi zaradi tega, ker so bile v preveč slabem stanju.

4. Obdelava raziskovalnih podatkov

Člani raziskovalne skupine, ki smo digitalne nadomestke pridobili od zgoraj navedenih javnih zavodov, le-teh nismo mogli takoj uporabiti v svoji raziskavi, temveč smo jih morali za potrebe svoje raziskave temu primerno najprej obdelati. Kot smo videli, pridobljeni metapodatki namreč niso bili nujno narejeni po enotnih standardih, predvsem pa so bili s stališča izvajanja raziskave v mnogih primerih tudi pomanjkljivi oziroma neprimerni. Ustanove s področja varstva kulturne dediščine glede na svoje poslanstvo z ustvarjanjem digitalnih nadomestkov praviloma zadovoljujejo potrebe širše javnosti, ki se zanima za kulturno dediščino in ne specifičnih interesov posameznih raziskovalnih skupin in njihovih projektnih vprašanj. Te interese v prvi vrsti pokrivajo raziskovalne infrastrukture (Blanke et al., 2018).

4.1. Podatkovni model

Obenem je bilo precej posmrtnih mask dostopnih samo v analogni obliki, zato smo morali veliko digitalnih nadomestkov najprej šele ustvariti. V skladu z raziskovalnimi praksami v digitalni humanistiki smo se odločili, da za digitalne nadomestke ustvarimo podatkovni model, ki bo povsem ustrezal specifičnim raziskovalnim potrebam našega projekta. Podatkovni model ni opis realnega sveta, temveč je interpretacija (analognega) objekta. Podatkovno modeliranje je v prvi vrsti ustvarjalni in kreativen proces, pri čemer funkcija digitalnega nadomestka določa, katere aspekte je potrebno modelirati (Flanders in Jannidis, 2015).

Raziskovalna skupina se je odločila ustvariti zbirko digitalnih objektov (posmrtnih mask), kjer ima vsak digitalni objekt nič ali več digitalnih fotografij in sledeče metapodatke:[13] naslov digitalnega objekta, opis, upodobljenec, avtor maske, naročnik, leto naročila, verzija odlitka, število znanih odlitkov, tehnika, institucija/lokacija, zbirka/nahajališče, inventarna številka, stanje predmeta in oznake, provenienca/zgodovina predmeta, viri (in literatura).

Ta podatkovni model je po eni strani relativno zelo enostaven. Lahko rečemo, da je povsem običajen glede dojemanja kulturne dediščine. Zato smo ga lahko tudi relativno enostavno kot zbirko digitalnih objektov vključili v portal Zgodovina Slovenije – SIstory, ki ga upravlja Raziskovalna infrastruktura slovenskega zgodovinopisja.[14] Digitalna zbirka omogoča iskanje in brskanje po digitalnih objektih ter pregled vseh metapodatkov in digitalnih fotografij.[15]

Vendar ima ta enostavni podatkovni model tudi nekatere pomanjkljivosti, ki bi lahko imele negativen vpliv na nadaljnji potek raziskav. Pri zapisih metapodatkov o upodobljencih, avtorjih mask in pogojno tudi naročnikih smo prvotno pri vsakem digitalnem objektu zapisovali imena in priimke teh oseb, njihove poklice ter datume rojstva in smrti. Takšna rešitev je imela sledeče pomanjkljivosti:

- oseba je lahko imela več kot en poklic;
- če je bila ista oseba prisotna pri več kot enemu digitalnemu objektu, bi bilo pri napačnih ali pomanjkljivih zapisih potrebno iste spremembe vnašati pri vseh teh objektih;
- podobno bi bilo potrebno nove vrste metapodatkov (npr. spol, kraj rojstva ali smrti) o posamezni osebi enotno vnesti pri vseh digitalnih objektih, kjer se ta oseba omenja.

Zato smo se odločili, da prvotni podatkovni model dopolnimo z dodatnimi entitetami. Poleg osnovne entitete object (objekt: posmrtna maska) smo v podatkovnem modelu začeli uporabljati še entiteto person (oseba), v načrtu pa imamo še razširitev podatkovnega modela z entitetama organization (javni zavodi, ki hranijo posmrtne maske) in place (kraji rojstva in/ali smrti). Relacije med objektom in osebo so lahko treh vrst (type): subject (oseba, ki je predmet upodobitve: upodobljenec), production (oseba ki je izdelala masko: avtor) in commissioning (oseba ali organizacija, ki je naročila izdelavo maske: naročnik).

4.2. Standardizacija

V naslednjem koraku smo se odločili, da bomo specifičen podatkovni model naše raziskave v čim večji možni meri uskladili z obstoječimi standardi s področja humanistike in umetnosti. S hitrim naraščanjem količine digitalnega gradiva v humanistiki in umetnosti je standardizacija praktično postala nuja za vse raziskovalce, ki želijo svoje digitalne podatke primerjati in deliti z ostalimi podobnimi digitalnimi podatki. Ker pa je standardizacijo mogoče uspešno izpeljati samo na podlagi ustreznega tehnično strokovnega znanja, se ji raziskovalci s področja umetnosti in humanistike pogosto poskušajo izogniti (Romary et al., 2016). V primeru naše raziskave nam je uspelo standardizacijo izpeljati relativno hitro in enostavno. Pri tem smo se oprli na obstoječe postopke in izkušnje raziskovalne infrastrukture.

Portal SIstory podobno kot veliko ostalih digitalnih knjižnic uporablja zelo razširjen Dublin Core metapodatkovni standard. Zaradi specifičnih potreb tega portala smo v Raziskovalni infrastrukturi slovenskega zgodovinopisja razvili aplikacijski profil, ki sloni na razširjenem Dublin Core (DCMI Metadata Terms).[16] Po vzoru projekta HOPE[17] smo mu dodali še nekatere elemente iz drugih metapodatkovnih shem, ki so potrebni pri opisu arhivskih, knjižničarskih, muzejskih in avdiovizualnih objektov (Pančur, 2013a; Pančur, 2013b). V okviru raziskave Odlivanje smrti so se za zelo primerne izkazali elementi iz sklopa muzejskih metapodatkov, katere smo prevzeli iz LIDO[18] in Spectrum[19] metapodatkovnega standarda. Bolj natančno je ta standardizacija prikazana v spodnji tabeli:

Standard	Element	Opis
DCMI	title	naslov objekta
DCMI	description	opis objekta
DCMIType	type	tip objekta: Physical Object
DCMI	creator	avtor maske
DCMI	contributor	naročnik
DCMI	created, date	leto naročila oz. izdelave
DCMI	hasVersion, isVersonOf	relacije med verzijami
LIDO	eventMaterialsTech	tehnika
SIstory	collection	ustanova analogne maske
DCMI	accessRights	zbirka/nahajališče
DCMI	identifier	inventarna številka
Spectrum	TechnicalAttributes	stanje objekta
LIDO	objectMeasurement	velikost objekta
DCMI	provenance	provenienca
DCMI	bibliographicCitation	viri in literatura

Tabela 2: Metapodatkovni standardi zbirke digitalnih objektov Odlivanje smrti na portalu SIstory

Z razširitvijo podatkovnega modela z entiteto person obstoječi metapodatkovni aplikacijski profil portala SIstory ni več ustrezal vsem potrebam raziskave Odlivanje

[13] Primerjaj prvotno poskusno postavitev baze posmrtnih mask http://ddr.si/sl/mask/.

[14] Odlivanje smrti / Casting of Death, http://hdl.handle.net/11686/menu196.

[15] Andrej Pančur, Zbirka posmrtnih mask na portalu SIstory [blog], 5. 8. 2017, http://ddr.si/sl/zbirka-posmrtnih-mask-na-portalu-sistory/.

[16] Dublin Core Metadata Initiative, http://dublincore.org/documents/dcmi-terms/.

[17] HOPE: Heritage of the People's Europe, http://hopewiki.socialhistoryportal.org/.

[18] LIDO: Lightweight Information Describing Object, http://network.icom.museum/cidoc/working-groups/lido/what-is-lido/.

[19] Spectrum, https://collectionstrust.org.uk/resource/the-spectrum-standard-v4-0/.

smrti. SIstory za upravljanje metapodatkov uporablja MySQL relacijsko bazo podatkov. Relacijske baze podatkov so sicer najpogosteje uporabljena tehnologija baz podatkov[20] in so zlasti primerne za upravljanje velike količine medsebojno povezanih entitet kot so osebe, predmeti, procesi ipd. Toda po drugi strani je pri relacijskih podatkovnih bazah potrebno vnaprej določiti strukturo njenih podatkov, podatkovno shemo in podatkovne tipe. Vsake naknadne spremembe so zelo zahtevne in jih je potrebno skrbno načrtovati.

Podobno kot pri mnogih digitalno humanističnih projektih je tudi pri naši raziskavi podatkovni model prilagojen specifičnih potrebam raziskave, hkrati pa mora biti dovolj fleksibilen za dodatne nadgradnje podatkovnega modela v skladu z vedno novimi raziskovalnimi vprašanji. Zato smo namesto preveč toge relacijske baze podatkov portala SIstory raje uporabili veliko bolj fleksibilno XML podatkovno strukturo.

V ta namen smo v okviru raziskovalne infrastrukture razvili postopek, ki omogoča pretvorbo podatkov iz datotek XML v statične HTML spletne strani in njihovo vključitev v portal SIstory. Pri kodiranju datotek XML uporabljamo Smernice Text Encoding Initiative (TEI) (TEI Consortium, 2018). Smernice TEI so predvsem v digitalni humanistiki *de facto* standard za kodiranje besedil. Smernice med drugim vključujejo tudi modul za kodiranje imen, datumov, oseb in krajev. Ta modul smo uporabili za kodiranje podatkov o osebah (entiteta *person* našega podatkovnega modela), v bodoče pa ga nameravamo uporabiti še za kodiranje entitet *organization* in *place*. V našem primeru je bila odločitev za uporabo TEI še toliko lažja, ker je večina oseb iz naše raziskave vključenih v Slovensko biografijo,[21] ki za kodiranje podatkov o osebah tudi uporablja TEI (Erjavec et al., 2011). Te podatke smo zato lahko samo z manjšimi spremembami relativno enostavno vključili v našo raziskavo.

Dokument TEI raziskave Odlivanje smrti vsebuje dva seznama entitet projektnega podatkovnega modela (object in person). Oba seznama sta kot <div> vključena v element <body>. Seznam oseb <listPerson> vključuje elemente <person> s podatki o osebah našega podatkovnega modela. Ta element nato vsebuje enega ali več elementov za kodiranje vseh možnih različic osebnih imen te osebe (<persName>), kodiranje vrednosti za spol osebe <sex> (vrednost atributa @value M za moške in F za ženske), podatki o enemu ali več poklicih <occupation>, podatke o rojstvu <birth> in smrti <death> ter nenazadnje URL identifikator <idno> spletnega mesta z dodatnimi metapodatki o teh osebah. Elementa o rojstvu in smrti lahko vsebujeta podatek o datumu <date> in kraju <placeName> rojstva ali smrti. Kot primer dobre prakse sodelovanja med raziskovalci in raziskovalnimi infrastrukturami smo iz projekta Slovenska bibliografija prevzeli taksonomijo poklicev. Izvorno taksonomijo, ki ni javno dostopna, smo vključili v <teiHeader> našega TEI dokumenta. Elementi <occupation> se na njo navezujejo preko atributa @code.

Večji izziv kot kodiranje podatkov o osebah je predstavljala druga entiteta našega podatkovnega modela: object (posmrtna maska). TEI je namreč namenjen kodiranju besedil in ne objektov, zato predstavlja kodiranje objektno usmerjenih zbirk nebesedilne kulturne dediščine precejšen izziv (Nelson, 2017). Mi smo se odločili uporabiti zaporeden seznam <list> objektov, kjer vsak objekt kot postavka <item> vsebuje svoj seznam glavnih metapodatkov. Ta seznam je kodiran kot glosar seznam izrazov <item> in njihovih opredelitev <label>. Pri tem smo enotno kodirali samo sledeče opredelitve:

- naziv posmrtne maske: Dublin Core title element;
- avtor maske: notranja povezava <ref> na element <person>;
- upodobljenec: notranja povezava <ref> na element <person>;
- SIstory: zunanja povezava <ref> na digitalni objekt te posmrtne maske v zbirki portala SIstory;
- LIDO metapodatki: zunanja povezava <ref> na LIDO metapodatke.

Ti seznami objektov torej razen naslova vsebujejo samo reference na izvorne digitalne objekte portala SIstory, na ostale entitete (*person*) kodirane v TEI in nenazadnje na vse metapodatke objekta posmrtne maske, ki smo jih kodirali v skladu s standardom LIDO. Ta standard je zlasti primeren za opisovanje muzejskih objektov, med drugim tudi analognega nebesedilnega gradiva kot so posmrtne maske. Kot takšnega ga uporabljajo tudi sorodne raziskovalne infrastrukture s področja digitalne humanistike (Steiner in Stigler, 2017).

Vsak objekt (digitalna maska) ima svojo datoteko XML z LIDO metapodatki. Za izdelavo teh metapodatkov uporabljamo izvoz metapodatkov o digitalnih objektih zbirke Odlivanje smrti iz SIstory relacijske baze podatkov v datoteko XML, (Pančur, 2013c) ki jo potem s posebej napisanim programom XSLT pretvorimo v datoteke XML z LIDO zapisom. LIDO metapodatki zapis vsebuje identifikator <lidoRecID> (uporabljamo SIstory handle identifikator), kategorijo <category> (opredelimo, da je fizični objekt), opisne in administrativne metapodatke. Slednji vsebujejo podatke o zapisu (<recordWrap>), za katerega v skladu z LIDO terminologijo[22] določajo, da ta zapis opredeljuje posamezen objekt (Item-level record), ki ga je z SIstory identifikatorjem (<recordID>) prispevalo Društvo za domače raziskave (<recordSource>). Vsak tak objekt ima lahko tudi eno ali več fotografij (<resourceRepresentation>).

Najbolj obsežni so opisni metapodatki (<descriptiveMetadata>). Z njimi najprej klasificiramo (<objectClassificationWrap>) objekt kot posmrtno masko (<objectWorkTypeWrap>),[23] ki upodablja konkretno osebo (<classificationWrap>). Potem identificiramo objekt (objectIdentificationWrap>) z njegovim nazivom (<titleSet>), ustanovo izvornega analognega gradiva (<repositoryWrap>), se pravi naziv ustanove (<repositoryName>), signaturo (<workID>) in lokacijo hrambe (<repositoryLocation>) ter še različne vsebinske opise maske in njenega stanja (<objectDescriptionWrap>) in njene mere (<objectMeasurementsWrap>).

Naposled sledi del (<eventWrap>), ki opisuje tri ključne dogodke (<eventSet>), ki so povezani z objektom. Za označbo vrste (<eventType>) vsakega od njih se uporablja ustrezna LIDO terminologija.

[20] DB-Engines Ranking, https://db-engines.com/en/ranking.

[21] Slovenska biografija, http://www.slovenska-biografija.si/.

[22] LIDO-Terminologie, http://terminology.lido-schema.org.

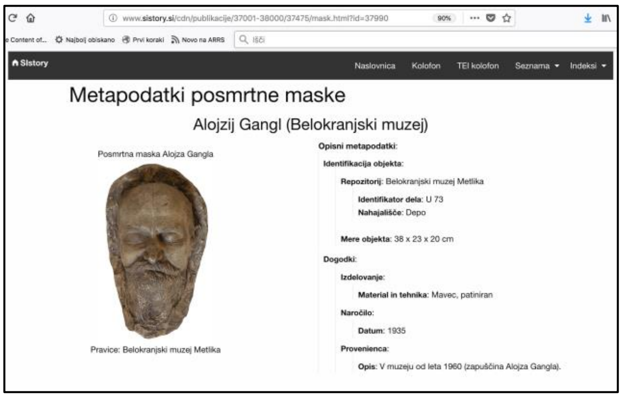
[23] Pri tem uporabimo tudi Getty Art & Architecture Thesarus, http://vocab.getty.edu/page/aat/300047724.

- Izdelovanje (Production): vsebuje podatke o izdelovalcu (<eventActor>) maske in materialu, iz katerega jo je izdelal (<eventMaterialsTech>).
- Naročilo (Commissioning): vsebuje podatke o naročniku (<eventActor>) in datumu naročila (<eventDate>).
- Provenienca (Provenience): opisih vseh znanih menjav lastništva in hrambe maske.

4.3. Digitalna izdaja

Tako kodirani dokumente TEI in LIDO so dostopni v GitHub repozitoriju.[24] Za izdelavo HTML digitalne izdaje[25] smo uporabili standardne pretvorbe XSLT konzorcija TEI,[26] ki smo jih nadgradili v skladu s potrebami vključitve HTML statičnih spletnih strani v portal SIstory.[27]

Te generične pretvorbe XSLT je za vsako digitalno izdajo mogoče prilagoditi specifičnim potrebam posamezne raziskave. V okviru raziskave Odlivanje smrti smo to fleksibilnost našega sistema poskusili izkoristiti v čim večji meri in smo statičnim spletnim stranem dodali še nekatere dinamične funkcionalnosti. Za prikazovanje LIDO metapodatkov samo o eni posmrtni maski na posamezni spletni strani (unikatni URL) smo uporabili Saxon-JS.[28]



Metapodatki posmrtne maske

Alojzij Gangl (Belokranjski muzej)

Posmrtna maska Alojza Gangla

Pravice: Belokranjski muzej Metlika

Opisni metapodatki:

Identifikacija objekta:

Repozitorij: Belokranjski muzej Metlika

Identifikator dela: U 73

Nahajališče: Depo

Mere objekta: 38 x 23 x 20 cm

Dogodki:

Izdelovanje:

Material in tehnika: Mavec, patiniran

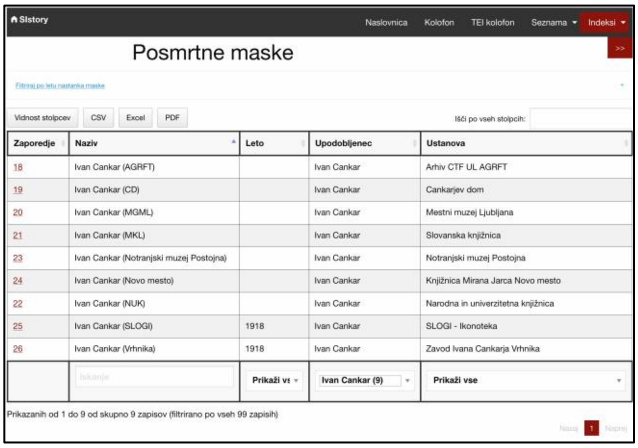
Naročilo:

Datum: 1935

Provenienca:

Opis: V muzeju od leta 1960 (zapuščina Alojza Gangla).

Slika 2: Statična spletna stran digitalne izdaje z dinamičnim prikazom LIDO opisnih metapodatkov o eni posmrtni maski



Posmrtne maske

Zaporedje	Naziv	Leto	Upodobljenec	Ustanova
18	Ivan Cankar (AGRFT)		Ivan Cankar	Arhiv CTF UL AGRFT
19	Ivan Cankar (CD)		Ivan Cankar	Cankarjev dom
20	Ivan Cankar (MGML)		Ivan Cankar	Mestni muzej Ljubljana
21	Ivan Cankar (MKL)		Ivan Cankar	Slovanska knjižnica
23	Ivan Cankar (Notranjski muzej Postojna)		Ivan Cankar	Notranjski muzej Postojna
24	Ivan Cankar (Novo mesto)		Ivan Cankar	Knjižnica Mirana Jarca Novo mesto
22	Ivan Cankar (NUK)		Ivan Cankar	Narodna in univerzitetna knjižnica
25	Ivan Cankar (SLOGI)	1918	Ivan Cankar	SLOGI - Ikonoteka
26	Ivan Cankar (Vrhnika)	1918	Ivan Cankar	Zavod Ivana Cankarja Vrhnika
	Iskanje	Prikaži vs	Ivan Cankar (9)	Prikaži vse

Prikazanih od 1 do 9 od skupno 9 zapisov (filtrirano po vseh 99 zapisih)

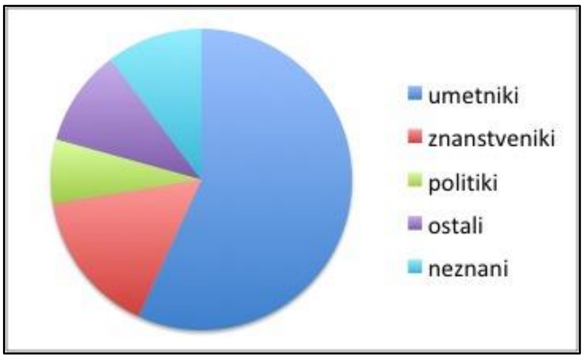
Slika 3: Interaktivna DataTables tabela

Za najbolj koristnega pa se je za potrebe naše raziskave izkazala odprtokodna DataTables, ki je vtičnik za jQuery JavaScript knjižnico.[29] Z njegovo pomočjo smo HTML tabelam (Posmrtne maske, Osebe, Posmrtne maske z znanimi upodobljenci, Upodobljenci posmrtnih mask) dodali številne funkcionalnosti, ki so nam omogočale filtriranje, razporejanje, iskanje in izvažanje želenih podatkov. (Slika 3) S pomočjo teh tabel smo se lahko uspešno lotili naslednje stopnje našega življenjskega cikla podatkov - analize raziskovalnih podatkov.

5. Analiza raziskovalnih podatkov

S pomočjo teh tabel smo lahko zelo enostavno prišli do nekaterih rezultatov kot npr. kdo je najpogostejši upodobljenec (Ivan Cankar - 9 kopij posmrtnih mask, primerjaj sliko 2) ali katera ustanova hrani največ posmrtnih mask (Mestni muzej Ljubljana - 17).

Za nekoliko bolj zapletene izračune pa te tabele omogočajo tudi izvoz vseh ali samo filtriranih podatkov v CSV format (za manj zahtevne uporabnike tudi Excel), ki ga nato lahko uporabimo v nadaljnjih statističnih izračunih v poljubnem statističnem programu. Na ta način smo tako izvozili podatke o vseh prvo navedenih poklicih upodobljencev. Te poklice smo nato klasificirali v skladu s preprosto shemo (slika 4).



Slika 4: Poklicne skupine upodobljencev

6. Diseminacija rezultatov

Sredi triletnega raziskovalnega obdobja smo se odločili za javni prikaz delnih rezultatov raziskave. Pripravili smo razstavo,[30] ki je fenomen posmrtnih mask predstavila skozi tri sklope: Zbiranje, Odlivanje[31] in Oživljanje. Prvi sklop je bil posvečen zbranim podatkom o historičnih posmrtnih maskah iz slovenskih javnih zbirk. Ker je bil odziv nagovorjenih institucij v času odprtja razstave le 55-odstoten, je bil eden od njenih namenov spodbuditi odziv še pri preostalih. Prvenstveno pa smo želeli javnost seznaniti z nekaterimi delnimi izsledki. Izbrali smo tri kriterije za prikaz baze podatkov v obliki infografik:

- posmrtne maske glede na poklic upodobljencev,

[24] Odlivanje smrti, https://github.com/SIstory/publications.

[25] Odlivanje smrti: Pregled objav na portalu Zgodovina Slovenije – SIstory, http://hdl.handle.net/11686/37475.

[26] TEI XSL Stylesheets, https://github.com/TEIC/Stylesheets.

[27] SIstory TEI Stylesheets, https://github.com/SIstory/Stylesheets.

[28] Saxon-JS, http://www.saxonica.com/saxon-js/index.xml.

[29] DataTables, https://datatables.net/.

[30] Društvo za domače raziskave v sodelovanju z Viktorjem Gojkovičem: Odlivanje smrti, Galerija Vžigalica (MGML), Ljubljana, 1. 11. – 24. 12. 2017, http://ddr.si/sl/odlivanje-smrti-v-vzigalici/.

[31] Dejanske (»analogne«) posmrtne maske smo razstavili izključno v okviru predstavitve kiparja Viktorja Gojkoviča, ki se s prakso odlivanja ukvarja od leta 1963.

- upodobljenci z največjim številom posameznih odlitkov posmrtne maske,
- posmrtne maske po letu smrti.

Prvi statistični prikaz podpira teorijo o kulturnem svetništvu (Dović, 2016), saj sta več kot dve tretjini upodobljencev umetnikov. Drugi kriterij je vzpostavil lestvico priljubljenosti posameznih osebnosti za nacionalno identiteto (z devetimi odlitki je na prvem mestu Ivan Cankar, sledijo Rihard Jakopič, Simon Gregorčič, Oton Župančič in Ivan Levar). Najbolj pa je presenetil tretji prikaz, saj smo pričakovali, da bo največ upodobljencev s konca 19. in začetka 20. stoletja, ko je bila posmrtna maska kot medij najbolj priljubljena (Mattl-Wurm, 1992). Dejansko pa je največ mask v slovenskih javnih zbirkah iz 50-ih let 20. stoletja, čeprav sprememba režima ni bistveno vplivala na izbiro upodobljencev (politiki ostajajo redki, vseh skupaj je le 5 %).

Tako kot nastajajoča baza podatkov na portalu SIstory, je bila tudi razstava zastavljena delovno, t. j. nereprezentančno, in je sprožila različne odzive: obiskovalci so prispevali dodatne informacije o posmrtnih maskah v drugih zbirkah, z njihovo pomočjo nam je uspelo tudi identificirati 3 neznane upodobljence. Dober odziv medijev in institucij v nadaljevanju raziskave pa daje slutiti, da posmrtne maske odslej ne bodo več del spregledane kulturne dediščine.

Po večkratnih poizkusih smo naposled uspeli pridobiti podatke o posmrtnih maskah od 118 institucij (93%) iz različnih delov Slovenije. Le 32 (27%) izmed njih ima v svoji zbirki tovrstne predmete. Izkazalo se je, da so muzeji, galerije in spominske sobe (22 ustanov) najpogostejše lokacije, kjer se danes nahajajo maske, saj hranijo 66% vseh evidentiranih mask, tj. 70 primerkov. Sledijo jim knjižnice (5 ustanov) – v njihovih depojih se nahaja 14% vseh zabeleženih odlitkov oz. 16 primerkov. Dvajset pa jih je našlo svoje mesto v zbirkah drugih kulturnih ustanov: na Akademiji za gledališče, radio, film in televizijo (AGRFT) ter SAZU, v Cankarjevem domu, Slovenskem gledališkem inštitutu in v arhivu Studia Slovenica.

7. Zaključek

Trenutno so delovne verzije zbirke digitalnih objektov in digitalne izdaje hranjene na portalu SIstory. Po zaključku projekta nameravamo končne verzije teh raziskovalnih rezultatov dolgoročno shraniti tudi v naši aplikaciji Archivematica (Pančur in Rožman, 2016).

Zaradi sprotnega nadgrajevanja in dopolnjevanja zbirke in vmesnega objavljanja raziskovalnih rezultatov je potrebno imeti jasno zastavljeno in pregledno kontrolo nad različnimi verzijami zbirke raziskovalnih podatkov o kulturni dediščini. Vsaka od teh verzij je namreč zbirka novih digitalnih nadomestkov. Zato smo se odločili, da bomo repozitorij portala SIstory nadgradili na način, ki bo omogočal dolgoročno in čim bolj trajnostno hrambo digitalnih izdaj. Raziskovalni rezultati projekta Odlivanje smrti nam bodo v tem primeru služili kot odličen testni primer.

8. 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 ter skozi projekt TRACES, ki je prejel sredstva iz programa Evropske unije za raziskave in inovacije Obzorje 2020 po pogodbi št. 693857.

9. Literatura

Anne Baillot, Mike Mertens in Laurent Romary. 2016. Data Fluidity in DARIAH – Pushing the Agenda Forward. *BIBLIOTHEK – Forschung und Praxis*, 40(2): 151-164.

Anne Baillot, Marie Puren, Charles Riondet in Laurent Romary. 2017. Access to cultural heritage data: a challenge for the Digital Humanities. V: *Digital Humanities 2017: Conference Abstracts*, str. 157-159. McGill University & Université de Montréal: Montréal, Canada.

Tobias Blanke, Conny Kristel in Laurent Romary. 2018. Crowds for Clouds: Recent Trends in Humanities Research Infrastructures. V: A. Benardou, E. Champion, C. Dallas in L. M. Hughes (ur.), *Cultural Heritage Infrastructures in Digital Humanities*, str. 48-62. Routledge, New York in Oxon.

Sandra Collins, Natalie Harrower, Dag Trygve Truslew Haug, Beat Immenhauser, Gerhard Lauer, Tito Orlandi, Laurent Romary in Eveline Wandl-Vogt. 2015. *Going Digital: Creating Change in the Humanities: ALLEA E-Humanities Working Group Report*. ALLEA. https://hal.inria.fr/hal-01154796.

George Didi-Huberman. 2013. *Podobnost prek stika: Arheologija, anahronizem in modernost odtisa*. Studia humanitatis.

Marijan Dović (ur.). 2016. *Kulturni svetniki in kanonizacija*. ZRC SAZU.

Tomaž Erjavec, Jan Jona Javoršek, Matija Ogrin in Petra Vide Ogrin. 2011. Od biografskega leksikona do znanstvenokritične izdaje: vprašanje trajnosti elektronskih besedil. *Knjižnica*, 55(1): 103-114. https://knjiznica.zbds-zveza.si/knjiznica/article/view/6004.

Julia Flanders in Fotis Jannidis. 2015. *Knowledge Organization and Data Modeling in the Humanities*. White paper of the Conference Proceeding. http://www.wwp.northeastern.edu/outreach/conference/kodm2012/flanders_jannidis_datamodeling.pdf.

Dafydd Gibbon. 2012. Resources for technical communication systems. V: Alexander Mehler in Laurent Romary (ur.), *Handbook of Technical Communication*, str. 255-284. Walter de Gruyter, Berlin/Boston.

Sylvia Mattl-Wurm. 1998. Die Totenmaskensammlung des Historischen Museums der Stadt Wien. V: Norbert Stefenelli (ur.), *Körper ohne Leben: Begegnung und Umgang mit Toten*, Böhlau Verlag.

Sylvia Mattl-Wurm (ur.). 1992. *Bilder vom Tod*. Sonderausstellung des Historischen Museums der Stadt Wien.

Brent Nelson. 2017. Curating Object-Oriented Collections Using the TEI. *Journal of the Text Encoding Initiative*, (9). DOI : 10.4000/jtei.1680

Andrej Pančur. 2013a. *Metapodatki portala Zgodovina Slovenije – SIstory: Navodila za uporabo orodja za vnos metapodatkov*. Inštitut za novejšo zgodovino. http://hdl.handle.net/11686/36151.

Andrej Pančur. 2013b. *SIstory Dublin Core XML Schema 1.0*. Inštitut za novejšo zgodovino. http://hdl.handle.net/11686/37479.

Andrej Pančur. 2013c. *SIstory Basic XML Schema 2.0*. Inštitut za novejšo zgodovino. http://hdl.handle.net/11686/37478.

Andrej Pančur in Bogomir Rožman. 2016. Dolgotrajno ohranjanje raziskovalnih podatkov v manjših raziskovalnih infrastrukturah: Uporaba odprtokodne aplikacije Archivematica. V: Tomaž Erjavec in Darja Fišer (ur.), *Zbornik konference Jezikoslovne tehnologije in digitalna humanistika, 29. september – 1. oktober 2016, Filozofska fakulteta, Univerza v Ljubljani, Slovenija*. Ljubljana: Znanstvena založba Filozofske fakultete. http://www.sdjt.si/wp/dogodki/konference/jtdh-2016/zbornik/.

Johanna Puhl, Peter Andorfer, Mareike Höckendorff, Stefan Schmunk, Juliane Stiller in Klaus Thoden. 2015. *Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften*. Göttingen: GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität (DARIAH-DE working papers 11). http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2015-4-4.

Laurent Romary. 2014. Sustainable data for sustainable infrastructures. V: A. Duşa, D. Nelle, G. Stock in G. G. Wagner, *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*. SCIVERO Verlag.

Laurent Romary, Emiliano Degl'innocenti, Klaus Illmayer, Adeline Joffres, Emilie Kraikamp, Nicolas Larrousse, Maciej Ogrodniczuk, Marie Puren, Charles Riondet in Dorian Seillier. 2016. *Standardization survival kit (Draft)*. [Research Report], Inria. https://hal.inria.fr/hal-01513531.

Dorian Seillier, Anne Baillot, Marie Puren in Charles Riondet. 2017. *Survey on researchers requirements and practices towards Cultural Heritage institutions: Documentations and analysis*. [Technical Report] Inria Paris. https://hal.inria.fr/hal-01562860.

Elisabeth Steiner in Johannes Stigler. 2017. *GAMS and Cirilo Client: Policies, documentation and tutorial*, verzija 2017-04-10. Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities. http://hdl.handle.net/11471/521.1.1.

TEI Consortium, ur. 2018 TEI P5: *Guidelines for Electronic Text Encoding and Interchange 3.3.0.* TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Analiza slovničnih napak v korpusu spisov učencev japonščine na osnovni ravni

Miha Pavlovič,* Rena Ito†

* Oddelek za Azijske študije, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
miha.pavlovic1@gmail.com
† Fakulteta za medkulturno komunikacijo, Univerza Rikkyō
Nishi-Ikebukuro 3-34-1, Toshima-Ku, Tokyo, 171-8501 Japonska
rn_yis7@yahoo.co.jp

Povzetek

V prispevku opisujemo izgradnjo korpusa spisov študentov japonščine na začetni stopnji in začetno analizo slovničnih napak v njem. Namen raziskave je ugotoviti, katere slovnične strukture povzročajo največ težav slovenskim učencem japonskega jezika na osnovni ravni. V 143 zbranih besedilih smo identificirali in kategorizirali 516 napak ter določili deset najpogostejših tipov napak. Pričakujemo, da lahko te ugotovitve koristijo tako učencem kot učiteljem japonščine pri izvajanju učnega procesa, hkrati pa je tako nastali korpus lahko prvi korak k izgradnji obsežnejšega, označenega in javno dostopnega korpusa besedil slovenskih učencev japonščine za nadaljnje raziskave o učenju japonščine kot tujega jezika.

Analysis of grammatical errors in a Japanese beginner learner corpus

We present the construction of a corpus of Japanese texts written by beginners, learners of Japanese as a foreign language at the University of Ljubljana, and a preliminary analysis of grammatical errors found in the corpus. The aim of our research is to determine which grammatical constructions are most difficult for Slovene beginner learners of Japanese. In the 143 texts we collected, we identified and categorised 516 grammatical errors, and determined the ten most common error types. These results can be useful to both learners and teachers, and the corpus can be a first step towards the construction of a larger, annotated and public corpus of texts by Slovene speaking learners of Japanese, as a basis for research on learning Japanese as a foreign language.

1. Uvod

V raziskavi ameriškega Inštituta za zunanje zadeve (Foreign Service Insitute -FSI) leta 2007 je bil japonski jezik uvrščen v skupino najzahtevnejših jezikov na svetu za maternega govorca angleščine, za katerega predvidevajo, da potrebuje 2200 in več ur aktivnega učenja, da doseže t. i. nivo S-3 (professional working proficiency), kar približno ustreza nivoju C1-C2 po lestvici CEFR, medtem ko je bil na primer slovenski jezik uvrščen v drugo kategorijo, za katero je potrebnih 1100 ur, kar je kar dvakrat manj od japonščine. Tu gre sicer za oceno s stališča maternega govorca angleščine, a po vsej verjetnosti ocena za maternega govorca slovenščine ne bi bila zelo drugačna.

Japonski jezik se razlikuje od slovenskega na vseh ravneh: ima drugačno stavčno strukturo (pri japonščini je povedek vedno na koncu stavka, medtem ko je pri slovenščini v povednem stavku glagol po navadi na drugem mestu); skloni, ki se v slovenskem jeziku izražajo s pregibanjem samostalnika, se v japonščini s pomočjo t. i. sklonskih členkov (*kakujoshi* 格助詞), ki se »pripnejo« samostalniku in mu tako določijo sklon; zapisuje se s štirimi nabori znakov, ki se uporabljajo za zapis različnih besednih vrst oz. morfemov itd.

Tako seveda ni nič čudnega, da pri učencih japonskega jezika pogosto prihaja do napak, predvsem na področjih, kjer se oba jezika poglavitno razlikujeta. Vendar so napake, ki nastopijo ob rabi jezika, eden izmed sestavnih delov učenja le-tega. Do njih pride zaradi pomanjkanja znanja o določenem področju oz. elementu jezika, ali pa zaradi napačne uporabe pridobljenega znanja. To je povsem naraven proces in v kolikor se te napake prepozna in posledično odpravi, lahko govorimo o pozitivnem učinku na učenje jezika. Obratno pa v primeru, da napake niso prepoznane in se jih ne odpravi, lahko imajo takšne pomanjkljivosti v znanju jezika negativen vpliv na nadaljnje učenje. Zato je pomembno, da se napake pravočasno prepozna in odpravi in s tem zagotovi trdne temelje oz. dobro osnovo za nadaljnje učenje jezika.

Sistematično učenje japonščine kot tujega jezika na visokošolski ravni se je v Sloveniji začelo leta 1995, ko je bil na Filozofski fakulteti v Ljubljani ustanovljen Oddelek za azijske študije in z njim prvi (in zaenkrat še edini) študijski program japonologije v Sloveniji. Tako japonologija sodi med eno mlajših strok Filozofske fakultete. Kljub temu, da je bilo v dvaindvajsetih letih delovanja izvedenih mnogo raziskav in projektov, ki se ukvarjajo s temo japonskega jezika, trenutno še ni na voljo vira ali orodja, s pomočjo katerega bi bil učencem in učiteljem japonskega jezika omogočen enostaven vpogled v podatke o tipih napak, ki učencem najpogosteje predstavljajo težave, in primerih le-teh. Tako smo kot prvi korak proti temu cilju v akademskem letu 2016/2017 zbrali korpus spisov študentov japonščine na začetni stopnji in v njem analizirali napake[1].

V okviru te raziskave smo iz pisnih besedil študentov prvega letnika programa Japonologija na Filozofski fakulteti v Ljubljani ustvarili in analizirali korpus 143 besedil, v katerem smo ugotovili 516 slovničnih napak. Napake smo najprej identificirali, kategorizirali in prešteli, nato pa smo določili deset najpogostejših tipov napak. Raziskavo smo izvedli z namenom, da ugotovimo, katere

[1] Raziskava je nastala pri predmetu Timsko raziskovalno delo pod mentorstvom dr. Kristine Hmeljak v letu 2016/17, ko je bila Rena Ito na študijski izmenjavi na Oddelku za azijske študije Filozofske fakultete Univerze v Ljubljani.

slovnične strukture slovenskim učencem japonskega jezika na osnovni ravni povzročajo največ težav. Pričakujemo, da lahko te ugotovitve koristijo tako učencem kot učiteljem japonščine pri izvajanju učnega procesa, hkrati pa je tako nastali korpus lahko prvi korak proti cilju izgradnje obsežnega, javno dostopnega korpusa besedil slovenskih učencev japonščine ter napak, ki se v njih pojavljajo, in s tem gradnik za nadaljnje raziskave.

1.1. Osnovne značilnosti japonske slovnice

V raziskavi smo uporabili pojme in termine, ki jih za opis japonske slovnice uporablja Bekeš (2005).

1.1.1. Členki

Členki (*joshi* 助詞) so heterogena zvrst. V skladnji igrajo pomožno vlogo. Eno skupino členkov dodajamo samostalnikom, ki se tako spremenijo v dopolnila in se vežejo na povedek. Zopet druga skupina členkov veže med seboj besede, besedne zveze ali stavke. Glede na vlogo, ki jo členki igrajo v zgradbi stavka, jih delimo na: sklonske členke (*kaku joshi* 格助詞), tematske členke (*teidai joshi* 提題助詞), besedilne členke (*toritate joshi* 取り立て助詞), vezne členke (*setsuzoku joshi* 接続助詞), povedne členke (*shūjoshi* 終助詞) ipd.

1.1.2. Sklonski členki

Sklonski členki (*kaku joshi* 格助詞) v stavku nakazujejo, kakšno vlogo igra dopolnilo v dejanju ali stanju, ki ga izraža povedek. Dopolnila so načeloma sestavljena iz samostalnika (ali samostalniške fraze) ter iz sklonskega členka. Pri rezultatih se podrobneje omenjajo sklonski členki *ga*, *ni*, *de* in *wo*. Sklonske vloge v slovenščini se resda pomensko delno prekrivajo z vlogo dopolnil, ki jih spremljajo sklonski členki v japonščini, vendar ne popolnoma, zato ne bomo uporabljali slovenske terminologije (imenovalnik, tožilnik, dajalnik ipd.), pač pa japonska imena: sklon *ga*, sklon *wo*, sklon *ni* itd.

Sklonski členek *ni* ima podobno funkcijo kot slovenski dajalnik (primer 1), lahko pa tudi označuje kraj (primer 2).

(1) *Hon wo tomodachi ni ageru*.
knjiga-*wo* prijatelj-*ni* dati
Knjigo dam prijatelju.

(2) *Jan wa heya ni iru*.
Jan-*wa* soba-*ni* biti
Jan je v sobi.

Sklonski členek *de* で deluje podobno kot slovenski orodnik (primer 3), lahko pa podobno kot členek *ni* に označuje tudi lokacijo (primer 4).

(3) *Kuruma de iku*.
Avto-*de* iti.
Grem z avtom.

(4) *Resutoran de matte iru*.
Restavracija-*de* čakati [kontinuativ]
Čakam v restavraciji.

Sklonski členek *wo* deluje podobno kot slovenski tožilnik in po navadi stoji ob predmetu, sklonski členek *ga* pa stoji ob osebku oz. deluje podobno kot imenovalnik (primer 5).

(5) *Yan ga hon wo kau*.
Jan-*ga* knjiga-*wo* kupiti
Jan kupi knjigo.

1.1.3. Tematski členek

Členek, ki zaznamuje temo povedi, imenujemo tematski členek (*teidai joshi* 提題助詞）. Tema v povedi se običajno pojavi kot kombinacija samostalnika (samostalniške zveze) in tematskega členka. Najbolj pogost tematski členek je členek *wa* (primer 6).

(6) Novak-san wa watashi no tomodachi desu.
Novak-gospod-*wa* jaz[atributiv] prijatelj biti
Gospod Novak je moj prijatelj.
(dobesedno: Kar se tiče gospoda Novaka, je moj prijatelj.)

1.1.4. Vezni členek *no*

Členek *no* je t. i. vezni členek (*rentaijoshi* 連体助詞). Vezni členki povezujejo besede in stavke. Členek *no* povezuje dva samostalnika v samostalniško zvezo, kjer običajno izraža pripadnost, hkrati pa deluje tudi kot sklonski členek, ki ga lahko primerjamo s slovenskim rodilnikom (primer 7).

(7) *Jan no hon*
Jan-*no* knjiga
Janova knjiga

1.1.5. Glagoli

Glagol lahko samostojno uporabimo kot povedek. Kategorije pregibanja niso oseba in število kot v indoevropsklih jezikih, pač pa se glagol pregiba predvsem glede na razne vidike modalnosti.

Glede na to, kako se glagoli pregibajo, jih delimo v dve veliki skupini: enostopenjske glagole (*ichidandōshi* 一段動詞) s samoglasniško osnovo in petstopenjske glagole (*godandōshi* 五段動詞) s soglasniško osnovo.

1.1.6. Pomožni glagol oz. kopula *desu*

Pomožni glagol *desu* s samostalniki tvori samostalniške povedke in deluje kot neke vrste »obešalnik« za informacijo, ki jo sicer pri glagolskih povedkih izražajo paradigme pregibanja.

1.1.7. Formalni samostalniki

V japonščini imamo posebno skupino samostalnikov, ki so po pomenu zelo abstraktni, nanašajo se na abstraktne odnose bolj kot na konkretne pomene in jih zato ne moremo uporabljati samostojno, brez modifikatorjev. Takim samostalnikom pravimo formalni samostalniki (*keishiki meishi* 形式名詞). Stavki, ki so jim na koncu dodani formalni samostalniki, v povedi delujejo kot odvisniki (npr. dopolnilni, prislovni ipd.). Eden izmed pomembnejših je formalni samostalnik *koto* (primer 8).

(8) *Kare ga shinda koto wo shirimasen deshita*.
On-*ga* umreti[pret.] dejstvo-*wo* vedeti[nik.pret.]
Nisem vedel, da je on umrl.
(dobesedno: Nisem vedel za dejstvo, da je umrl.)

2. Pregled obstoječe literature in sorodnih raziskav

Zaradi pomanjkanja predhodnih raziskav, ki bi se ukvarjale z slovničnimi napakami slovenskih učencev japonščine, smo ob zbiranju literature najprej poskusili najti podobne raziskave, ki se ukvarjajo z napakami slovenskih učencev drugih jezikov. A ker se japonska slovnica precej razlikuje od evropskih jezikov, smo se tu namesto k predhodnim raziskavam, ki se ukvarjajo z napakami slovenskih učencev v drugih jezikih, raje zatekli k predhodnim raziskavam, ki se ukvarjajo z napakami učencev japonščine (prevladujejo predvsem analize napak kitajskih, tajvanskih, korejskih in ameriških učencev).

Pri uporabljeni literaturi gre omeniti predvsem dve sorodni deli japonskih avtorjev. Prvo je delo z naslovom *Mali slovar primerov napak v japonščini* (Ichikawa, 1997), ki kategorizira preko tisoč povedi s slovničnimi napakami na dve glavni skupini, osem podskupin in 87 posameznih slovničnih kategorij. Najširše loči napake na t. i. skupino modifikatorjev (*shūshokubu* 修飾部) in skupino prilastkov (*jutsugo* 述語). Med modifikatorje uvršča razne odvisnike (časovne, vzročne, namerne, dopolnilne itd.) in členke; skupina napak v povedkih pa vsebuje podskupine, kot so glagolski vid, glagolski način in druge kategorije povezane z glagolom. Ta način kategorizacije napak se uporablja tudi v jezikovnem korpusu učencev japonščine[2] (Umino et al. 2012), ki je zelo podobnem našemu končnemu cilju. Ta način kategorizacije napak omogoča jasen pregled nad slovničnimi elementi japonskega jezika glede na njihovo funkcijo v stavku, zato smo se tudi sami odločili za enak tip kategorizacije.

Drugo delo, ki je bilo zelo pomembno za to raziskavo, pa je *Zbirka napak v japonščini tujih učencev* (Teramura, 1990) ena pomembnejših raziskav na področju japonskega korpusnega jezikoslovja v tistem času. Delo vsebuje korpus (še v analognem zapisu) z več kot 7000 primeri stavkov iz različnih besedil, ki so jih ustvarili učenci japonščine iz različnih držav. V njih so označene in kategorizirane napake, skupek besedil pa tako tvori prvi korpus učencev japonščine.

3. Predmet raziskave

Gre za korpusno raziskavo, predmet katere so slovnične napake, ki se pojavljajo v spisih učencev japonščine na začetnem nivoju. Korpus tu tvori skupek 143 besedil, katerih avtorji so študentje, ki so v akademskem letu 2016/2017 na Oddelku za azijske študije na Filozofski Fakulteti v Ljubljani v okviru programa Japonologija obiskovali pouk pri predmetih *Japonski jezik v praksi 1* in *Sodobna japonščina 1*. Spisi niso bili napisani v testni situaciji, temveč so nastali kot domača naloga, kar pomeni, da avtorji ob pisanju niso bili pod časovnim pritiskom ter so si lahko ob pisanju besedil pomagali z raznimi učbeniki, slovarji in drugimi pripomočki. Ta informacija je pomembna, saj to dejstvo zmanjšuje verjetnost, da je do napak v spisih prišlo zaradi površnosti, temveč zaradi dejanskega pomanjkanja znanja o določenem slovničnem elementu oz. njegovi uporabi. Pri analizi napak pa smo se omejili zgolj na slovnične, saj bi analiza ostalih tipov napak, kot na primer ortografske, stilistične in napake, povezane z pisanjem pismenk itd. zahtevala popolnoma drugačen pristop.

Spisi, ki tvorijo korpus, zajemajo devet tem. Prva izmed tem je »opis sobe«; tu učenec opisuje lastno sobo ter predmete v njej, zaradi česar je v teh spisih poudarek predvsem na opisovanju lokacij premetov v prostoru in uporabi ustreznih glagolov stanja. Druga izmed tem je »družina«; tu je poudarek na naštevanju in opisovanju lastnosti družinskih članov in rabi pridevnikov ter vezni obliki le-teh. Tretja tema je »hobi«; tu je poudarek na naštevanju oz. nizanju glagolov in nominalizaciji le-teh. Četrta tema je dnevnik aktivnosti v času programa SloTan[3]; tu je poudarek na opisovanju preteklih dogodkov, uporabi preteklih oblik glagolov in pridevnikov.[4] Tri tematsko podobne teme pa predstavljata dva dnevnika branja in pa predstavitev knjige. V teh spisih učenci obnavljajo in komentirajo vsebine prebranih knjig. Spisi torej obsegajo različne teme, ki zahtevajo uporabo različnih slovničnih struktur. Le-to je za takšen tip raziskave pomembno, saj bi v nasprotnem primeru, zaradi precej majhne velikosti korpusa negativno vpliva na verodostojnost rezultatov (prevladuje namreč samo raba nekaterih slovničnih elementov).

Tekom same raziskave je bila izvedena anketa, s pomočjo katere so bili pridobljeni nekateri osnovni podatki o učencih. Tu gre predvsem za jezikovno ozadje. Povprečna starost učencev je bila 20 let; 80 % učencev se japonskega jezika ni učilo pred vpisom v prvi letnik; pri vseh učencih gre za naravne govorce slovenskega jezika, ki po lastni oceni angleški jezik obvladajo vsaj na nivoju B1, ter v povprečju govorijo vsaj še en drug tuj jezik; kot odgovor na vprašanje, kateri del japonske slovnice jim po lastnem mnenju povzroča največ težav, pa je velika večina učencev navedla členke. To je zanimivo, saj nas zanima, v kolikšni meri se učenci sami zavedajo svojih šibkosti.

4. Metoda raziskave

Na roko napisane spise smo najprej digitalno posneli in shranili kot slike. Nato smo s pomočjo programa Excel ustvarili tabelo s šestimi stolpci, ki je služila kot ogrodje za naš korpus napak. V prvi stolpec smo prepisali stavek, takšen kot se je pojavil v spisu, vključno z napakami, katere smo posebej označili. V drugi stolpec smo vpisali slovnično pravilno različico. V tretjem stolpcu smo označili kategorijo slovnične napake po Ichikawi. V četrtega temo spisa, oštevilčeno od ene do devet. V petega smo vpisali zaporedno številko vrstice, v kateri se napaka v spisu pojavi, in v zadnjega dvomestno šifro, dodeljeno vsakemu izmed učencev.

V tej raziskavi napak ne kategoriziramo glede na napačno uporabljen slovnični element, temveč glede na element, ki bi bil moral biti uporabljen. Tu zagovarjamo stališče, da je bolj kot to, da je učenec na določenem mestu uporabil napačen vzorec, pomembno to, da ni uporabil pravilnega, saj le-to pomeni, da učencu primanjkuje znanja o tem slovničnem elementu.

Nato smo vsakemu primeru napake določili šestmestno šifro, pri kateri prvi dve številki označujeta temo spisa, srednji dve označujeta avtorja in zadnji dve vrstico, v kateri se pojavlja napaka. Zaporedno število vrstice se tu nanaša

[2] http://cblle.tufs.ac.jp/llc/ja/index.php?menulang=en

[3] Kratek program za japonsko govoreče študente na izmenjavi na FF UL, ki ga vsako leto organizira oddelek AŠ. V okviru programa imajo japonski in slovenski študentje predstavitve o raznih temah, debatirajo, organizirajo vodene oglede idr.

[4] V japonščini se pridevniki pregibajo glede na čas.

na zaporedne vrstice v skeniranih verzijah spisov, pisanih na roko. V kolikor bi spisi bili pisani računalniško, bi namesto zaporednega števila vrstice uporabili zaporedno številko povedi.

Tako ustvarjeno tabelo smo nato analizirali s pomočjo Excelovih funkcij. Tako smo dobili podatke o najštevilnejših napakah. A pri teh rezultatih je šlo zgolj za podatke o najštevilnejših napakah in ne najpogostejših. Ker se na primer sklonski členki, kot na primer členek *ga*, ki v stavku označuje osebek, v povprečju v tekstih pojavljajo precej pogosteje kot na primer pogojni členek *tara*, je samoumevno, da bo napak pri uporabi takšnega slovničnega elementa več. Zato nam samo število napak tako ni povedalo še ničesar. Zanimale so nas namreč najpogostejše napake.

Da bi pridobili podatke o najpogostejših napakah, smo morali tu ugotoviti razmerje med številom primerov, ko bi določen slovnični element moral biti uporabljen, a ni bil, in pa med številom primerov, v katerih je oz. bi moral ta slovnični element oz. vzorec biti uporabljen, da bi bil stavek slovnično pravilen. Zanimalo nas je torej razmerje med količino pravilne in napačne rabe posameznega elementa. Za ta izračun smo potrebovali podatke o tem, kolikokrat je oz. bi moral biti posamezen slovnični element oz. vzorec uporabljen znotraj našega korpusa. Štetja nismo mogli izvesti strojno, saj lahko v japonskem jeziku enaki členki imajo več različnih funkcij, kot na primer členek *de*, ki lahko v naslednjem kontekstu označuje orodje (*kuruma de – z avtom*) ali pa lokacijo (*kuruma de – v avtu*). Zato je moral ta del analize biti izveden ročno.

Ko smo pridobili vse podatke, smo izračunali razmerje med pravilno in napačno rabo oz. pogostost pojavljanja napak pri določenem slovničnem elementu. V dobljenih rezultatih so izstopali slovnični vzorci, ki se v spisih skoraj niso pojavljali. Razlog za to je bil namreč ta, da je šlo za vzorce, ki so bili oz. bi morali biti znotraj spisov uporabljeni tako malokrat, da je bil vzorec napak premajhen, da bi bili podatki lahko verodostojni. Zaradi tega smo tu določili dodatni kriterij, oz. spodnjo mejo. Pogoj, da se napaka lahko upošteva kot kandidat za najpogostejšo napako, je ta, da mora količina primerov te napake presegati 1 % skupnega števila vseh napak znotraj korpusa, ki znaša 516, kar pomeni več kot pet primerov tega tipa napake. Po upoštevanju tega kriterija so pridobljeni rezultati delovali veliko bolj realistično.

5. Rezultati raziskave

Po uporabi dodatnega kriterija za določitev najpogostejših tipov napake smo ponovno določili deset najpogostejših tipov napak. V skupini desetih najpogostejših napak so kar šest mest zasedli členki. Ti členki so bili: tematski členek *wa*, vezni členek *no* in sklonski členki: *ga*, *ni*, de in *wo*. Ostale štiri najpogostejše tipe napak so predstavljale napake povezane z vezno obliko glagola (*dōshi heiritsu* 動詞並立), z glagolom stanja *aru*, kopulo *desu* ter s formalnim samostalnikom *koto*.

SLOVNIČNI ELEMENT	**ODSTOTEK NAPAČNE RABE**
vezna oblika glagolov V-te	17,24 %
formalni samostalnik *koto*	16,67 %
sklonski členek *wo*	10,56 %
sklonski členek *de*	9,95 %
sklonski členek *ga*	8,96 %
sklonski členek *ni*	8,36 %
glagol stanja *aru*	7,27 %
vezni členek *no*	7,17 %
tematski členek *wa*	5,83 %
kopula *desu*	5,46 %

Tabela 1: Deset najpogostejših tipov napak

6. Diskusija

Kar se tiče samih členkov, sklepamo, da je razlog, da se jih je med najpogostejše napake uvrstilo toliko, najbrž dejstvo, da v slovenskem jeziku stavčne vloge izražamo s sklanjanjem in ne z dodajanjem členkov kot v japonščini, zato je takšen način izražanja funkcij besed slovenskim učencem tuj ter posledično težje razumljiv. Hkrati pa japonski členki slovenskim sklonom ne ustrezajo vedno popolnoma. Medtem ko se na primer za izražanje pripadnosti in rodilnika nedvoumno uporablja členek *no*, in v primeru dajalnika členek *ni*, imamo v primeru mestnika oz. izražanju lokacije možnost uporabe členka *de* in členka *ni*. V primeru imenovalnika uporabljamo členek *ga*, katerega pa mnogokrat lahko nadomestimo z tematskim členkom *wa*, ki določuje temo stavka. Takšni pari členkov so slovenskim učencem japonščine še posebej težavni. Kot rešitev oz. način za zmanjšanje pogostosti pojava takšnih napak lahko predlagamo predvsem, da učitelj ob razlagi problematičnih členkov le-te predstavi skupaj z njihovimi pari ter pri razlagi več pozornosti posveti poudarku na razliki med podobnima členkoma, predvsem s pomočjo prevodov v slovenščino.

Glagol stanja *aru* deluje podobno kot slovenski glagol *biti*, le da se le-ta uporablja le v primeru, da gre pri osebku stavka za neživo stvar ali rastlino (torej ne za ljudi ali živali, saj se zanje uporablja glagol *iru*). Kot omenjeno v točki 1.1, se kopula *desu* prav tako uporablja podobno kot glagol *biti*. Tako ima učenec ob tvorbi stavka na voljo tri različne možnosti, ki pa naravnemu govorcu slovenščine (kjer kopuli *desu* podobnega slovničnega elementa ni) delujejo zelo podobno (v nekaterih primerih lahko v določenem stavku glagole stanja *iru* in *aru* tudi nadomestimo z *desu*), zaradi česar pride do napak. Ker moramo v nekaterih primerih kopulo *desu* pripeti stavkom, ki na prvi pogled izgledajo slovnično popolni tudi brez nje, tudi primerov, ko je do napake prišlo zaradi neuporabe kopule, ni bilo malo.

Kar se tiče napak pri formalnem samostalniku *koto*, je verjetno razlog za težave ta, da ta slovnični element v japonskem jeziku deluje precej drugače kot posamostaljenje v slovenščini. V slovenščini glagol nominaliziramo besedotvorno (delati – delo), v japonskem jeziku pa glagol nominaliziramo tako, da mu pripnemo t. i. formalni samostalnik *koto*. V največ primerih je do napake prišlo zaradi neuporabe tega slovničnega elementa. Tudi tu gre predvsem za to, da učenci niso navajeni na rabo tega slovničnega elementa, ker v slovenskem jeziku nominalizacija deluje drugače. Tudi ta napaka bi morala s časom samodejno izginiti.

Procentualno najpogostejši tip napak je bil tip napak povezan s. t. i. vezno obliko glagolov V-*te*. Pri tem tipu napak gre predvsem za napake pri spreganju glagolov (predvsem pet-stopenjskih), oz. pomanjkanju rabe te slovnične strukture. Menimo, da je do večine napak prišlo, ker se učenci teh glagolskih oblik učijo dokaj pozno, zaradi česar niso navajeni na njihovo rabo.

Pri samih rezultatih moramo seveda omeniti tudi dejstvo, da korpus vendarle obsega le 143 besedil, kar vpliva na verodostojnost samih rezultatov (zaradi česar je tudi prišlo do težav, omenjenih v točki 4). A kljub temu menimo, da so rezultati in ugotovitve te raziskave, pridobljeni po uporabi dodatnega kriterija, verodostojni in lahko služijo kot vodilo pri načrtovanju učenja.

Način kategorizacije se ni izkazal za najprimernejšega. Razlog za to je, da Ichikawa te kategorizacije ne uporablja za analiziranje napak znotraj korpusa, temveč samo za prikaz najpogostejših tipov napak. Tako smo v primeru redkejših tipov napak bili primorani sami dodati nove kategorije. Zaradi tega bo v primeru nadaljevanja raziskave potrebno primerno prilagoditi Ichkawino kategorizacijo, ali pa uporabiti kakšno primernejšo kategorizacijo napak.

7. Zaključek in bodoče delo

V sklopu te raziskave smo ustvarili korpus 143 besedil slovenskih učencev japonščine ter na podlagi analize le-tega ugotovili, kateri slovnični elementi slovenskim učencem japonščine povzročajo največ težav. S tem smo naredili prvi korak proti končnemu cilju izgradnje obsežnejšega korpusa besedil učencev japonščine, ki bo učiteljem jezika omogočil vpogled v tipe napak, ki se pojavljajo pri učencih, in s tem izpostavil slovnične elemente, ki jim je treba pri uvajanju snovi nameniti več časa, ter predstavljal osnovo za bodoče raziskave na tem področju, tako v slovenskem prostoru kot širše v primerjavi z učenci japonščine iz drugih jezikovnih okolij.

Zaenkrat so spisi digitalizirani zgolj kot slike, digitalizirani pa so le stavki, ki vsebujejo napake. Tako je eden izmed nadaljnjih ciljev najprej popolna digitalizacija vseh 143 besedil, kar bo omogočalo lažji pregled, več možnosti analize in pa vpogled v kontekst posamezne povedi, kar je pri visoko kontekstualnem jeziku, kot je japonščina, zelo pomembno za razumevanje vsebine povedi.

Še eden izmed ciljev je razširiti sam korpus, s čimer bomo povečali verodostojnost statističnih podatkov. Ker so v korpus vključena zgolj besedila učencev na osnovni ravni, bomo poskušali vključiti najprej besedila učencev na srednji in nato še na višji ravni. S tem nameravamo omogočiti razne primerjave tipov napak, ki se pojavljajo na posamezni ravni, z nadaljnjim osvajanjem jezika pa izginjajo ali ostajajo. Prav tako pa nameravamo v prihodnosti povečati tudi količino vključenih besedil učencev na osnovni ravni.

V primeru razširitve korpusa bo potreben hitrejši in enostavnejši način obdelave oz. analize podatkov, zaradi česar nameravamo korpus popolnoma digitalizirati, vsem besedilom dodati oznake posameznih tipov napak, ter celoten korpus objaviti tako za dostop preko spletnega konkordančnika kot tudi v javno dostopnem repozitoriju jezikovnih virov.

8. Literatura

Bekeš, Andrej. 2005. *Pregled slovnice japonskega jezika (skripta).* Oddelek za azijske in afriške študije, Filozofska fakulteta Univerze v Ljubljani.

Foreign Service Institute, 2007. *Learning Expectations*. http://web.archive.org/web/20071014005901/http://www.nvtc.gov/lotw/months/november/learningExpectations.html Dostop: 10. 4. 2018.

Ichikawa Yasuko 市川保子. 1997. *Nihongo goyou reibun shoujiten* 日本語誤用例文小辞典 [*Mali slovar primerov napak v japonščini*]. Tokyo: Bonjinsha.

NINJAL (n.d.) Learner Corpus Study of Aquisiton of Japanese as a Second Language, http://lsaj.ninjal.ac.jp/, Dostop: 10. 4. 2018.

Teramura, Hideo 寺村秀夫. 1990. *Gaikokujingakushuusha no nihongo goyoureishuu* 外国人日本語学習者の日本語誤用例集 [*Zbirka napak tujih učencev japonščine*], http://teramuradb.ninjal.ac.jp/teramura.goyoureishu.pdf, Dostop: 15. 1. 2018.

Umino, Tae et al. 2012. *Learners' Language Corpus of Japanese*. Tokyo University of Foreign Studies. http://cblle.tufs.ac.jp/llc/ja/index.php?menulang=en.

Building a corpus of the Croatian parliamentary debates using UDPipe open source NLP tools and Neo4j graph database for creation of social ontology model, text classification and extraction of semantic information

Benedikt Perak,[*] Filip Rodik,[†]

[*] Department of Cultural Studies, Faculty of Humanities and Social Sciences, University of Rijeka
Sveučilišna avenija 4, 51000 Rijeka, Croatia
bperak@uniri.hr
[†] Tune informacijske tehnologije d.o.o.,
Levanjska 5, 10040 Zagreb, Croatia
filip.rodik@gmail.com

Abstract

This paper describes a process of creating morphosyntactically tagged corpus of the Croatian parliamentary debates using NLP tool UDapi for tokenization, morpho-syntactic parsing and processing Universal Dependencies data to process over 300 thousand transcribed parliamentary speech utterances produced over the period from 2003-2017 and store the data in a Neo4j graph database.

Introduction

This paper[1] describes a pipeline for creating morphosyntactically tagged corpus of the Croatian parliamentary debates using open source NLP tool UDapi (https://github.com/udapi) for tokenization, morpho-syntactic parsing and processing Universal Dependencies data. The pipeline was used to process over 300 thousand transcribed parliamentary speech utterances produced over the period from 2003-2017 and store the data in a Neo4j graph database. The aim of using the graph database is to create a complex representation of the social ontology of the political behaviour involving various social entities, communication processes as well as to apply basic statistic summarization, text classification, community and centrality graph analytics for the research of social, linguistic and conceptual networks.

This computational linguistic and data science research is valuable for the humanities because these parliamentary texts represent one of the biggest available transcribed corpus of public speech in Croatian, while the graph analytics and social model of communication can be valuable for the research in different social sciences because The Croatian Parliament (Croatian: Hrvatski sabor) is one of the most important representative and legislative body of the citizens of the Republic of Croatia. The Parliament is composed of 151 members elected to a four-year term that convene regularly twice a year, the first session runs between 15 January and 15 July, while the second session runs from 15 September to 15 December. The Croatian Parliament can also hold extraordinary sessions (http://www.sabor.hr/Default.aspx?sec=713).

The parliament debates are transcribed and published on the http://www.sabor.hr/ web site. The site comprises of current debates in the 9th term of the Parliament, along with the material from the previous 5th, 6th, 7th and 8th terms of the Parliament, covering sessions from the year 2003-2017. However, this type of repository is not suitable for extensive analysis of the communicative or linguistic features of the delivered speeches. Therefore, we developed a pipeline for tokenization, lemmatization, syntactic parsing of dependencies and meta data integration for creation of complex queries and exploration of linguistic features related to speakers, topics, and sessions.

Goal of the paper

The goal of the paper is to present tools, methods and resources used for the a) data harvesting and extraction of the Croatian Parliament speeches, b) tokenization, lemmatization and syntactic parsing of the files, and c) data storing, modelling and integration. The structure of the paper follows these steps.

Data gathering

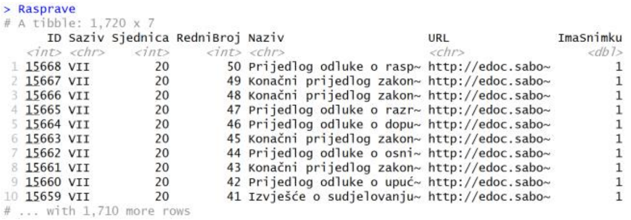
The texts of the Croatian parliamentary debates corpus are gathered using a RSelenium scraper (https://github.com/ropensci/RSelenium) on the Parliament web-repository (http://edoc.sabor.hr/). The data gathering process is published as a github

[1] This research is part of the EmoCNet project http://emocnet.uniri.hr, supported by the University of Rijeka's initial grant for the researchers 2017-18.

project (https://github.com/rodik/Sabor). The debates of the 5th to 9th Parliamentary Assembly are downloaded as datasets in a CSV format (Table 1-2). The structure of the data representation on the web-repository yielded two datasets – a session dataset and a transcripts dataset, each with unique metadata features.

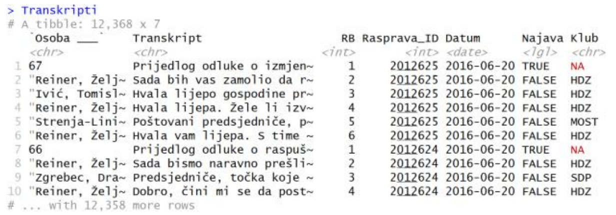
The sessions dataset (table 1.) collected the information about the Parliamentary sessions with features: 1) unique ID, 2) number of the parliamentary assembly, 3) number of the parliamentary session, 4) identifying number, 5) title of the session, 6) url of the session, 7) logical value on the existence of the recording (illustration 2).

The transcripts dataset (table 2.) harvested the transcripts data with the following features: 1) person, 2) transcript, 3) number of the utterance, 4) unique ID of the session, 5) date, 6) announcement, 7) parliamentary club.



```
> Rasprave
# A tibble: 1,720 x 7
      ID Saziv Sjednica RedniBroj Naziv                     URL                  ImaSnimku
   <int> <chr>    <int>     <int> <chr>                     <chr>                    <dbl>
 1 15668 VII         20        50 Prijedlog odluke o rasp~ http://edoc.sabo~           1
 2 15667 VII         20        49 Konačni prijedlog zakon~ http://edoc.sabo~           1
 3 15666 VII         20        48 Konačni prijedlog zakon~ http://edoc.sabo~           1
 4 15665 VII         20        47 Prijedlog odluke o razr~ http://edoc.sabo~           1
 5 15664 VII         20        46 Prijedlog odluke o dopu~ http://edoc.sabo~           1
 6 15663 VII         20        45 Konačni prijedlog zakon~ http://edoc.sabo~           1
 7 15662 VII         20        44 Prijedlog odluke o osni~ http://edoc.sabo~           1
 8 15661 VII         20        43 Konačni prijedlog zakon~ http://edoc.sabo~           1
 9 15660 VII         20        42 Prijedlog odluke o upuć~ http://edoc.sabo~           1
10 15659 VII         20        41 Izvješće o sudjelovanju~ http://edoc.sabo~           1
# ... with 1,710 more rows
```

Table 1: Example of sessions dataset CSV file.



```
> Transkripti
# A tibble: 12,368 x 7
   `Osoba ___`      Transkript                  RB Rasprava_ID Datum      Najava Klub
   <chr>            <chr>                    <int>       <int> <date>     <lgl>  <chr>
 1 67               Prijedlog odluke o izmjen~     1     2012625 2016-06-20 TRUE   NA
 2 "Reiner, Želj~ Sada bih vas zamolio da r~     2     2012625 2016-06-20 FALSE  HDZ
 3 "Ivić, Tomisl~ Hvala lijepo gospodine pr~     3     2012625 2016-06-20 FALSE  HDZ
 4 "Reiner, Želj~ Hvala lijepa. Žele li izv~     4     2012625 2016-06-20 FALSE  HDZ
 5 "Strenja-Lini~ Poštovani predsjedniče, p~     5     2012625 2016-06-20 FALSE  MOST
 6 "Reiner, Želj~ Hvala vam lijepa. S time ~     6     2012625 2016-06-20 FALSE  HDZ
 7 66               Prijedlog odluke o raspuš~     1     2012624 2016-06-20 TRUE   NA
 8 "Reiner, Želj~ Sada bismo naravno prešli~     2     2012624 2016-06-20 FALSE  HDZ
 9 "Zgrebec, Dra~ Predsjedniče, točka koje ~     3     2012624 2016-06-20 FALSE  SDP
10 "Reiner, Želj~ Dobro, čini mi se da post~     4     2012624 2016-06-20 FALSE  HDZ
# ... with 12,358 more rows
```

Table 2: Example of transcript dataset CSV file.

Initial data integration and modelling

The aim of the project is to integrate existing data with an ontology that can intuitively represent the entities and their relations in the process of the Parliamentary debates and possible future data enrichments with some other informational structures and corpora. Two csv datasets were integrated with a Python script using Py2Neo, a client library and toolkit for working with Neo4j from within Python applications (https://py2neo.org/v4/). Neo4j is is one of the most used open-source, fully transactional database, a persistent Java engine where it is possible to store structures in the form of graphs instead of tables (Webber, 2012). It has its own programmatic Cypher language, created by the Neo4j company for developing unique approach to graph query methods. Through the combination of Python code and Cypher queries this language that we can store and get the data from the graph database (Panzarino, 2014). The ontology of the data is represented in the illustration 1. It has 6 structurally different nodes stored with different properties and labels. These structures are connected using the partonomic type of description: 1) Parliament Assembly HAS Session with unique ID, 2) Session HAS Number of the parliamentary session, 3) Number of the parliamentary session HAS Utterance, 4) Person IS_MEMBER_OF Parliamentary club, and a process type description: 5) Person DELIVERED Utterance.

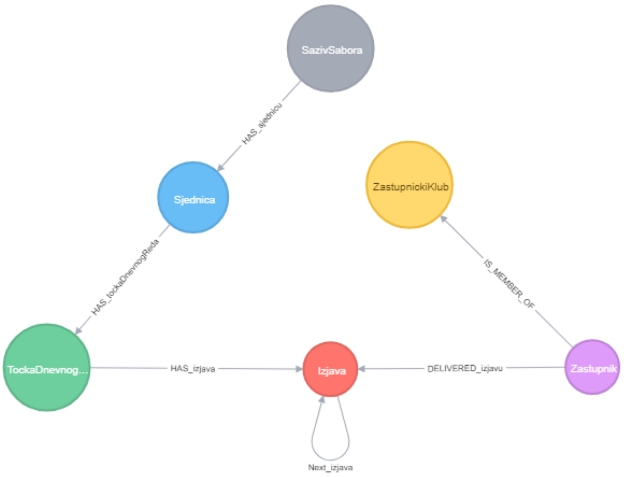
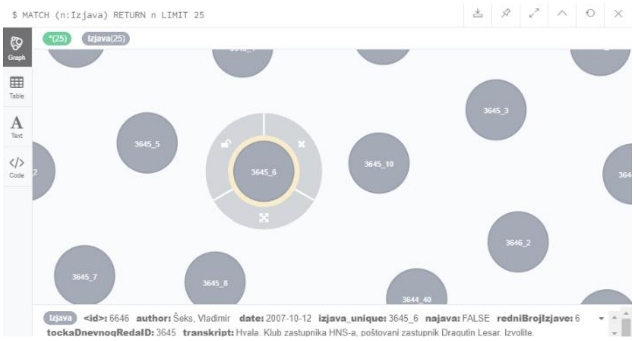


Figure 1. Ontological model of the sessions and trancrips datasets: Parliamentary Assembly – HAS –> Session – HAS –> Discussion point –> HAS –> Utterance <– DELIVERED – Representative –> IS_MEMEBER_OF –> Parliamentary club



```
$ MATCH (n:Izjava) RETURN n LIMIT 25
```

Izjava <id>: 6646 author: Šeks, Vladimir date: 2007-10-12 izjava_unique: 3645_6 najava: FALSE redniBrojIzjave: 6 tockaDnevnogRedaID: 3645 transkript: Hvala. Klub zastupnika HNS-a, poštovani zastupnik Dragutin Lesar. Izvolite.

Figure 2. Screenshot of the Neo4j graph data base application browser presenting random 25 nodes labelled *Izjava* 'Utterances' with respective properties.

Corpus creation

The data from the transcripts have been extracted and stored as a batch of separate files with unique identifier of the session and number of the utterance, for example: 2012726_35.txt. These files have been sent to local installation of the UDPipe (http://ufal.mff.cuni.cz/udpipe), and specifically R package for Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing Based on the UDPipe Natural Language Processing Toolkit (https://bnosac.github.io/udpipe). The model used for parsing was croatian-ud-2.0-170801.udpipe from Universal Dependencies 2.0 Models for UDPipe repository at (https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2364). The parsed files were stored with _conllu.txt endings, for example: 2012726_35_conllu.txt. The output of each transcribed utterance uses a revised version of the CoNLL-X format called CoNLL-U. Annotations are encoded in plain UTF-8 encoded text files with three types of lines: word lines containing the annotation of a word/token in 10 fields separated by single tab characters. Blank lines marking sentence boundaries. Comment lines starting with hash (#). The 10 fields are respectively: 1) ID: Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes. 2) FORM: Word form or punctuation symbol. 3) LEMMA: Lemma or stem of word form. 4) UPOS: Universal part-of-speech tag. 5) XPOS: Language-specific part-of-speech tag; underscore if not available. 6) FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available. 7) HEAD: Head of the current word, which is either a value of ID or zero (0). 8) DEPREL: Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one. 9) DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs. 10) MISC: Any other annotation. (http://universaldependencies.org/format.html). The example of the structure is the following:

```
# newdoc id = doc1
# newpar
# sent_id = 1
# text = Poštovani predsjedniče, poštovani premijeru članovi Vlade uz čestitke za pozitivnom aviju.
1	Poštovani	poštovan	ADJ	_	Case=Nom|Definite=Def|Degree=Pos|Gender=Masc|Number=Plur	2	amod	_	_
2	predsjedniče	predsjednik	NOUN	_	Case=Voc|Gender=Masc|Number=Sing	0	root	_	SpaceAfter=No
3	,	,	PUNCT	_	_	4	punct	_	_
4	poštovani	poštovati	ADJ	_	Case=Nom|Definite=Def|Degree=Pos|Gender=Masc|Number=Plur|VerbForm=Part	2	acl	_	_
5	premijeru	premijer	NOUN	_	Case=Dat|Gender=Masc|Number=Sing	4	iobj	_	_
6	članovi	član	NOUN	_	Case=Nom|Gender=Masc|Number=Plur	4	nsubj	_	_
7	Vlade	Vlada	NOUN	_	Case=Gen|Gender=Fem|Number=Sing	6	nmod	_	_
8	uz	uz	ADP	_	Case=Acc	9	case	_	_
9	čestitke	čestitka	NOUN	_	Case=Acc|Gender=Masc|Number=Plur	6	nmod	_	_
10	za	za	ADP	_	Case=Acc	12	case	_	_
11	pozitivnom	pozitivan	ADJ	_	Case=Ins|Definite=Def|Degree=Pos|Gender=Fem|Number=Sing	12	amod	_	_
12	aviju	avija	NOUN	_	Case=Acc|Gender=Fem|Number=Sing	9	nmod	_	SpaceAfter=No
13	.	.	PUNCT	_	_	2	punct	_	_
```

From these files additional two structures: a) Sentences and b) Tokens, were created in the graph database. The Sentences nodes were connected to the utterances nodes using the HAS_Sentence relation, and every sentence in a utterance was connected with the NEXT_sentence relation (illustration 3)

The words of a sentences have been stored as Token nodes with all ten data fields from the UDPipe parser as properties. Each token relates to a Sentence node with HAS_token relation and stores mutual dependency information with other Tokens in a sentence using HAS_dependency relation. The graph representation is depicted in the figure 4.

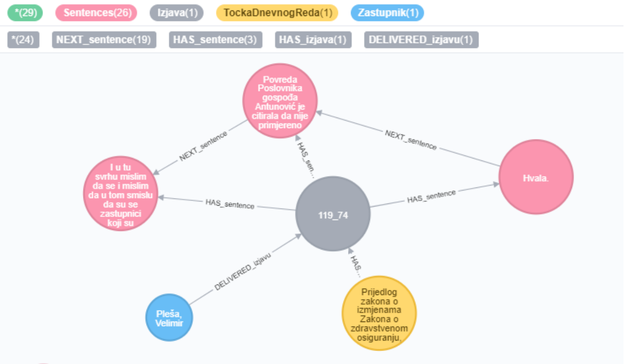
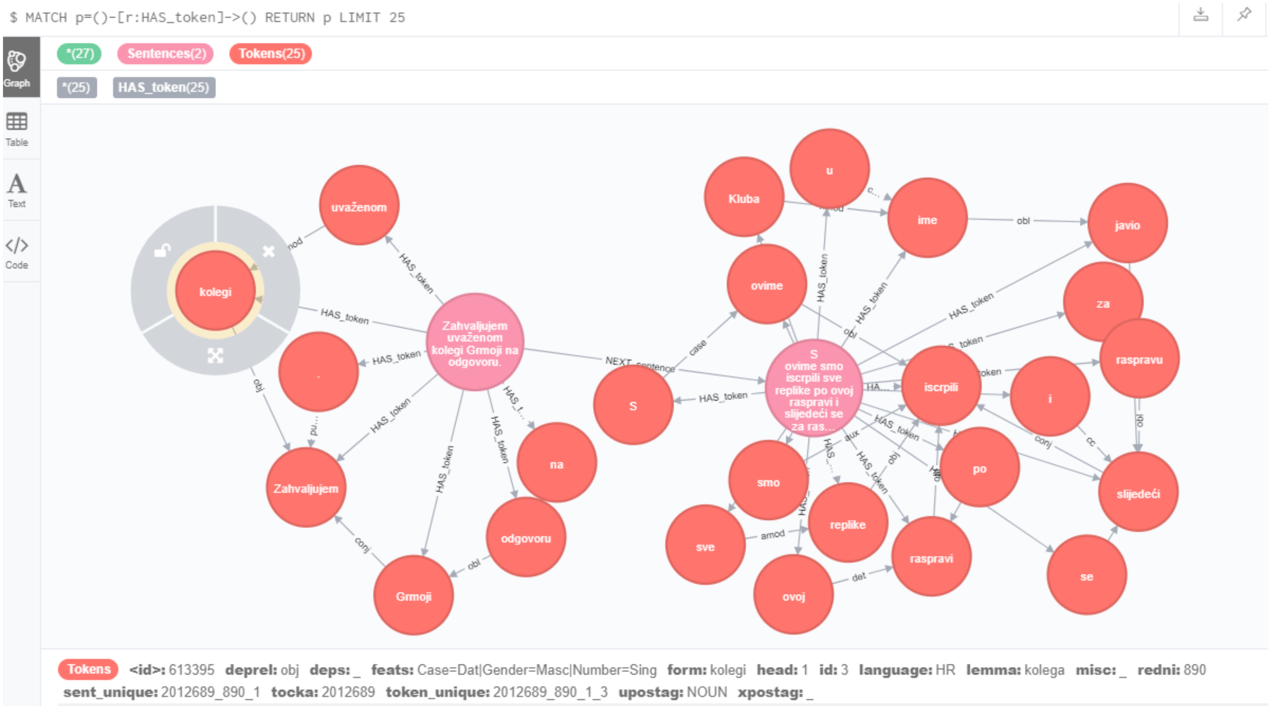


Figure 3 Screenshot of the Neo4j graph data base application browser representing relation of the nodes labelled Izjava 'Utterances' to Sentence.



```
$ MATCH p=()-[r:HAS_token]->() RETURN p LIMIT 25
```

Tokens <id>: 613395 deprel: obj deps: _ feats: Case=Dat|Gender=Masc|Number=Sing form: kolegi head: 1 id: 3 language: HR lemma: kolega misc: _ redni: 890 sent_unique: 2012689_890_1 tocka: 2012689 token_unique: 2012689_890_1_3 upostag: NOUN xpostag: _

Figure 5 Screenshot of the Neo4j graph data base application browser representing random token nodes of two Sentences with respective properties and dependency relations.

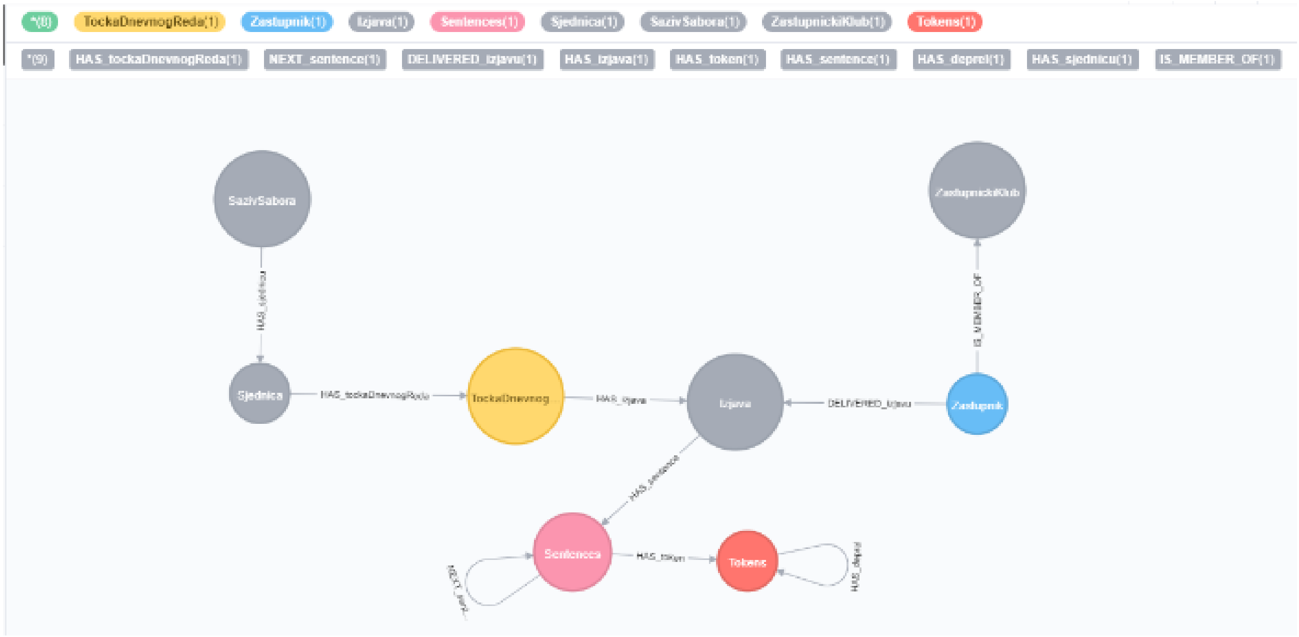


Figure 4. Ontological model stored in the Neo4j database: Parliamentary Assembly – HAS –> Session – HAS –> Discussion point –> HAS –> Utterance (-HAS->Sentence– HAS->Token)<– DELIVERED – Representative –> IS_MEMEBER_OF –> Parliamentary club

The final ontology in the database has the structure as depicted in the figure 5.

Further research of the application

The graph database enables highly connected storage of the disparate datasets thus allowing for an intuitive and yet complex structural development of the data according to the custom created ontology. Besides creating the statistical summarization of the entities, the graph data structure allows creation of complex queries about relations between the interconnected levels within a single text or for multiple texts. In this manner, a local corpus with universally described features can be created allowing for the analysis of the various informational features within the patterns that form the linguistic corpus and its metadata. A special feature of the Neo4j graph database is related to the native graph algorithms library (https://neo4j.com/developer/graph-algorithms/) that extends the basic summarization procedures to allow the community detection and centrality tagging for any given set of patterns. Furthermore, the graph storage of the parsed text enables the data enrichment for each level of the entities and relations. This means that the level of texts can be enriched with connections the new structures (mentioned Persons, Institution, and Organization) that can be used for further ontological description and contextualization of the text (Perak forthcomming).

References

Onofrio Panzarino2014. *Learning Cypher*. Packt Publishing Ltd.

Benedikt Perak (forthcoming) "Ontological and constructional approach to the discourse analysis of the commemorative speeches in Croatia". In: *Framing the Nation*. Pavlaković, Vjeran, Pauković, Davor (eds.) Routledge.

Jim Webber 2012. A programmatic introduction to neo4j. In: *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity* (pp. 217-218). ACM.

https://neo4j.com/developer/graph-algorithms

https://bnosac.github.io/udpipe

http://ufal.mff.cuni.cz/udpipe

https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2364

https://github.com/ropensci/RSelenium

https://github.com/rodik/Sabor

https://github.com/udapi

http://www.sabor.hr/Default.aspx?sec=713

Samopromocija na Instagramu: Primer predsednikovega profila

Dan Podjed,* Ajda Pretnar†

* Inštitut za slovensko narodopisje
Znanstvenoraziskovalni center SAZU
Novi trg 2, 1000 Ljubljana
dan.podjed@zrc-sazu.si

† Laboratorij za bioinformatiko
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
ajda.pretnar@fri.uni-lj.si

Povzetek

Prispevek predstavi profil slovenskega predsednika na družbenem omrežju Instagram. Na podlagi kvalitativne in kvantitativne analize podatkov avtorja pojasnita, kakšne samopromocijske strategije so na tem omrežju najbolj uspešne in katere predsednikove objave imajo največji odziv med sledilci. Izpostavita tri kategorije objav: 1. upodabljanje z zvezdniki in družino, 2. ustvarjanje ljudskosti, 3. nenavadne in izstopajoče samoupodobitve. S tovrstnimi objavami predsednik ustvarja prepoznavno osebnost, ki obstaja tako v fizični kot digitalni realnosti, ter nadgrajuje svoje javno predstavljanje političnega udejstvovanja z drugimi dejavnostmi, od športa in razvedrila do družine.

Self-Promotion on Instagram: A Case of President's Profile

The paper presents the Instagram profile of the Slovenian president. On the basis of a combination of qualitative and quantitative data analysis, the authors explain which self-promotion strategies are the most successful on the social network and which posts have the greatest response among president's followers. In this context, three categories of posts are highlighted: 1. portrayal of the president with celebrities and family, 2. creating an impression of being approachable to public, 3. unusual and conspicuous self-depictions. With such posts on the profile, the president creates a recognisable personality that exists both in physical and digital reality, exceeds the dullness of politics, and upgrades it with other activities, from sports and entertainment to the family.

1. Uvod

»Kralj Instagrama« in »instagramski predsednik« sta naziva, ki si ju je v domačih in tujih medijih prislužil slovenski predsednik Borut Pahor (Jager 2017; Associated Press 2017). Na predsedniški položaj se je prvič povzpel decembra 2012, profil na Instagramu pa je začel uporabljati septembra 2013. Od takrat do začetka leta 2018 je bilo na njegovem profilu objavljenih več kot 600 fotografij in posnetkov, predsednik pa je v tem času pridobil več kot 50 tisoč sledilcev, ki spremljajo, kaj se dogaja v javnem in zasebnem življenju najvišjega političnega predstavnika Republike Slovenije in vrhovnega poveljnika Slovenske vojske.

V prispevku predstavljava analizo Pahorjevega profila na Instagramu in ob tem ugotavljava, kako se predsednik s pomočjo družbenega omrežja predstavlja v vsakdanjem življenju (prim. Goffman, 2014) ter skrbi za samopromocijo. Tovrstna analiza je pomembna za boljše razumevanje sodobne politične scene in načina življenja v skupnosti, v kateri se fizične in digitalne vsebine neločljivo povezujejo. Iz prispevka je namreč razvidno, kako ljudje – posebej pa izstopajoči posamezniki – predstavljajo sebe po omrežjih ter kako se pri tem prepletata njihovo javno in zasebno življenje.

2. Teoretski razmislek o Instagramu

Pomenljivo in pomembno je, da si je predsednik za osrednjo (samo)promocijsko platformo izbral Instagram. Če sledimo misli medijskega teoretika Marshalla McLuhana (1964), medij sam po sebi vsebuje specifično sporočilo, njegove omejitve in prednosti pa zakoličijo način sporočanja in predstavljanja v javnosti. Uporabniki Twitterja, denimo, so bolj kot na vizualna sporočila vezani na kratke tekste, ki so bili v preteklosti omejeni na 140 znakov, od leta 2017 pa je njihova dolžina lahko dvakrat daljša. Sporočila na tem mediju, ki ga redno uporablja aktualni ameriški predsednik Donald Trump, so posledično kratka in samoomejujoča, zaradi česar včasih delujejo ekspresivno in eksplozivno, kot bi bila napisana v naglici. Facebookova komunikacijska platforma dopušča ljudem več prostora in svobode pri pisanju ter dodajanju fotografij in posnetkov, zaradi česar udarnost sporočil pogosto razvodeni. Instagram se razlikuje tako od Facebooka kot od Twitterja, saj se osredotoča na uporabnike mobilnih telefonov, s katerimi ti fotografirajo, kaj se dogaja okoli njih ali pa se ovekovečijo s t.i. *selfiji* (Deeb-Swihart et al., 2017; Diefenbach in Christoforakos, 2017; Gómez Cruz in Thornham, 2015; Senft in Baym, 2015). Glavna vsebina na omrežju Instagram torej ni tekst, temveč slika.

Instagramovo omrežje, ki je nastalo leta 2010, je bilo na začetku namenjeno predvsem deljenju fotografij s pripisi, ki jih je avtor lahko preoblikoval z različnimi filtri. Ti so fotografijam bodisi dali bolj profesionalen videz bodisi so jih navidezno postarali, da so bile videti, kot do so jih posneli s polaroidnim fotoaparatom ali drugo analogno fotografsko pripravo. Šele od leta 2013, torej približno od takrat, ko se je slovenski predsednik pridružil Instagramu, dopušča omrežje objavljanje krajših posnetkov. Kljub temu na omrežju še vedno prevladujejo fotografije, ob katerih so kratki pripisi, ki so pogosto dopolnjeni s *hashtagi*, torej ključnimi besedami ali *ključniki*, kot lahko prevedemo izraz v slovenščino. Z njimi uporabniki opišejo oziroma pojasnijo, s katerimi temami je povezana podoba.

Način predstavljanja sebe v javnosti, pri katerem ima podoba prednost pred tekstom, izpostavljen pa je vizualni del vsebine, slovenskemu predsedniku očitno ustreza, prednosti medija pa zna uspešno uporabiti pri utrjevanju lastne politične pozicije in transformaciji socialnega v politični kapital, če parafrazirava Pierra Bourdieuja (1986), ki je razdelal različne oblike kapitala. Vsaka predsednikova objava, s katero si po Instagramu kopiči socialni kapital in pridobiva nove sledilce, je tudi javni nastop, pripomoček za samouprizarjanje in predstava za javnost, pred katero si nadene »masko«, preden stopi na »oder«, če uporabiva izrazje, ki ga je Erving Goffman razdelal v prelomnem delu *Predstavljanje sebe v vsakdanjem življenju* (2014).

3. Raziskovalna metodologija

Predsednikov profil na Instagramu sva analizirala s kombinacijo kvalitativnih in kvantitativnih pristopov, pri čemer sva bila pozorna tako na vizualne kot tekstualne vsebine. Pregledala sva celoten profil in evidentirala vse objave od septembra 2013, posebej pa sva se posvetila objavam iz leta 2017, ko je potekala kampanja pred predsedniškimi volitvami.

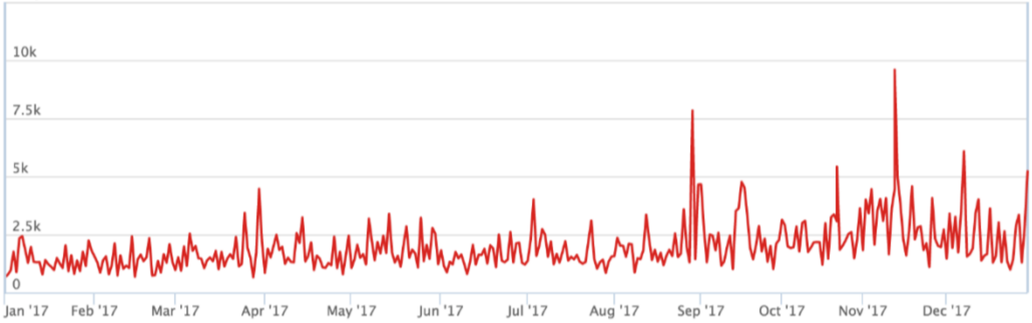
Instagram raziskovalcem ne omogoča, da bi v gručah zajemali podatke z uporabniških profilov preko vmesnikov za programiranje aplikacij (API). Zato sva uporabila orodje 4kStogram, ki slike prenese v lokalno mapo, hkrati pa omogoča razmeroma hiter izvoz pripisov k sliki. Zbrala sva 403 primere slik in opisov, ki sva jih opremila še s številom všečkov in komentarjev za vsako sliko. Za analizo sva ohranila le slike, ki so bile objavljene v letu 2017, in izločila video posnetke, ki pomenijo drugačno obliko komuniciranja po Instagramu in jih ne moremo zlahka primerjati ali enačiti s fotografijami.

Na koncu sva za računalniško analizo ohranila 357 primerov objav, ki sva jih analizirala v orodju Orange (Demšar et al., 2013). Splošen pregled nad aktivnostjo profila sva izvedla z analizo časovnih vrst, pripisov k slikam pa sva se lotila z rudarjenjem besedil.

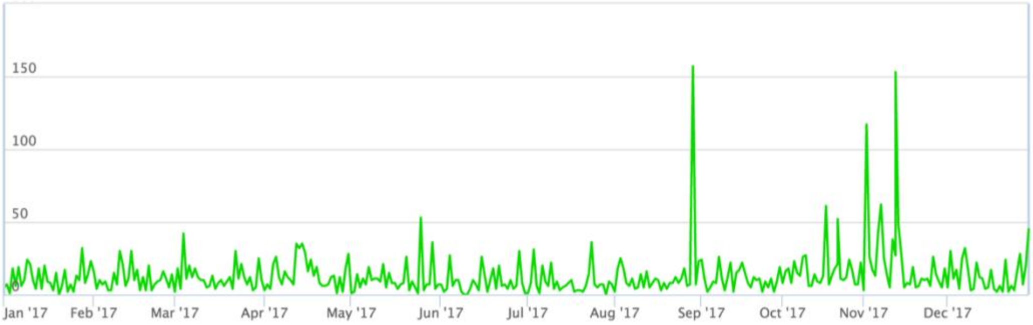
4. Analiza profila na omrežju

Ob začetnem pregledu predsednikovega profila na Instagramu od septembra 2013 do marca 2018, ki sva ga izvedla še brez programskih orodij, sva opazila, da so bile z vsebinskega vidika objave na začetku razmeroma nezanimive in bolj suhoparne, povezane predvsem z državnimi proslavami ter protokolarnimi obiski različnih vojaških in civilnih prireditev. Prva Pahorjeva objava na Instagramu iz leta 2013 je, denimo, posnetek s športne prireditve, na kateri predsednik navdušeno vije slovensko zastavo in spodbuja slovensko reprezentanco. Nacionalni simboli – od zastave in grba do Triglava na dresu slovenske reprezentance – se pojavljajo tudi na številnih njegovih kasnejših objavah, kar ne preseneča glede na njegovo politično funkcijo.

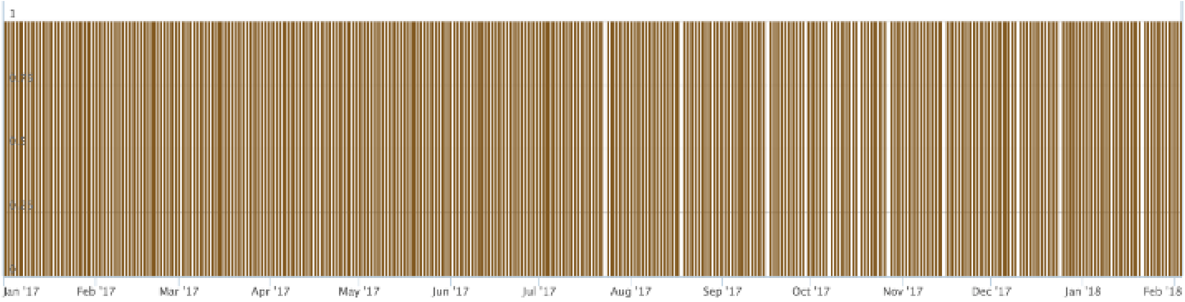
Postopen obrat k bolj osebnim prizorom, ki so pritegnili večjo pozornost javnosti in medijev, opazimo leta 2016, predvsem pa 2017, namreč sočasno z vnovično kandidaturo za predsednika Republike Slovenije. V tem obdobju so postajali bolj osebni, premišljeni in izstopajoči



Slika 1: Število všečkov na profilu v letu 2017.



Slika 2: Število komentarjev na profilu v letu 2017.

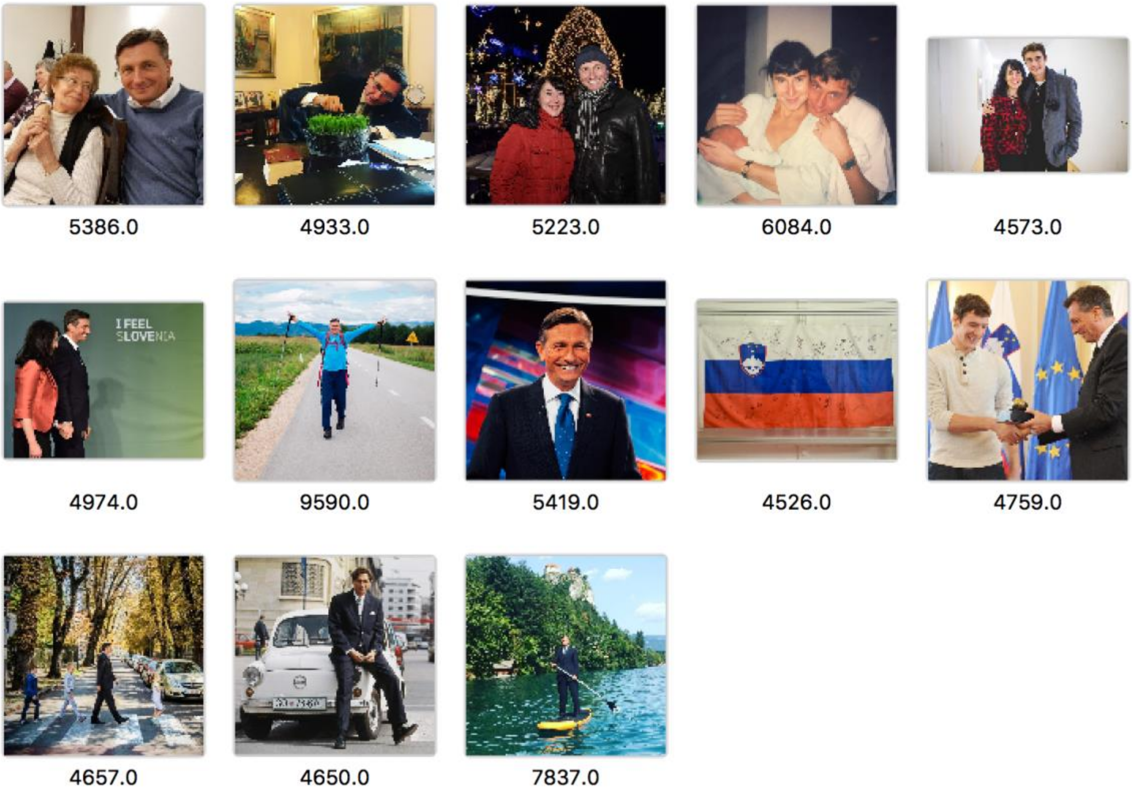


Slika 3: Pogostost objavljanja na profilu (aktivnost) med januarjem 2017 in februarjem 2018.

tudi pripisi k podobam na profilu, kar je predvidoma vplivalo na vse večjo priljubljenost in komentiranost objav – oboje doseže vrh novembra 2017, torej v času drugega kroga volitev predsednika (Slika 1, Slika 2). Iz tretjeosebnih navedb, da je predsednik obiskal vojaško prireditev ali se udeležil lokalnega slavja, so opisi postali bolj prvoosebni in čustveno zaznamovani. To seveda ne pomeni nujno, da te objave od nekega trenutka na Instagramovi časovnici piše in objavlja predsednik sam; prej kaže na to, da je njegovo spletno predstavljanje sebe v vsakdanjem življenju, pri katerem mu pomaga podporna ekipa, postalo bolj dovršeno in prefinjeno. Objave so enakomerno tempirane in se na Instagramu pojavijo ob enakem času dneva, in to skoraj brez izjeme po ena na dan (Slika 3). Pahorjeva stran na Instagramu je v tem obdobju vse bolj posvečena enemu samemu subjektu – predsedniku samemu. Vse drugo, kar se znajde na fotografijah, je upodobljeno predvsem kot ozadje in podpora njegovemu liku in delu, v fokusu pa je skoraj dosledno predsednik.

4.1. Upodabljanje z zvezdniki in družino

Zanimivo je spremljati, katere Pahorjeve objave požanjejo največ zanimanja med sledilci. Med objavami po priljubljenosti izstopajo *grupiji*, in sicer predvsem takšni, na katerih se predsednik pojavlja z drugimi slavnimi osebnostmi, na primer s hrvaško pevko Severino, pevcem skupine U2 Bonom, smučarskim trenerjem Andreo Massijem, prvakom v smučarskih skokih Petrom Prevcem, vodilnim slovenskim motokrosistom Timom Gajserjem, filmskim zvezdnikom Johnom Malkovichem, manekenko Naomi Campbell itn. Pri pojavljanju z zvezdniki – ali pa pri objavljanju z njimi povezanih predmetov in suvenirjev – je pomembno tempiranje posnetka. Fotografija nacionalne zastave s podpisi košarkarjev je, na primer, doživela izjemen odziv na predsednikovem profilu v času, ko je slovenska reprezentanca zmagala na evropskem prvenstvu, podobno pa se Pahor s pevko Heleno Blagne na fotografiji

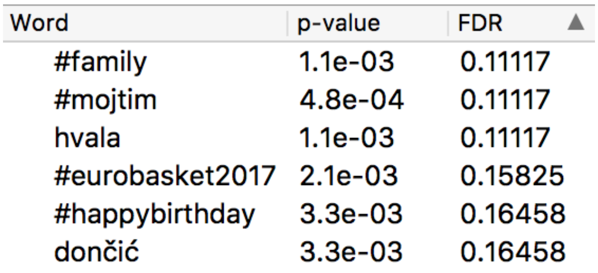


Slika 4: Najbolj priljubljene objave na profilu Boruta Pahorja z navedbo števila všečkov.

pokaže v času, ko se zvezdnica po nekajletnem zatišju vrne na odre.

Še bolj kot fotografije z zvezdniki so med sledilci priljubljene podobe iz predsednikovega zasebnega življenja, ki ga prikažejo kot družinskega človeka, ki namenja pozornost svoji dolgoletni partnerki, ostareli materi in najstniškemu sinu (Slika 4). Več kot štiri tisoč srčkov, s katerimi uporabniki Instagrama označijo, da jim je neka objava všeč, si je prislužila fotografija mladega Pahorja, ki prižema k sebi svojo partnerko. Če sodimo po zapisu ob objavi iz marca 2017, njuna sreča še vedno traja. Ob fotografiji namreč piše: »Že 30 let neperfekten par.«

Analiza pripisov ob najbolj priljubljenih Pahorjevih objavah med januarjem 2017 in februarjem 2018 je pokazala še nekaj zanimivega. Če ima objava na profilu več kot tri tisoč všečkov, zelo verjetno vsebuje eno od naslednjih ključnih besed: dončić, #happybirthday, #eurobasket2017, hvala, #mojtim in #family. Pri tem je prav slednja beseda, torej družina (po angleško *family*) najpomembnejša in vplivnejša tudi od športa in nacionalne košarkarske ekipe. To sva ugotovila z analizo obogatitve besed, kjer sva pripise tokenizirala s tokenizatorjem tvitov iz modula nltk, ki obdrži tudi ključnike in emotikone, nato pa ohranila le tiste pojavnice, ki so prisotne v manj kot 90 odstotkih pripisov. Statistično signifikantne besede sva nato določila z obogatitvijo besed, ki določi tiste pojavnice, ki se v izbranem vzorcu (v najinem primeru v najbolj všečkanih objavah) pojavijo izrazito bolj pogosto kot v celotnem korpusu (Tabela 1). Kadar Pahor izpostavi družino kot univerzalno vrednoto, in to dodatno označi s ključnikom #family, je objava zelo verjetno med najbolj priljubljenimi.



Word	p-value	FDR ▲
#family	1.1e-03	0.11117
#mojtim	4.8e-04	0.11117
hvala	1.1e-03	0.11117
#eurobasket2017	2.1e-03	0.15825
#happybirthday	3.3e-03	0.16458
dončić	3.3e-03	0.16458

Tabela 1: Statistično značilni ključniki ob objavah, ki imajo več kot 3000 všečkov.

4.2. Ustvarjanje ljudskosti

Pahor se na Instagramu pogosto pokaže kot nepopoln in človeški predsednik, kot del ljudstva. Ob fotografiji, na kateri je videti nekoliko zdelan in utrujen, tako pripiše: »Zguban, a ne zguba.« Ko sedi zgrbljen za delovno mizo in zamišljeno zre v papir, piše zraven: »Sklonjen, a neuklonljiv.« Svojo ljudskost in pripravljenost pokazati se javnosti v neurejeni in nedodelani podobi prikaže fotografija iz januarja 2017, ki kaže njegov nasmejan in neobrit obraz; na glavi ima slamnik, na nosu pa sončna očala. Ob fotografiji piše: »Zanemarjen, a ne zapuščen.« Izmed 357 objav jih kar 25 sledi podobni formi »a ne« v pripisih, od teh pa jih je 5 podanih na predlog zunanjih opazovalcev oziroma komentatorjev. Ta jezikovna forma je očitno postala prepoznavna fraza in neformalna blagovna znamka predsednikovega profila na omrežju, ki s svojo preprostostjo in zabavnostjo še podkrepi predsednikovo ljudskost, dostopnost in pripravljenost na šegavo komuniciranje z javnostjo.

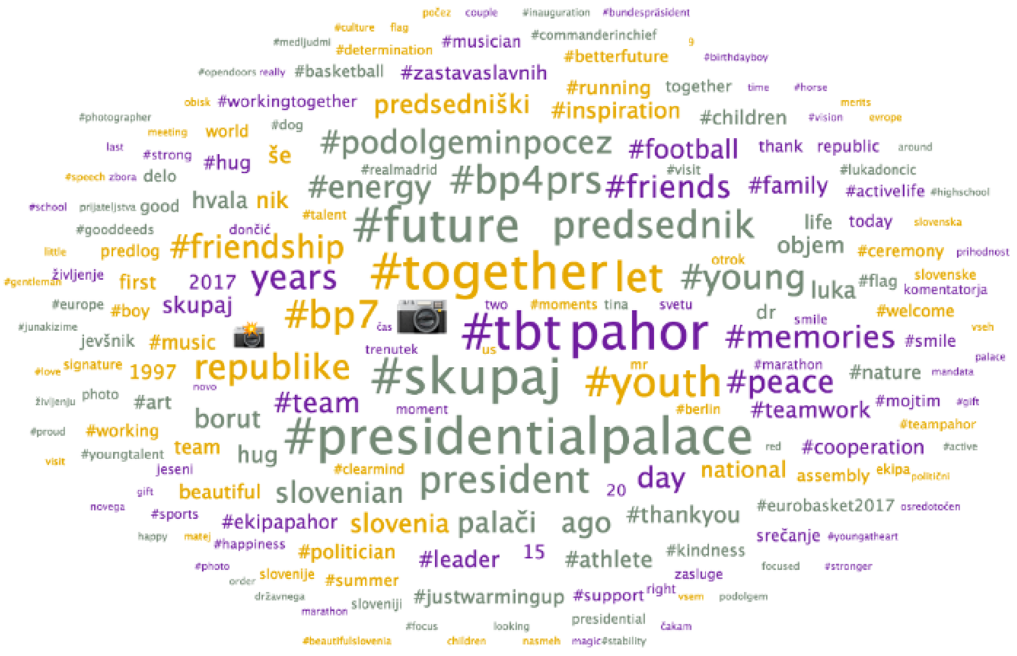
Kot topel in dobrosrčen človek se predsednik pokaže tudi, ker ima očitno rad pse. Kadar se fotografira z njimi, ravno tako pridobi na tisoče všečkov, podobno priljubljene pa so njegove nostalgične objave s starimi avtomobili, ki jih je vozil ali pa se v njih še vedno prevaža, kakršna sta *fičko* in *katrca*, ki sta se zapisala v spomin predvsem ljudem, rojenim v nekdanji Jugoslaviji. S fotografiranjem ob ali v teh avtomobilih znova opozori na svojo skromnost in ljudskost, hkrati pa se pokaže kot odprta oseba, ki je ne zanima le politika.

4.3. Nenavadne in izstopajoče objave

Posebna kategorija Pahorjevih objav so njegove upodobitve na fotografijah v nenavadnih in nepričakovanih položajih in situacijah. Eden takšnih ima pripis »Domotožje v Kairu«, prikazuje pa predsednika, ki zasanjano sloni na ograji razkošnega stopnišča v egiptovski palači in gleda naravnost v objektiv kamere. Ta njegova objava iz decembra 2016 je doletela na izjemen odziv v medijih in postala celo internetna senzacija. Ljudje so po tej objavi tekmovali, kdo se bo bolj izkazal v dejavnosti, ki so jo poimenovali *boruting*, namreč v naslanjanju na ograje, pri čemer so se fotografirali in svoje šaljive podobe razširili po spletu. Skoraj leto kasneje, namreč septembra 2017 se je iz lastne objave ponorčeval tudi Pahor in objavil še eno fotografijo, očitno starejšo, na kateri se naslanja na leseno stopniščno ograjo, ob njej pa je pripis »Pra-domotožje, 2008.« Še ena tovrstna fotografija, objavljena januarja 2018, je postala izjemno priljubljena. Na njej je predsednik v svoji pisarni, na nosu ima očala, v rokah pa škarje, s katerimi vestno in natančno striže travico, posajeno v kristalni skledi. »Po vrnitvi iz tujine sem se takoj lotil domačih nalog, tudi urejanja pisarniškega vrta,« zapiše v pojasnilo. Zakaj je ta fotografija postala uspešnica, ni povsem jasno.

Največ všečkov med vsemi Pahorjevimi objavami sta zbrali dve, na katerih je predsednik prikazan kot junak, ki zmore nemogoče in si drzne več kot drugi. Prva, s katero je Borut Pahor avgusta 2017 na svojem profilu nabral skoraj osem tisoč všečkov, ga prikazuje v poslovni obleki, ko z veslom v rokah stoji sredi blejskega jezera na deski za *supanje*, torej stoječe veslanje. Ob tej podobi, za katero so mnogi mislili, da je fotomontaža, je kratek pripis: »Zanesljivost.« Zgovorni so tudi ključniki ob tej podobi, ki izpostavljajo stabilnost (#stability), zanesljivost (#dependability), predsednikov značaj (#character) in vodstvene sposobnosti (#leader) ter poudarijo njegovo zmožnost povezovanja Slovenije in Slovencev (#together, #skupaj, #presidentpahor, #president, #slovenia).

Še bolj priljubljen je prizor po objavi rezultatov drugega kroga predsedniških volitev leta 2017. Na fotografiji, ki je zbrala več kot devet tisoč všečkov, je Pahor v športni opremi, z nahrbtnikom na ramah in s pohodniškimi palicami v rokah, ki jih široko nasmejan širi v zrak. Z iztegnjenim kazalcem in sredincem kaže znak V, ki označuje dvojno zmago: najprej za prehojeno dolgo predvolilno pešačenje po Sloveniji, med katerim je bivši in bodoči predsednik obiskal številne kraje in ljudi, poleg tega pa seveda še za zmago na volitvah, ki je bila cilj njegove kampanje tako v fizičnem kot digitalnem prostoru.



Slika 5: Najbolj pogoste besede iz pripisov k slikam. Na prikazu so izpostavljene zgolj besede, ki se pojavijo v manj kot 90 odstotkih objav, torej takšne, ki niso prisotne ob vsaki objavi.

Pahorjev profil zajema vrsto tematik in je po vsebini precej heterogen. To kaže tudi preprost oblak besed (Slika 5), ki kot pogoste besede izpostavlja, na primer, #tbt,[1] ki označuje nostalgične slike, #together oziroma #skupaj, kar je bil slogan Pahorjeve predvolilne kampanje, ter #future, ki kaže na predsednikovo zavezanost skupni prihodnosti.

5. Diskusija

Kot sva že omenila, Pahorjeva izbira Instagrama kot ključnega orodja za prikazovanje vrlin in dosežkov ni naključna. Instagramova glavna značilnost je, da poudarja podobo in izpostavlja posameznika z njegovimi vizualnimi specifikami, s čimer lahko uspešno pomaga pri gradnji politične kariere človeka, ki je vajen stati pred objektivom. Predsedniku Borutu Pahorju je to verjetno nekoliko lažje kot drugim politikom tudi zaradi njegovih manekenskih izkušenj iz študentskih let (Pirc, 2007). Še bolj pomembno pa je, da Instagram nudi izvrstno platformo za ustvarjanje *psevdo-dogodkov*, kot jih je poimenoval zgodovinar Daniel J. Boorstin, torej navideznih, insceniranih, umetno uprizorjenih dogodkov. V delu *The Image: A Guide to Pseudo-Events in America*, je pojasnil, da psevdo-dogodke ustvarjajo predvsem zvezdniki, in sicer preprosto zato, ker so »po svoji naravi bolj zanimivi in privlačni kot spontani dogodki« (Boorstin, 1992: 37). Pahorjevo stanje na deski sredi jezera in opiranje na ograjo nedvomno lahko uvrstimo med tovrstne dogodke, ki skoraj zagotovo niso spontani, temveč inscenirani, a so kljub temu – ali pa ravno zato – izjemno odmevni. Podobna je tudi objava iz marca 2016, ob kateri piše: »Nocoj v ringu dva borca.« Na njej vidimo predsednika v vodi, ki v slogu delfina plava proti robu bazena, kjer so odložene boksarske rokavice. Težko si predstavljamo, da tak prizor ni insceniran in da je namenjen čemur koli drugemu kot (samo)promociji.

Odzivi sledilcev in nasploh uporabnikov Instagrama na tovrstne samopromocijske objave so pogosto mešani. Ob omenjeni fotografiji v bazenu je najprej nadvse pozitivno sporočilo: »Uuu kok kjut.« Odziv očitno pomeni, da je sledilki prizor všeč. Drugo sporočilo je kratko in ostro: »Narcisoid.« Tretji komentator stopi predsedniku v bran in pravi: »Pa ti si celi narcisoid. Človek, ki lahko pokaže svoje dosežke, je kvečjemu samo uspešen človek, če se zna z njimi še pohvaliti, pa pomeni, da jih ceni, ker konstanto napreduje. Človek, ki mu to oporeka, tako kot ti, je pa zavisten in reven v duši.«[2] Takšna polariziranost komentarjev ob Pahorjevih objavah je pogosta, veliko pa se jih obregne prav ob razkazovanje po spletu, ki je po mnenju nekaterih v sodobnem političnem prostoru nujno, drugim pa se zdi neresno in nepredsedniško.

6. Sklep

Pahorjeva strategija samouprizarjanja in gradnje lastne blagovne znamke se ne zdi nič posebnega in novega v svetu, ki je obseden z medijskimi podobami, prežet z narcizmom in impregniran s hipervizibilnostjo. Takšne strategije dejansko še precej bolje kot politiki obvladajo instant zvezdniki, med katere lahko v tujini štejemo Kim Kardashian in Paris Hilton, pri nas pa na primer Damjana Murka, Denise Dame in Urško Hočevar Čepin (Podjed, 2012; Podjed, 2013). Poleg tega Borut Pahor ni prvi ali edini slovenski politik, ki se je izmojstril v uporabi družbenih medijev in kreiranju ter širjenju psevdo-

[1] #tbt je oznaka za t.i. *Throwback Thursday*, ki na Instagramu pomeni objavo podob iz preteklosti, ob katerih obujamo nostalgične spomine.

[2] Pri tretjem sporočilu so za lažje branje dodane kljukice na šumnikih, ki jih v izvirniku ni.

dogodkov po spletu. Pri njem pa je bistveno drugače to, da tudi sam ustvarja trende na področju uporabe omrežij in načinov predstavljanja sebe po spletu, namesto da bi jim zgolj sledil in jih kopiral od domačih in tujih zvezdnikov. Z lastnim pristopom, temelječim na izpostavljanju vizualne podobe po omrežju Instagram, ki je posebej priljubljeno med mladimi, je ustvaril prepoznavno osebnost, ki obstaja tako v fizičnem kot digitalnem prostoru in je sočasno heterogena in celovita, saj presega suhoparnost politike in jo nadgrajuje z drugimi dejavnostmi: od športa in razvedrila do družine.

V nadaljnjih raziskavah bi bilo potrebno to analizo nujno nadgraditi in slovenskega predsednika primerjati z drugimi politiki in izstopajočimi posamezniki – glasbeniki, športniki, igralci itd. Tako bi namreč dobili bolj celovit pogled na samopromocijske strategije, utemeljene na Instagramu. Zanimiva bi bila še primerjava med Slovenijo in tujino, s čimer bi lahko ugotovili, kako družbeno in kulturno okolje vpliva na posameznikovo predstavljanje po družbenem omrežju in kako njegovo udejstvovanje na omrežju sooblikuje lokalno in globalno družbeno realnost.

7. Literatura

Associated Press. 2017. 'Barbie, and Not a Bad Guy': Meet Borut Pahor, Slovenia's Instagram President. *The Guardian,* 8. 3. 2017. https://www.theguardian.com/world/2017/mar/08/borut-pahor-slovenia-instagram-president.

Daniel J. Boorstin. 1992 (1961). *The Image: A Guide to Pseudo-Events in America*. Vintage Books, New York.

Pierre Bourdieu. 1986. The Forms of Capital. V: John G. Richardson, ur., *Handbook of Theory and Research for the Sociology of Education*, str. 241–258. Greenwood, New York.

Julia Deeb-Swihart, Christopher Polack, Eric Gilbert in Irfan Essa. 2017. Selfie-Presentation in Everyday Life: A Large-Scale Characterization of Selfie Contexts on Instagram. V: *Proceedings of the Eleventh International AAAI Conference on Web and Social Media ICWSM 2017*, str. 42–51. AAAI, Montréal. https://www.aaai.org/Library/ICWSM/icwsm17contents.php.

Janez Demšar, Tomaž Curk et al. 2013. Orange: Data Mining Toolbox in Python. *The Journal of Machine Learning Research*, 14(1): 2349–2353.

Sarah Diefenbach in Lara Christoforakos. 2017. The Selfie Paradox: Nobody Seems to Like Them Yet Everyone Has Reasons to Take Them. An Exploration of Psychological Functions of Selfies in Self-Presentation. *Frontiers in Psychology* 8(7): 1–14.

Erving Goffman. 2014 (1959). *Predstavljanje sebe v vsakdanjem življenju.* Studia humanitatis, Ljubljana.

Edgar Gómez Cruz in Helen Thornham. 2015. Selfies Beyond Self-Representation: The (Theoretical) F(r)ictions of a Practice. *Journal of Aesthetics and Culture* 7: 1–10.

Vasja Jager. 2017. Kralj Instagrama. *Mladina,* 18. 8. 2017. http://www.mladina.si/181458/kralj-instagrama/

Marshall McLuhan. 1964. *Understanding Media: The Extensions of Man*. McGraw-Hill, London, New York in Toronto.

Vanja Pirc. 2007. Po tretji poti od manekena do liderja. *Mladina,* 5. 7. 2007. http://www.mladina.si/95402/po-tretji-poti-od-manekena-do-liderja/.

Dan Podjed. 2012. Slovenske instant zvezde: Ustvarjanje in ohranjanje slave po svetovnem spletu. *Glasnik SED* 52(1/2): 72–81.

Dan Podjed. 2013. Konec zgodovine Vélikih Mož? Samopromocijske strategije instant zvezde Urške Hočevar Čepin in njihov odsev v politiki. V: Božidar Jezernik, ur., *Heroji in slavne osebnosti na Slovenskem*, str. 151–172. Znanstvena založba Filozofske fakultete, Ljubljana.

Theresa M. Senft in Nancy K. Baym. 2015. What Does the Selfie Say? Investigating a Global Phenomenon. *International Journal of Communication* 9: 1588–1606.

Data Mining Workspace Sensors: A New Approach to Anthropology

Ajda Pretnar,[*] Dan Podjed[†]

[*]Laboratory of Bioinformatics
Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, SI-1000 Ljubljana
ajda.pretnar@fri.uni-lj.si

[†]Institute of Slovenian Ethnology
Research Centre of the Slovenian Academy of Sciences and Arts
Novi trg 2, SI-1000 Ljubljana
dan.podjed@zrc-sazu.si

Abstract

While social sciences and humanities are rapidly including computational methods in their research, anthropology seems to be lagging behind. However, this does not have to be the case. Anthropology is able to merge quantitative and qualitative methods successfully, especially when traversing between the two. In the following contribution, we propose a new methodological approach and describe how to engage quantitative methods and data analysis to support ethnographic research. We showcase this methodology with the analysis of sensor data from a University of Ljubljana's faculty building, where we observed human practices and behaviours of employees during working hours and analysed how they interact with the environment. We applied the proposed circular mixed methods approach that combines data analysis (quantitative approach) with ethnography (qualitative approach) on an example of a smart building and empirically identified the main benefits of our methodology.

1. Introduction

Social sciences and humanities are rapidly adopting computational approaches and software tools, resulting in an emerging field of digital humanities (Klein and Gold, 2016). Among these is anthropology, which is particularly suitable for traversing between quantitative and qualitative methods. Anthropologists study and analyse human behaviours and cultures, with a particular focus on long-term fieldwork as a methodological cornerstone of the discipline. With an increasing availability of data coming from social networks and wearable devices among other sources (Miller et al., 2016; Gershenfeld and Vasseur, 2014), anthropologists can easier than ever dive into data analysis and study humans and their societies, subcultures and cultures quantitatively as well as qualitatively.

With this contribution, we tentatively place anthropology in the field of digital humanities, mostly because the suggested approach is multidisciplinary and by analogy similar to the shifts between distant and close reading (Jänicke et al., 2015) in literary studies. Just like distant reading can offer an abstract (over)view of the corpus, quantitative analyses can give a researcher a broad understanding of the population she is investigating. And just like distant reading needs close reading to understand the style, themes, and subtle meanings of a literary work, so does data analysis need an ethnographic approach to contextualize the information and extract subtle meanings of individual human experience.

As Pink et al. (2017) suggest, there is value in investigating everyday data that reveal what is ordinary, what extraordinary and how to contextualize the two. In this contribution we expand the idea by employing the so-called *circular mixed methods approach* that combines qualitative research from anthropology and quantitative analysis from data mining. We consider mixed methods (Creswell and Clark, 2007; Teddlie and Tashakkori, 2009) as an integrative research that merges data collection, methods of research and philosophical issues from both quantitative and qualitative research paradigms into a singular framework (Johnson et al., 2007). We also stress the need for a circular research design, where we intentionally traverse between methods to continually verify and enhance our knowledge of the field. Circularity gives research flexibility and enables shifting perspectives in response to new information.

Our research began in October 2017 and currently monitors 14 offices at one of the University of Ljubljana's faculty buildings. We retrieved measurements from approximately 20 sensors from the SCADA monitoring system for the year 2016 and extrapolated behavioural patterns for different rooms and, more generally, room types through data visualization and exploratory analysis. The analysis showed specific patterns emerging in several rooms - there were some definite outliers in terms of working hours and room interaction.

We used computational methods to gauge new perspectives on human behaviour and invoke potentially interesting hypotheses. Data analysis provided several distinct patterns of behaviour and defined the baseline for workspace use. However, this approach was unable to provide us with a context for the data. Quantitative methods can easily answer the 'what', 'where' and 'when' type of questions, but struggle with the 'why'. At that stage, we employed long-term fieldwork and ethnography as the main methods of anthropology. We conducted interviews with room occupants to explain what the uncovered patterns mean and why people behave the way they do.

The main goal of our study was to demonstrate how anthropologists can use statistics and data visualisation to establish the essential facts of the observed phenomena and how the traditional anthropological methods, which have not significantly changed since the early 20th century (Malinowski, 1922), can be complemented and upgraded by data analysis. We call this a circular mixed methods approach, where circular implies continual traversing between qualitative and quantitative methods, between fieldwork and data analysis. The present contribution applies the proposed methodology to sensor data obtained from a smart building and with a combination of data mining and ethnographic fieldwork establishes both a wide and deep understanding of human behaviour in a workplace setting.

2. Related Work

While digital humanities became a full-fledged field in the last couple of decades (Hockey, 2004), anthropology seems to be left of out its spectrum. Some authors suggest anthropology would be more concerned with digital as an object of analysis rather than as a tool (Svensson, 2010). However, there have been several attempts to include computational methods and quantitative analyses into anthropological research. Already in the 1960s, anthropologists looked at using computers for organisation of anthropological data and field notes (Kuzara et al., 1966; Podolefsky and McCarty, 1983). Progress in text analysis, coding facts, and comparative studies in linguistics (Dobbert et al., 1984; White and Truex, 1988) followed suit.

However, only lately there has been some digital shift in the discipline. Digital anthropology turned disciplinary attention to the analysis of online worlds, virtual identities, and human relationships with technology. For example, Bell (2006) gave a cultural interpretation of the use of ICTs in South and Southeast Asia, Nardi (2010) explored gaming behaviour of the World of Warcraft, Boellstorff (2015) investigated alternate online worlds of the Second Life, and Bonilla and Rosa (2015) described how to use hashtags for ethnographic research. Moreover, a discussion has been opened on what does 'big data' mean for social sciences and how to ethically address its retrieval and analysis (Boyd and Crawford, 2012; Mittelstadt et al., 2016).

There was a discussion on the methodological front as well. Anderson et al. (2009) argue for a method that combines the ethos of ethnography with database mining techniques, something the authors call 'ethno-mining'. Similarly, Blok and Pedersen (2014) look at the intersection of 'big' and 'small' data to produce 'thick' data and include research subjects as co-producers of knowledge about themselves. Finally, Krieg et al. (2017) not only elaborate on the usefulness of algorithms for ethnographic fieldwork, but also show in detail how to conduct such research in an example of online reports of drug experiences.

3. Anthropology vs. Data Analysis

For an anthropologist, statistical and computational analysis is not the first thing that comes to mind when developing research design and methodology. Anthropologists are trained to observe phenomena in the field, talk to people, spend time with them, participate in daily activities, and immerse themselves in topics of interest (Kawulich, 2005; Marcus, 2007). This type of information gives us detailed stories of human lives, uncovers meanings behind rituals, habits, languages, and relationships, and provides a coherent explanation of the researched phenomena. So why would anthropologists even have to include data analysis in their studies? Why and when is such an approach relevant?

Sometimes, the phenomena that anthropologists are trying to explain occur in different places at the same time and are impossible to observe simultaneously. It could be that anthropologists know little of the topic they are exploring and have yet to generate their research questions. Or the nature of the phenomenon lends itself nicely to computational analysis. For example, behaviour of many individuals is difficult to observe in real time, especially if we want to observe them at once in different locations. Sensors, on the other hand, can track behaviours of these individuals independently (Patel et al., 2012) and therefore enable a detailed comparative analysis. With a large amount of measurements, researchers can also observe seasonal variations, similarity of users, and changes through time.

Data analysis also helps us define the parameters of our research field and establish what is an ordinary and what is an extraordinary behaviour. Visualisations in particular are excellent tools for exploring and understanding frequent patterns of behaviour and outliers. When done well, visualisations harness the perceptual abilities of humans to provide visual insights into data (Fayyad et al., 2002, p. 4). Moreover, they provide a new perspective on a phenomenon and help generate research questions and hypotheses. Once we know how our research participants behave (or communicate if we are observing textual documents or establish social ties if we are observing social networks), we can enter the field equipped with knowledge and information to verify and contextualise.

Finally, large data sets are particularly appropriate for computational analysis. While 'big data‘ became a popular buzzword in data science, anthropologists most likely will not be dealing with millions of data points that can be analysed only with graphics processing units (GPUs). However, even ten thousand observations are too much for a researcher to make sense of. For such data, we need software tools and visualisations, which provide an overview of the phenomenon, plot typical patterns, and enable exploring different sub-populations.

4. Data Ethics and Surveillance Technologies

Data ethnography inevitably raises questions of ethics, just like sensor data inevitably raise the question of surveillance. Both topics are too broad for the scope of this contribution but let us briefly touch upon them. Research ethics, in particular sensitivity to the potential harm a study could elicit, is one of the core questions of anthropology, which is so deeply immersed in the personal human experience. A solid deontological paradigm is crucial for working with not only sensitive data but any human-produced data. In this sense, we follow the principles of positivist ethics which call for human dignity, autonomy, protection, maximizing benefits and minimizing harm, respect, and justice (Markham et al., 2012; Halford, 2017).

As for surveillance, we propose a distinction between surveillance and monitoring. Surveillance implies guiding actions of surveilled subjects, while monitoring proposes a more passive stance of observing behaviour. The present study was not designed to guide behaviour but to observe and understand, hence being more monitoring than surveillance focused. And even if we consider it surveillance-like, Marx (2002) proposes ”a broad comparative measure of surveillance slack which considers the extent to which a technology is applied, rather than the absolute amount of surveillance”, meaning that the extent to which surveillance is harmful is the power it holds for the user. The case of sensor data of a smart building that monitors only neutral human behaviour [1], falls to the soft side of power, which, in the opinion of the authors, deserves some surveillance slack. Nevertheless, we strived to uphold high ethical standards for handling the data and disseminating the results, mostly by employing ”ongoing consensual decision-making” (Ramos, 1989) by informing our participants of the purpose of the research, which data are being collected and how the findings are going to be presented.

5. Data Preprocessing

In our study, we have observed sensor measurements from a faculty which is considered to be a state-of-the-art smart building in Slovenia. Each room in the building is equipped with a temperature sensor and sensors on windows that track when they are open or closed. Doors have electronic key locks that track when the room is occupied. There were altogether 11 sensor measurements, with an additional 8 measurements coming from the weather station located on the building’s rooftop. In-room sensor reports the room temperature, set temperature, ventilation speed, daily regime, and so on, while the weather station reports the external temperature, light, rainfall, etc. One of the most important measurements is the daily regime, which has four values, each representing a state of the overall room setting. When a person is present in the room, the regime is comfort (value = 0) and when a window is open, the regime is off (value = 4). If the room is vacant, the regime goes to night (value = 1) or standby (value = 3)[2].

We retrieved 55,456 recordings for 14 rooms of different types, namely 5 laboratories, 6 cabinets, and 3 administration rooms. Measurements are recorded bi-hourly and stored in the SCADA monitoring system. We decided to observe the year 2016 and later compare it to 2017. The results in the paper refer only to 2016. The rooms are anonymised to ensure data privacy and results for two of the rooms are not reported at the request of their occupants.

We performed extensive data cleaning and preprocessing and removed data points with missing values (Table 1). We considered *daily regime* as our most important variable since it reports a presence in the room or the opening of windows. Concurrently, we removed data points where the daily regime was *comfort* throughout the day[3].

For the analysis, we retained only one feature, namely *daily regime*, since, as mentioned above, this was the feature that registered human behaviour the best. We also generated additional features, such as the day of the week and room type (cabinet, laboratory, and administration).

In the second part of the analysis, we created a transformed data set where we merged daily readings for a room into one ‘daily behaviour‘ vector (Table 2). In the new data set, each room has a daily recording, where the new features are values of the daily regime at each hour. Since sensors only record the state every two hours, we filled missing values with the previous observed state. For example, if the original vector was $\{0, ?, 0, ?, 1, ?, 1\}$, we imputed missing values to get $\{0, 0, 0, 0, 1, 1, 1\}$. As we were interested only in the presence in the room, we put 0 where daily regime was 1 (night) or 3 (standby) and 1 where it was 0 (comfort) or 4 (window open), discarding the information on specific temperature regimes. This gave us the final daily behaviour vector which we could compare in time and between rooms.

6. Results

First, we wanted to see how rooms differ by room occupancy alone. We hypothesised there will be a significant difference in occupancy between laboratories and cabinets since the presence of more people in a space extends the occupancy hours (no complete overlap of working time).

We took the first data set with bi-hourly recordings and removed readings where the daily regime was either 1 (night) or 3 (standby) because these readings indicate the room was not occupied. Afterwards, we computed the contingency matrix of room occupancy by the day of the week, which shows how many times per year a room was occupied on a certain day. We visualised the result in a line plot (Figure 1). We can notice that laboratories have a higher presence on Saturday and Sunday than the other rooms.

Moreover, N and O are the top two rooms by occupancy. We know that these two rooms belong to a single laboratory and are separated with a permanently open door. These two rooms are occupied by the largest number of people and since the employees of the faculty have a somewhat flexible working time, the dispersion of working time is expectedly the highest in rooms with the most occupants (smallest overlap in working time among employees). N and O are also among the few rooms where occupancy goes up towards the end of the week.

F and B are also laboratories, both displaying similarly high presence across the week. On the bottom of the plot there are cabinets, namely G, K, F. Unsurprisingly, cabinets display lower occupancy rates than laboratories, since cabinets are used by a single person and hence no overlap is

[1] We consider neutral human behaviour a behaviour which does not explicitly convey sensitive information.

[2] Standby is activated on workdays as a transitory setting between night and comfort regime.

[3] We considered a constant comfort regime an error in the reading since a 24-hour workday is highly unlikely.

Date	Room temp	Daily regime	Room
2016-01-01 02:10:00	20.94266	1	C
2016-01-01 02:10:00	21.65854	1	B
2016-01-01 02:10:00	20.63234	1	K
2016-01-01 02:10:00	22.41270	1	D
2016-01-01 02:10:00	20.25890	1	M
2016-01-01 07:10:00	21.45220	3	C

Table 1: Original data

Date	0am	1am	2am	3am	...	Room	Day	Type
2016-01-01	0	1	1	0	...	C	Fri	laboratory
2016-01-01	0	0	0	0	...	B	Fri	laboratory
2016-01-01	0	0	1	1	...	K	Fri	administration
2016-01-01	0	0	0	1	...	D	Fri	cabinet
2016-01-01	0	1	1	1	...	M	Fri	administration
2016-01-02	0	0	1	1	...	C	Sat	laboratory

Table 2: Data transformed into a behaviour vector. 1 denotes presence in the room, meaning daily regime value was either 0 (comfort) or 4 (window open).

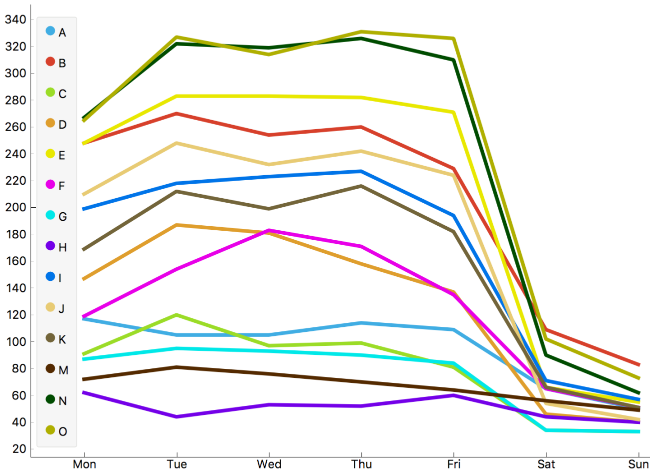


Figure 1: Occupancy of the rooms for each day of the week.

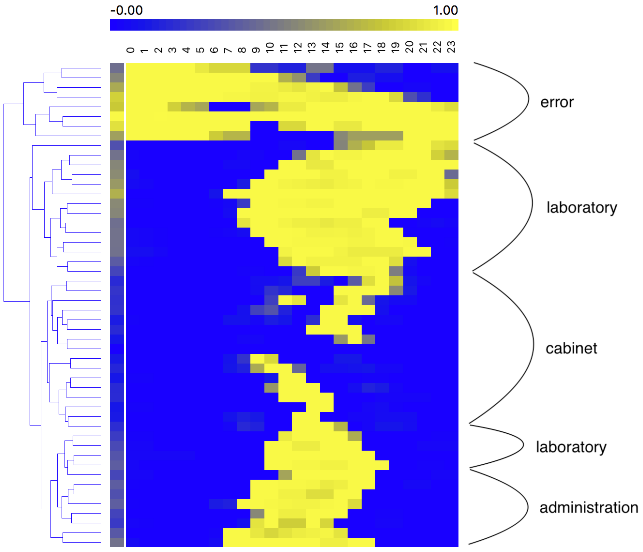


Figure 2: Occupancy by the hour of the day. Distinctive room-type patterns emerge.

possible. They are also functional rooms, used predominantly for meetings, office hours, and other intermittent work of professors.

With the second room occupancy data set, we made an analysis of behavioural patterns by the time of the day. We observed occupancy by room type in a heat map where 1 (yellow) means presence and 0 (blue) absence. Visualisation in Figure 2 is simplified by merging similar rows with k-means (k=50) and clustering by similarity (Euclidean distance, average linkage and optimal leaf ordering). Such simplification joins identical or highly similar patterns into one row and rearranges them so that similar rows are put closer together.

Clustering revealed that occupancy sequence highly depends on the room type. There were some error data, where sensors recorded presence at unusual hours (for example during the night consistently across all rooms). But despite some noise in our data, we can distinguish between typical laboratory, administration and cabinet behaviour, since our error data constitute a separate cluster (Dave, 1991). Cabinets again show the lowest occupancy with presence recorded sporadically across the day. Normally, university lecturers spend a large portion of their time in lecture rooms and in their respective laboratories. This is why occupancy of cabinets is so erratic and does not display a consistent pattern. Laboratory occupants, on the other hand, usually come late and stay late, while administration staff work regularly from 7:00 a.m. to 4:00 p.m. They both display fairly consistent behaviour.

We visualised the same data set in a line plot, which shows the frequency of attributes on a line. This way, we can better observe differences between individual rooms at each time of the day and where specific peaks (high frequencies) happen. Figure 3 displays the occupancy ratio at a specific time of the day, while Figure 4 shows the ratio of window opening [4]. Several interesting observations emerge. In both cases, room O is skewed to the right, meaning its occupants work at late hours and open windows while working. Conversely, room J is skewed to the right, indicating its occupants start work earlier than most. There is also a distinct peak in window opening at around lunch time.

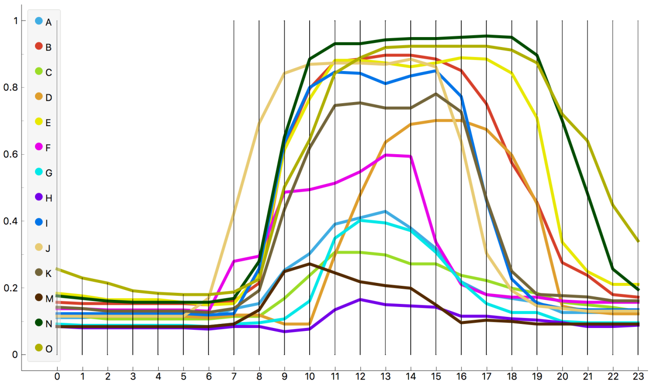


Figure 3: Room occupancy by the time of the day.

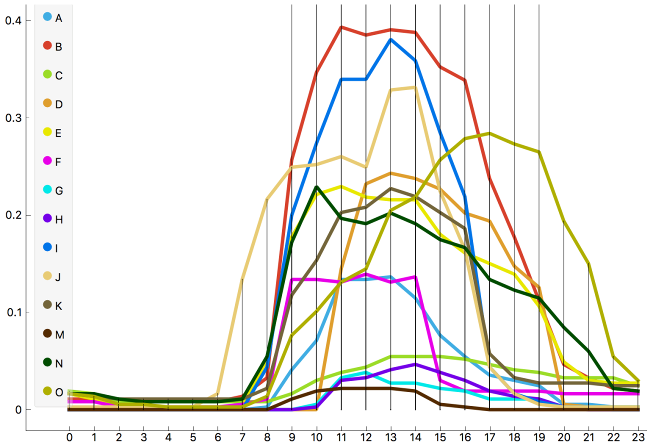


Figure 4: Window opening frequency by the time of the day.

In most rooms, people are opening windows from late morning to early afternoon. Again, not surprising, considering this is their peak working time. This is a great indicator for an ethnographer if he or she wants to observe window interaction (who does it, is there a consensus on whether or not it should be opened, does this happen more frequently after lunch...). Looking at the data, the best time for observing the specified behaviour is between 10:00 a.m. and 1:00 p.m. Accordingly, data analysis can also serve as a guide for ethnographic field work.

7. Ethnography Comes In

Data analysis revealed some interesting patterns in the use of working spaces:

- laboratories work more on weekends,
- rooms N and O work late,
- room J starts the day early and opens windows at lunchtime, and
- in rooms H, N and O the occupancy goes up towards the end of the week

How can we explain this? While the data gave us clues, the answers lie with the people. Substantiating analytical findings with fieldwork ethnography is crucial for understanding the data. We conducted semi-structured interviews with the rooms' occupants to discover what those patterns mean and why a certain behaviour occurs.

Laboratories have a higher weekend occupancy since they offer a quiet place to work for PhD students who are either catching deadlines for publishing papers or using their 'off time' for some in-depth research. Room B, in particular, seems to like working at weekends and we were able to identify an individual who often comes to work on Saturdays. In the interview, he[5] told us this was time when he was able to really focus on his dissertation.

Rooms N and O are quite similar in terms of presence although room N displays a tendency to work the latest. By observing inhabitants in this room and talking to them, we identified an individual who preferred to work in the late afternoon and evening. Since, as mentioned above, working time is flexible at the studied faculty, he adjusted his working hours to suit his preferences. He also prefers fresh air to artificial ventilation and opens the windows whenever possible. This accounts for the skew to the right for room O in Figure 4.

The increased productivity in rooms N, O, and H towards the end of the week is explained by the fact that Fridays are working sprints for occupants of these three rooms. The case of room H is particularly interesting. This is the room with the overall lowest occupancy, yet the room is most frequented on Fridays, unlike in most other rooms, where the occupancy decreases towards the end of the week. Room H is the cabinet of a professor who runs laboratories N and O. He is also a part of the Friday working sprints, hence the peak. Yet he is very social and prefers to work in the laboratory with colleagues, rather than alone in the cabinet. This explains the overall low and erratic occupancy of his room during the rest of the week.

The skewed peak for room J in Figure 3 is again interesting. The occupant of this room admitted he prefers coming to work earlier to make the most of the day. He stressed several times that daylight is important to him and by shifting working time to earlier hours, he was able to leave early and use the rest of the day for himself. He also said he was the most productive in early mornings since these were the quietest parts of the day. Personal preferences evidently affected the discovered patterns of workday behaviour.

Such a circular methodological approach, where the researcher traverses between data analysis, observation, and

[4] 1 would mean the room was always occupied and 0 that the room was never occupied at a specific time of the day.

[5] Pronoun he is used for both males and females.

ethnography, has several benefits. Large data collections can be effectively and rapidly analysed with computational means. Visualisations, moreover, substantiate the findings and enable researchers to uncover relations, patterns, and outliers in the data. Hence, data analysis can help generate hypotheses and questions for the research. This cuts down the time required to get familiar with the field. A researcher can come into the field equipped with potentially interesting hypotheses and test them almost immediately.

Looking at the data alone, however, we would be unable to determine what any of those patterns and outliers mean. To truly understand them, we need to immerse ourselves in the field, ask questions and observe how people behave and create their habits and practices. While quantitative analysis provides us with clues, qualitative approaches, such as ethnography and fieldwork, explain those clues and substantiate the superficial knowledge of the field acquired in the first research phase. Metaphorically speaking, data analysis is great for scratching the surface, while ethnography excels at digging deeper.

8. Conclusion

In this paper, we have shown the how to combine quantitative and qualitative methods for anthropological research. While the findings are still preliminary and based on a limited sample, they nevertheless pinpoint aspects of data analysis that benefit from ethnographic insight and vice versa.

With the increasing availability of data, especially from sensors, wearable devices, and social media, anthropologists can use computational methods and data analysis to uncover common patterns of human behaviour and pinpoint interesting outliers. Quantitative methods have proven useful when dealing with large data sets. In such cases, an analysis without digital tools is virtually impossible, while visualisations offer new insight into the problem and help present the data concisely. In addition, quantitative approaches also increase the reproducibility of research.

However, patterns emerging from such analysis can hardly ever be explained with data alone. We argue that data analysis can generate new hypotheses and research questions (Krieg et al., 2017) and provide a general overview of the topic. Conversely, ethnography substantiates analytical findings with the context and story behind the data. Going back and forth, from quantitative to qualitative and *vice versa*, enables researchers to establish a research problem as suggested by the data, gauge new perspectives on the known problems, and account for outliers and patterns in the data. Circular research design enhances the quality of information, which does not have to derive solely from a quantitative or qualitative approach. By combining the two, we are using a research loop that ensures both sets of data get an additional perspective - quantitative data are verified with ethnography in the field, while ethnographic data become supported with statistically relevant patterns.

Such methods are already, to a certain extent, employed in digital anthropology (Drazin, 2012), but they are gaining more prominence in mainstream anthropology as well (Krieg et al., 2017). By establishing a solid methodological framework for quantitative analyses in relation to qualitative ones, we do not only strengthen the subfield of computational anthropology, but also provide new perspectives and research ventures to anthropology and emphasise its relevance for studying lifestyles, habits, and practices in data-driven societies.

9. References

Ken Anderson, Dawn Nafus, Tye Rattenbury, and Ryan Aipperspach. 2009. Numbers have qualities too: Experiences with ethno-mining. In *Ethnographic Praxis in Industry Conference Proceedings*, volume 2009, pages 123–140. Wiley Online Library.

Genevieve Bell. 2006. Satu keluarga, satu komputer (one home, one computer): Cultural accounts of icts in south and southeast asia. *Design Issues*, 22(2):35–55.

Anders Blok and Morten Axel Pedersen. 2014. Complementary social science? quali-quantitative experiments in a big data world. *Big Data & Society*, 1(2):2053951714543908.

Tom Boellstorff. 2015. *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human*. Princeton University Press.

Yarimar Bonilla and Jonathan Rosa. 2015. # ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the united states. *American Ethnologist*, 42(1):4–17.

Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679.

John W Creswell and Vicki L Plano Clark. 2007. Designing and conducting mixed methods research.

Rajesh N Dave. 1991. Characterization and detection of noise in clustering. *Pattern Recognition Letters*, 12(11):657–664.

Marion Lundy Dobbert, Dennis P McGuire, James J Pearson, and Kenneth Clarkson Taylor. 1984. An application of dimensional analysis in cultural anthropology. *American Anthropologist*, 86(4):854–884.

Adam Drazin. 2012. Design anthropology: Working on, with and for digital technologies. *Digital Anthropology*, pages 245–65.

Usama M Fayyad, Andreas Wierse, and Georges G Grinstein. 2002. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann.

Neil Gershenfeld and JP Vasseur. 2014. As objects go online; the promise (and pitfalls) of the internet of things. *Foreign Aff.*, 93:60.

Susan Halford. 2017. The ethical disruptions of social media data: Tales from the field. In *The Ethics of Online Research*, pages 13–25. Emerald Publishing Limited.

Susan Hockey. 2004. The history of humanities computing. *A Companion to Digital Humanities*, pages 3–19.

Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, and Gerik Scheuermann. 2015. On close and distant reading in digital humanities: A survey and future challenges. In *Eurographics Conference on Visualization (EuroVis)-STARs. The Eurographics Association*.

R Burke Johnson, Anthony J Onwuegbuzie, and Lisa A Turner. 2007. Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, 1(2):112–133.

Barbara Kawulich. 2005. Participant observation as a data collection method. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 6(2).

Lauren F Klein and Matthew Gold. 2016. Digital humanities: The expanded field. *Debates in the Digital Humanities*.

Lisa Jenny Krieg, Moritz Berning, and Anita Hardon. 2017. Anthropology with algorithms? *Medicine Anthropology Theory*, 4(3).

Richard S Kuzara, George R Mead, and Keith A Dixon. 1966. Seriation of anthropological data: A computer program for matrix-ordering. *American Anthropologist*, 68(6):1442–1455.

Bronislaw Malinowski. 1922. *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea*. G. Routledge & Sons.

George E Marcus. 2007. Ethnography two decades after writing culture: From the experimental to the baroque. *Anthropological Quarterly*, 80(4):1127–1145.

Annette Markham, Elizabeth Buchanan, AoIR Ethics Working Committee, et al. 2012. Ethical decision-making and internet research: Version 2.0. *Association of Internet Researchers*.

Gary T Marx. 2002. What's new about the" new surveillance"? classifying for change and continuity. *Surveillance & Society*, 1(1).

Daniel Miller, Elisabetta Costa, Nell Haynes, Tom McDonald, Razvan Nicolescu, Jolynna Sinanan, Juliano Spyer, Shriram Venkatraman, and Xinyuan Wang. 2016. *How the World Changed Social Media*. UCL press.

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.

Bonnie Nardi. 2010. *My Life as a Night Elf Priest: An Anthropological Account of World of Warcraft*. University of Michigan Press.

Shyamal Patel, Hyung Park, Paolo Bonato, Leighton Chan, and Mary Rodgers. 2012. A review of wearable sensors and systems with application in rehabilitation. *Journal of Neuroengineering and Rehabilitation*, 9(1):21.

Sarah Pink, Shanti Sumartojo, Deborah Lupton, and Christine Heyes La Bond. 2017. Mundane data: The routines, contingencies and accomplishments of digital living. *Big Data & Society*, 4(1):2053951717700924.

Aaron Podolefsky and Christopher McCarty. 1983. Topical sorting: A technique for computer assisted qualitative data analysis. *American Anthropologist*, 85(4):886–890.

Mary Carol Ramos. 1989. Some ethical implications of qualitative research. *Research in Nursing & Health*, 12(1):57–63.

Patrik Svensson. 2010. The landscape of digital humanities. *Digital Humanities*.

Charles Teddlie and Abbas Tashakkori. 2009. *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Sage.

Douglas R White and Gregory F Truex. 1988. Anthropology and computing: The challenges of the 1990s. *Social Science Computer Review*, 6(4):481–497.

Crowdsourcing terminology: harnessing the potential of translator's glossaries

Ivanka Rajh*, **Siniša Runjaić†**

*Zagreb School of Economics and Management
Jordanovac 110, HR-10000Zagreb
ivanka.rajh@zsem.hr

†Department of General Linguistics, Institute of Croatian Language and Linguistics
Ulica Republike Austrije 16, HR-10000Zagreb
srunjaic@ihjj.hr

Abstract

This paper describes the history of relations, separate processes of identifying user needs and connectedness of the results of two separate surveys, a convergence of positions and the ultimate establishment of cooperation between members of the Translators and Interpreters Interest Group (TIIG) and experts from the Department of General Linguistics at the Institute of Croatian Language and Linguistics. We'll show how, regardless of the initial disparate positions, by forming smaller working groups, by accepting the idea of crowdsourcing and wider understanding of what terminological infrastructure means for two different types of stakeholders, a common goal can be defined leading to a relatively quick creation of a searchable database of translator's glossaries. The emphasis is on the fact that such a database becomes an important resource for the translator community, but at the same time provides an important contribution to the increase in the number of domains and to the quality of results on the national terminological portal created within a scientific institution.

1. Introduction

When talking about specialized languages and people who use them on everyday basis, there is always a certain disbalance present and even emphasized between linguistic theories, or theories of terminology, and usability of applied terminological resources (databases, online dictionaries, glossaries) created on the basis of theories. Although terminology is a relatively young interdisciplinary field, which emerged in the second part of the 20th century accompanied by a mass of knowledge from its "mother" lexicography and by the advances in technology upon which its application can be based, even in the most recent literature we find conclusions such as this: "[…] translation of terminological theories into real and working terminographical products has so far left a lot to be desired…" (Fuertes-Olivera and Tarp, 2014: 128). First discussions between today's collaborators, members of the Translators and Interpreters Interest Group (TIIG) and terminologists from the Department of General Linguistics at the Institute of Croatian Language and Linguistics, as well as answers of the participants in the survey (Gracin et al. 2016), revealed a lack of trust of the users from the translator community in the applicability of the terminological resources built within the scientific community, which will be examined further in the following chapters.

The basic goal of this paper is to show how initial obstacles were overcome, the role that separate user surveys played in that, the current situation and methods of gathering material for the terminological database of translators' glossaries and, finally, how that database is linked in a specific format of the metasearch engine within the Croatian Terminology Portal.

2. Translators as users of terminological resources

Translation professionals in Croatia have for a long time been organized into separate organizations depending on their specialization, such as the Croatian Association of Scientific and Technical Translators founded in 1957, The Croatian Society of Conference Interpreters (1974) and the Association of Court Interpreters and Translators (1989). Although the idea of an umbrella organization that would represent and promote interests of the entire translator community of Croatia was discussed on several occasions since Croatia gained its independence in 1990, it was only in 2009 that the Translators Group was formed within the Foreign Languages Association of the Croatian Chamber of Economy. The Translators Group was very active since the very beginning organizing its activities into eight special-interest areas, such as translation technologies, professional status and certification, and hosted several experts who held talks on topics related to translation, one of them being prof. Maja Bratanić from the Institute of Croatian Language and Linguistics, who presented the national terminology project Struna[1] at the beginning of 2010. Terminology was quickly recognized as one of the most important issues common to translators of all specializations, and the members of the Group participated in the translation into English and German of university degrees, during which they established a close cooperation with the Agency for Science and Higher Education (Pavuna, 2011).

Eventually, in 2013, the Translators Group dissociated itself from the Foreign Languages Association and became an autonomous interest group within the Croatian Chamber of Economy, i.e. Translators and Interpreters Interest Group (TIIG). One of the aims of TIIG is a „creation of a comprehensive database which would serve translators in their work, containing sources such as dictionaries, style

[1] http://struna.ihjj.hr.

guides, subject-specific glossaries, thesauruses, web forum, collaboratory, translation memories and special software…“ (Pavuna, 2016). In order to fulfill this aim, a working group named National Terminological Infrastructure was set up in 2015, its main idea being to support translators in their efforts to exert influence on the creation of terminological infrastructure and adapt that infrastructure to their practical needs. Already at that stage, members of the working group were aware of the complexity of the project, which would involve not only translators, but also IT specialists, and planned to apply for the EU funds for the purpose of its financing as well as to internationalize the project thought cooperation with experts from Slovenia and TermNet. The first task of the working group was to assess the current situation and needs of the translators' community. The following paragraphs describe the main conclusions of the survey conducted among the members of TIIG in early 2016 (Gracin et al., 2016). The survey collected answers to 37 questions from 99 translators within a period of around 40 days. The respondents were mainly freelancers (59.6%), followed by those employed in companies (27.3%), and the rest were part-time translators.

The survey showed that as many as 72% of translators find inconsistent, inappropriate or outdated solutions in terminological resources they use. When it comes to frequency of using particular resources, translators most commonly use the internet (73%), including online scientific and professional literature (65%), and multilingual texts (51%), followed by their own termbases and translation memories (40%), and inquiries among colleagues (32%), while print dictionaries represent the least consulted resource (21%). Taking into account that Croatian belongs to a group of non-dominant languages which create terms based on neology (Cabré, 1999:18), it is not surprising that translators sometimes (49.5%) or even often (18.2%) have to create neologisms. However, 82% of them think creation of neologisms is not the only solution, probably referring to the fact that terminology exists but is not readily available. This assumption is corroborated by the fact that as many as 57.5% of translators use between 20% and 40% of their time on terminological research.

Although 78% of translators cooperate with their clients in their search for best terminological solutions, it seems that the clients are not aware of the importance and value of terminological resources. Namely, clients rarely or never (63%) provide them with terminological databases relevant for the translation project, and 82% of them rarely or never ask for a submission of the database created during the project. This means that a wealth of terminological data created in a time-consuming process remains underexploited, which points to huge inefficiencies of the current (non-existent) system of terminology management. The necessity of setting up a centralized terminological system or a database is clear taking into account that 92% of respondents do not know about all terminological resources available on the internet. When it comes to two most significant online terminological resources for Croatian, the Croatian Terminology Portal and IATE, the results are equally worrying: as many as 42% of translators are not aware of their existence, and roughly 70% of those who are informed consider that those resources do not satisfy the needs of translators. Those who do use the Croatian Terminology Portal[2] are positive about it being reliable, but point to the fact that it contains a limited number of terminologies, that equivalents are available in only one or two major foreign languages, and that there are no colloquialisms and context, which translators find particularly useful. Respondents who use IATE confirm its exhaustiveness when it comes to the EU terminology, but complain about the limited number of Croatian terms, which are often inconsistent or downright wrong. It is important to emphasize that the quality of Croatian IATE has significantly improved since the beginning of 2016 when this survey among translators was conducted. According to the Croatian terminological team, the number of Croatian terms in IATE doubled from 10,005 to 20,526 between July 2015 and December 2017 (Miloš and Cimeša, 2017). Furthermore, Croatian IATE contains a low number of duplicates (only 2.5%) compared to “old” EU languages (11%), probably thanks to the inclusion of the Croatian team at the stage when system was already well functioning.

Finally, the survey shows that translators are willing to join forces in order to rectify this market failure, i.e. the lack of readily available, comprehensive and reliable terminological resources. The majority of respondents (78.8%) have their own terminological databases, and 65.7% share those with their colleagues. Furthermore, as many as 96% of translators would use a centralized terminological database and 77.8% would help in its creation, and even pay for its maintenance (65.7%). Additionally, translators consider that terminological databases, just like any knowledge, should be available to everyone regardless of their financial status. Some suggest that such database should be created and financed from the state budget and thus available to everyone for free or at a small fee like public libraries.

The results of the survey led to the overall conclusion that Croatian translators are not satisfied with the current state of terminological resources for Croatian language, especially by their suitability to translators' needs, their volume and quality. On the other hand, it was also revealed that many translators are not aware of the existence of different terminological resources for Croatian. Therefore, the working group suggested that TIIG should engage in the creation of a database of translator's glossaries, which would become an element of a web portal containing a comprehensive overview of information necessary for high quality translation work, including links to language resources, translation repositories and IT tools for translators.

3. Terminological infrastructure from the Institute of Croatian Language and Linguistics

This chapter shall clarify how, in the scientific community, the basis for the creation of the terminological infrastructure, Struna and the Croatian Terminology Portal was established, making it logical for TIIG, i.e. TIIG's

[2] http://nazivlje.hr.

working group National Terminological Infrastructure, to choose for their endeavor as partners and collaborators precisely those resources and their administrators. The attempts of establishing publicly-accessible infrastructure for specialized languages within the framework of the standard Croatian language were made on several occasions during the second half of the 20th century, when certain committees for creation of terminological dictionaries were set up, but it was only within the process of preparation for the accession of the Republic of Croatia to the EU, which included a major project of the translation of the *aquis communautaire* and accompanying documents, that a more organized terminological work on the national level began. In parallel to those processes, the scientific community realized that the establishment of a modern terminological database is a necessity, so the Croatian Science Foundation first financed a project within the program Sociocultural transition from industrial into the knowledge society, designating the Institute for Croatian Language and Linguistics as a national coordinator for the creation of Struna, a database of Croatian special-field terminology (Brač, Bratanić and Ostroški Anić, 2015: 10–15). Struna was founded in 2008 and represents a traditionally-organized normative terminological database, which is populated through a special system of scientific and professional projects financed by the Croatian Science Foundation. The projects are implemented by experts in different scientific fields assisted by linguists, i.e. terminologists-terminographers (Bratanić and Ostroški Anić, 2015: 59–68). Struna currently encompasses 20 finalized projects covering diverse fields in humanities, interdisciplinary, natural and applied sciences. Since February 2012, the general public can access more than 30,000 concepts, i.e. standardized, preferred terms, accompanied by different synonyms (admitted, deprecated, obsolete and jargon terms). Struna can serve translators as a stable source of information since every term must have at least an English equivalent, but often there are also equivalents in other European languages such as German, French or Russian. Due to the complexity of data in Struna, the search engine includes both simple and advanced types of search, with the help of special characters (wildcards) and Boolean operators. The Croatian Terminology Portal was conceived primarily as a user- and translator-oriented addition to the Struna system, as well as a central place of gathering diverse terminologies and it was publicly released after two years of preparatory work in July 2015. The portal search engine was designed as a metasearch engine, or aggregate search engine, and is simultaneously searching four separate resources: Struna, donated terminological dictionaries transformed into a terminological database form, digitally accessible resources of the Miroslav Krleža Institute of Lexicography and terminology collections of the Croatian Standards Institute (Bratanić, Ostroški Anić and Runjaić, 2017: 663–64). Since terminologies in the portal system are typologically more diverse than the Struna database, built according to the ISO recommendations and in the TBX format for data exchange, the search engine of the portal was completely simplified and made to simultaneously search all available terms in any language for a string of characters entered into a search bar, and the results can be further narrowed down and sorted depending on the user needs and preferences. There are more than 100,000 Croatian terms and more than 160,000 equivalents in foreign languages available in the portal, which contains a much wider range of terminologies than Struna. While Struna gained recognition among users as a reliable source of information since its public release in 2012, a new strategy had to be devised for the promotion of the Croatian Terminology Portal after it was launched in 2015. Since the Portal was made simple to use with the translator community in mind, the information on its launch was sent to all translator associations and translation agencies. Among the first to react was the Translators and Interpreters Interest Group (TIIG) of the Croatian Chamber of Economy, which invited authors to present the Portal during the annual meeting at the beginning of 2016. From then on, the idea of the creation of the database made up of translator's glossaries was gradually developed, including the idea of its inclusion into the Croatian Terminology Portal.

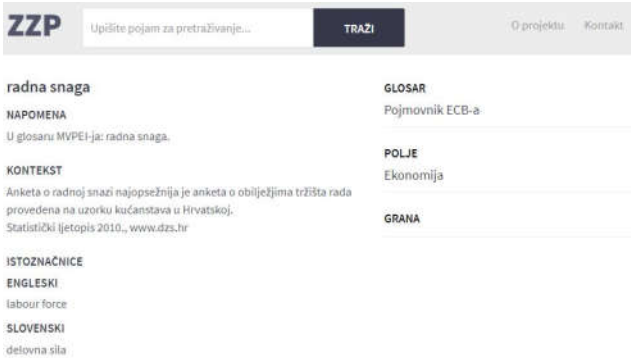
At the same time, a survey on needs and habits of the users of both terminological resources maintained by the Department of General Linguistics of the Institute of Croatian Language and Linguistics was carried out. The results of the survey were presented at the scientific conference Slavic Terminology Today in Belgrade in May 2016 (Lončar and Runjaić, 2016). For the purpose of this paper, we are going to focus on those results of the survey which correspond to the results obtained by the Translators and Interpreters Interest Group (TIIG) in their survey, and which prove that the cooperation between two organizations is a logical outcome of the answers provided by the respondents of survey conducted by the Institute[3]. The survey was made up of 32 questions, completed by 85 respondents. Demographic data point to the connection with the translators' survey due to the similar age and educational structure of respondents. Namely, the majority of respondents were in the 25-46 year age group (70.6%), of which 90.6% with a university-level diploma, while 79.8% said they were translators of different legal statuses. The most interesting information was that 63.4% of respondents use electronic dictionaries and terminological databases on (almost) daily basis, while only 20.2% search printed terminological resources. At the time, the Croatian Terminology Portal was open for public for less than 365 days, so the most important question was the one regarding frequency of visits to the Portal. The results showed that most respondents search Struna several times a month (71.8%), and 7.1% every day, while the Croatian Terminology Portal was still not as recognized in the public since only 46.3% respondents were searching it several times a month. Therefore, it was considered that the cooperation with the Translators and Interpreters Interest Group (TIIG), and especially the inclusion of translator's

[3] Of course, since the survey was taken by anonymous respondents, we cannot know whether the members of TIIG also took this survey.

glossaries alongside other resources as an integral part of the Portal search engine could significantly improve the number of visits by translators, thus improving its visibility with the aim of fulfilling its primary purpose on the one hand, and reducing the possible dissatisfaction of translators by an apparently limited scope of terminological resources publicly available in Croatia on the other.

4. Construction of the database and the search engine

Technical features and possibilities of search on the Croatian Terminology Portal were presented to the TIIG in April 2016, with a special emphasis on the new database "Terminology dictionaries and glossaries" available on the Portal. Unlike Struna, which is based on as many as 46 fields in accordance with the ISO terminology standards and the TBX (Term Base eXchange) standard and set up for a specific type of term processing in cooperation between experts and terminologists (Brač and Lončar, 2012: 261–66), the new database is significantly simplified in structure and adapted for the import of ready-made terminology manuals, which needed only minor adjustments and transformation into a terminological database form. That database is made up of basic metadata for individual terminological resource and the minimum terminological entry consists of at least one Croatian term and an equivalent in one of the European languages, depending on the type of document being converted. Thus, the database comprises both simple bilingual dictionaries without definitions[4] and, for instance, quadrilingual lexicons with longer definitions and notes, of course with the authors' permissions and respect of copyright[5]. Members of the Council of the Translators and Interpreters Interest Group (TIIG) recognized that such a system of a simplified database corresponds to their idea of terminological infrastructure for translators comprised of translator's glossaries (based on the crowdsourcing principle), and from that moment on, negotiations and preparations for the creation of the database and the search engine for translator's glossaries of the TIIG took off. It was arranged that the experts from the Department of General Linguistics, as the authors of the resource and active terminologists in the Struna and the Croatian Terminology Portal projects, provide their technical and expert know-how and experience in building the system, while the members of the Council of TIIG will invite interested members to prepare and send their glossaries to be imported into the database. The agreement was finalized at the beginning of 2017, when common efforts were invested into designing the final version of the database of translator's glossaries. We can conclude that, during the creation of the database, the potential of scientifically-based terminological specifications and the internal content management system (CMS) was used, so today we use a simple procedure of importing glossaries in their classic table formats (.xls., xlsx, etc.), which can subsequently be additionally edited by administrators in the database itself. The final searchable entry always contains the data relevant to translators such as the name and topic of a glossary, field classification, Croatian term (and possible synonyms), equivalents in foreign languages (depending on the language specialization of the translator who is the author of the glossary), and additional fields for possible information on the context and any additional remarks.



radna snaga

NAPOMENA

U glosaru MVPEI-ja: radna snaga.

KONTEKST

Anketa o radnoj snazi najopsežnija je anketa o obilježjima tržišta rada provedena na uzorku kućanstava u Hrvatskoj.
Statistički ljetopis 2010., www.dzs.hr

ISTOZNAČNICE

ENGLESKI

labour force

SLOVENSKI

delovna sila

GLOSAR

Pojmovnik ECB-a

POLJE

Ekonomija

GRANA

Figure 1. An illustration of the record radna snaga '*labour force*' in the search engine.

The searchable database was released to public in April 2018 on the website of the Croatian Chamber of Economy. The condition of the experts from the Institute of Croatian Language and Linguistics was for the database to be hosted on the same server as the other terminological resources. Furthermore, the TIIG database was linked through the application programming interface (API), thus becoming searchable on the Croatian Terminology Portal in July 2018 for further improvement of visibility and quality of results for all interested publics. Currently, it contains 19 glossaries with 11,124 Croatian terms and 14,854 equivalents in different foreign languages, and there are a few more glossaries in the pipeline. Its simple query system is in line with the needs of translators for simple term searches of any string of characters in any language[6].

5. Features of translator's glossaries

As expected, glossaries that have been sent for the inclusion in the TIIG's terminological database vary in structure, length and overall quality, so they require additional editing performed by the terminologists from the Institute. When invited to send their contributions, translators were given very broad guidelines as to the structure of their glossaries, which were further discussed during the workshop organized for that purpose in the Croatian Chamber of Economy in July 2017. An important aspect of that well-attended workshop was to remove any doubts and fears translators might have had regarding their glossaries and explain the overall benefit of their

[4] For instance Stefan Rittgasser and Ljiljana Kolenić. 2012. *Hrvatsko-njemački rječnik jezikoslovnoga nazivlja*. Hrvatsko filološko društvo.

[5] For instance Dubravka Bačun, Mirjana Matešić and Mislav Ante Omazić. 2012. *Leksikon održivog razvoja*. Hrvatski poslovni savjet za održivi razvoj.

[6] Which also corresponds to the results of the survey (Lončar and Runjaić, 2016) according to which 2/3 of the users of Struna and the Croatian Terminology Portal prefer simple search as opposed to advanced search with the help of Boolean operators.

publication for the authors themselves and the entire translator community. As to the guidelines, they were the following: glossaries should preferably be in a table format, contain a minimum of two columns (terms in Croatian and a foreign language), be consistent in the use of different symbols (e.g. semicolons to separate synonyms), cells should contain one piece of information, terms should be written in consistence with the language rules, the information on gender may be omitted. Translators were encouraged to send even short glossaries of as few as 20 terms and supply the name of the glossary, subject field and the author's name even though the authors may request to remain anonymous to the public.

The database was made very robust so as to accommodate different formats of glossaries. The content management system allows import of different categories of data (term, name of the glossary, subject field, synonym, context, source, remarks etc.), of which only four are mandatory, i.e. the term in Croatian, its equivalent in a foreign language, the title of the glossary and the subject field.

The glossaries received until now indeed vary in terms of their volume and structure: from a big Croatian-English-Slovene central banking glossary of 1720 entries with definitions to a small Croatian-Italian inheritance law glossary with 51 pairs of equivalents. Other subject fields covered by the received glossaries are: insurance, human bones, saltwater fish, winemaking, weather, real estate, world languages, scripts and regions, single euro payments area (SEPA), waters, and ecology. Editing performed on those glossaries included additional technical and orthographical corrections in the columns carried out by the Institute's members of the project, such as minor features that would not be compatible with the database CMS once the file is uploaded to the system (i.e. usage of brackets denoting the synonymy instead of two separate terms, missing or false diacritics in Croatian etc.).

Project coordinators are aware of the limitations of translator's glossaries, which shall be clearly stated on the webpage of the terminological database. Translator's glossaries are the result of a translation project and as such reflect the process of a terminology search translators find themselves in. Quite often, translators are faced with a lack of terminological resources for a particular subject or inadequacies of existing resources, so they are forced to compile their glossaries from a variety of sources of different quality and origin. Users of translator's glossaries should bear in mind that they contain terms found in the text that was being translated and not the terminology of an entire subject field. Furthermore, they may contain terms that were specific to a particular project or client requirements. However, it is expected that terms that are found in the database, even though they may not be the ones they are looking for, may help skilled translators turn their search in the right direction.

In order to populate the database with as many glossaries as possible, TIIG plans to continue promotional activities among the translator community in Croatia via different channels (workshops, website, Facebook). Additionally, project members shall contact translation services of various government institutions with the aim of detecting internal terminology resources and persuading the owners to enable the access to the general public either by their inclusion into the translator's glossary database or directly into the Croatian Terminology Portal. Namely, members of TIIG consider that terminological resources that were created by using taxpayers' money should be made available to those who paid for them.

Additional confirmation of such a belief came from the Slovenian Jožef Stefan Institute, whose researchers Simon Krek and Andraž Repar contacted TIIG in May 2018, expressed their interest in the translator's glossaries database and suggested cooperation on further collection of terminological resources, hoping that by common action, we could persuade the owners of such resources, especially public institutions, to offer them for public use. Ultimately, the collected terminological resources, including the translator's glossaries database, would be used within the eTranslation Termbank project[7], co-financed by the European Union's Connecting Europe Facility and implemented by a consortium of partners, among which Jožef Stefan Institute, is in charge of collecting resources in Slovenian, Croatian and Bulgarian.

6. Conclusion

In previous chapters, we have described a specific example of collaboration in which two separate ideas on the establishment of a comprehensive terminological infrastructure are brought together in a common IT project. Although initially the need for the creation of a database of translator's glossaries was expressed by a special-interest community, i.e. Translators and Interpreters Interest Group (TIIG), flexibility of the experts from the Department of General Linguistics of the Institute of Croatian Language and Linguistics and their technical capacities needed for the creation of an adapted terminological database provided an additional benefit for a wider translator community by inclusion of translator's glossaries into the system of the Croatian Terminology Portal. We have also demonstrated the importance of conducting user surveys for further planning of implementing activities by both the TIIG and the Institute. The readiness of the participants in the surveys to accept the idea of crowdsourcing[8] and open access to information as a necessary condition to enable the technical realization of the described process also proved to be of a great importance. The database of translator's glossaries described above, its search engine and linkage with the search engine of the Croatian Terminology Portal has been

[7] http://ettb.ijs.si/sl/etranslation-termbank/.

[8] According to the integrated definition (Estellés-Arolas and González-Ladrón-de-Guevara, 2012): "Crowdsourcing is a type of participative online activity in which an individual, an institution, a nonprofit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task; of variable complexity and modularity, and; in which the crowd should participate, bringing their work, money, knowledge **[and/or]** experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and use to their advantage that which the user has brought to the venture, whose form will depend on the type of activity undertaken."

made fully operational and open to public. The next step is to conduct surveys and similar activities in order to check the satisfaction of users with the performance and usefulness of the entire project, as well as to encourage them to participate in the creation of the database by contributing their glossaries to the project.

7. References

Ivana Brač, Maja Bratanić and Ana Ostroški Anić. 2015. Hrvatsko nazivlje i nazivoslovlje od Šuleka do Strune: hrvatski jezik i terminološko planiranje. In: *Od Šuleka do Schengena: terminološki, terminografski i prijevodni aspekti jezika struke*, pages 3–26. Institut za hrvatski jezik i jezikoslovlje and Pomorski fakultet u Rijeci.

Ivana Brač and Maja Lončar. 2012. Terminology Planning for the Croatian National Terminology Database STRUNA. In: *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012), 19–22 June 2012, Madrid, Spain*, pages 258–69.Universidad Politécnica de Madrid.

Maja Bratanić and Ana Ostroški Anić. 2015. Koncepcija i ustrojstvo terminološke baze Struna. In: *Od Šuleka do Schengena*: *terminološki, terminografski i prijevodni aspekti jezika struke*, pages 57–73. Institut za hrvatski jezik i jezikoslovlje and Pomorski fakultet u Rijeci.

Maja Bratanić, Ana Ostroški Anić and Siniša Runjaić. 2017. Od baze do portala – razvoj nacionalne terminološke infrastrukture. In: *Slovenska terminologija danas: zbornik referata izloženih na međunarodnom simpozijumu Slovenska terminologija danas*, pages 657–66. Institut za srpski jezik SANU.

Maria T. Cabré and Juan C. Sager. 1999. *Terminology theory, methods, and applications*. J. Benjamins Pub. Co.

Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara. 2012. Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2), 189–200.

Pedro A. Fuertes-Olivera and Sven Tarp. 2014. *Theory and practice of specialized online dictionaries*: *lexicography versus terminography*. Walter de Gruyter GmbH.

Damir Gracin, Tomislav Kojundžić, Vesna Kaniški, Ivanka Rajh and Damir Pavuna. 2016. Prevoditelji i terminologija. Survey conducted among the members of Translators and Interpreters Interest Group (TIIG) within the Croatian Chamber of Economy, Zagreb, from January 20 to March 8 2016.

Maja Lončar and Siniša Runjaić. 2016. Analiza korisničke upotrebe baze hrvatskoga strukovnog nazivlja Struna i Hrvatskoga terminološkog portala. Presentation held at conference *Slovenska terminologija danas*, Beograd, May 11 – 13, 2016. https://www.researchgate.net/publication/303436714_Analiza_korisnicke_upotrebe_baze_hrvatskoga_strukovnog_nazivlja_Struna_i_Hrvatskoga_terminoloskog_portala.

Irena Miloš and Saša Cimeša. 2017. Terminološki rad u Europskoj komisiji. Presentation held at conference *Translating Europe Workshop 2017*, Zagreb, December 7, 2017.

Damir Pavuna. 2011. Grupacija prevoditelja (pri HGK Zajednica za strane jezike) – krovna udruga prevoditelja. Internal document.

Damir Pavuna. 2016. O Zajednici za prevoditeljstvo. Retrieved on February 9, 2018 from http://www.hgk.hr/zajednica-za-prevoditeljstvo.

Evaluation of Statistical Readability Measures on Slovene texts

Tadej Škvorc,[*†] **Simon Krek,**[†◦] **Senja Pollak,**[†] **Špela Arhar Holdt,**[*◦] **Marko Robnik-Šikonja**[*]

[*] University of Ljubljana, Faculty of Computer and Information Science
Večna Pot 113, SI-1000 Ljubljana
tadej.skvorc@fri.uni-lj.si marko.robnik@fri.uni-lj.si

[◦] University of Ljubljana, Faculty of Arts
Aškerčeva 2, SI-1000 Ljubljana
spela.arharholdt@ff.uni-lj.si

[†] Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
simon.krek@guest.arnes.si senja.pollak@ijs.si

Abstract

The majority of existing readability measures are explicitly designed for and tested on English texts. The aim of our paper is to adapt and test the readability measures on Slovene. We test a set of 10 well-known readability formulas and 8 additional readability criteria on different types of texts: children's magazines, general magazines, daily newspapers, technical magazines, and transcriptions from the national assembly. As these groups of texts target different audiences, we assume that the differences in writing styles should also be reflected in different readability scores. Our analysis shows which readability measures perform well on this task and which fail to distinguish between the groups.

1. Introduction

In English, the problem of determining text readability (i.e. how easy a text is to understand) has long been a topic of research, with its origins in the 19th century (Sherman, 1893). Since then, many different methods and readability measures have been developed, often with the goal of determining whether a text is too difficult for it's target age group. Even though the question of readability is complex from a linguistic standpoint, a large majority of existing measures are based on simple heuristics. Nevertheless, it makes sense to apply these measures to Slovene and evaluate how well they perform, since there has been little work dedicated to this question.

There are several factors that might cause these measures to perform poorly on non-English languages, such as:

- Many measures are fine-tuned to correspond to the grade levels of the United States education system. It is likely a different fine-tuning would be needed for other languages, as a.) their education system is different from the US system, and b.) the differences in readability between grade levels are likely to be different between languages, meaning that each language would require specifically tuned parameters.

- Some measures utilize a list of common English words and their results depend on the definition of this list. For Slovene, there currently does not exist a publicly available list of common words, so it is not known how such measures would perform.

- The measures do not use the morphological information to determine difficult words but rely on syllable and character counts, or a list of difficult words. As Slovene is morphologically much more complex than English, words with a more complex morphology are likely harder to understand than those with a simple morphology, even if they have the same number of characters or syllables.

These are only a few of the reasons explaining why it is hard to evaluate the performance of the original measures on other languages. In this paper, we analyze the commonly used readability measures (as well as some novel measures) on Slovene texts and propose a word list needed for implementing the word-list-based measures. We calculate statistical distributions of scores for each readability measure across subcorpora and assess the ability of measures to distinguish between different subcorpora.

The paper is structured as follows. In Section 2. we present the related work on readability measures. In Section 3. we describe the readability measures used in our analysis. The methodology of the analysis is presented in Section 4. The results are contained in Section 5. and Section 6. concludes the paper.

2. Related Work

For English, there exists a variety of works focused on determining readability by using readability formulas. Those formulas rely on different features of the text such as average sentence length, percentage of difficult words, and the average number of characters per word. Examples of such measures are given in Section 3. and include the Coleman-Liau index (Coleman and Liau, 1975), LIX (Björnsson, 1968), and the automated readability index (ARI) (Senter and Smith, 1967). Some formulas, like the Flesch-Kincaid grade level (Kincaid et al., 1975) and SMOG (Mc Laughlin, 1969) use the number of syllables per word to determine if a word is difficult. Additionally, some measures (e.g., the Spache readability formula

(Spache, 1953) and Dale-Chall readability formula (Dale and Chall, 1948)) rely on a pre-constructed list of difficult words.

Aside from readability formulas, there exists a variety of other approaches that can be used to determine readability (Bailin and Grafstein, 2016). For example, various machine-learning approaches can be used to obtain better results than readability formulas, such as the approach presented in François and Miltsakaki (2012) which outperforms readability formulas on French text.

To the best of our knowledge, there is little existing work that attempts to apply these measures to Slovene texts. Most work dealing with readability of Slovene text is focused on manual methods. For example, Justin (2009) analyzes Slovene textbooks from a variety of angles, including readability. Works that focus on automatic readability measures are rare. Zwitter Vitez (2014) uses a variety of readability measures for author recognition in Slovene text, but we found no works that used them to determine readability.

In addition to Slovene, some related work evaluates readability measures on other languages. Debowski et al. (Debowski et al., 2015) evaluate readability formulas on Polish text and show that they obtain better results by using a more complex, machine-learning-based approach.

3. Readability Measures

In our analysis, we used two groups of readability measures:

Existing readability formulas for English: we focused mainly on popular methods that have been shown to achieve good results on English texts. These measures mostly rely on easy-to-obtain features such as number of difficult words, sentence length and word length).

Additional readability criteria: we used additional criteria that are not present in the existing readability formulas, such as the percentage of verbs, number of unique words, and morphological difficulty of words. In English formulas, such criteria are not used, but they might contain useful information for readability of Slovene texts.

In this section, we present these two groups of readability measures. In Section 3.1. we present the established readability measures for grading English text and in Section 3.2. we present the additional criteria.

3.1. Existing Readability Formulas

There exists a variety of ways to measure readability of texts written in English. For our analysis, we used 10 readability formulas given below. The entities used in the expressions correspond to the number of occurrences of a given entity, e.g., word corresponds to the number of words in a measured text.

Gunning fog index (Gunning, 1952) is calculated as:

$$\text{GFI} = 0.4\left(\frac{\text{words}}{\text{sentences}} + 100\frac{\text{complex words}}{\text{words}}\right),$$

where a word is considered complex if it contains three or more syllables [1]. The resulting score is calibrated to the grade level of the USA education system.

Flesch reading ease (Kincaid et al., 1975) is calculated as:

$$\text{FRE} = 206.835 - 1.015\frac{\text{words}}{\text{sentences}} - 84.6\frac{\text{syllables}}{\text{words}}.$$

The score does not correspond to grade levels. Instead, the higher the value is the easier the text is considered to be. A text with a score of 100 should be easily understood by 11-year-old students, while a text with a score of 0 should be intended for university graduates.

Flesch–Kincaid grade level (Kincaid et al., 1975) is similar to Flesch reading ease, but does correspond to grade levels. It is calculated as:

$$\text{FKGL} = 0.39\frac{\text{words}}{\text{sentences}} + 11.8\frac{\text{syllables}}{\text{words}} - 15.59.$$

Dale–Chall readability formula (Dale and Chall, 1948) is calculated as:

$$DCRF = 0.1579\frac{\text{difficult words}}{\text{words}} + 0.0496\frac{\text{words}}{\text{sentences}}.$$

The formula requires a predefined list of common (easy) words and the words which are not on the list are considered as difficult. The originality of the Dale-Chall Formula was that it did not use word-length counts but uses a count of 'hard' words, which are the words that do not appear on a specially designed list of common words. This list was defined as the words familiar to most of the 4th-grade students: when 80 percent of the fourth-graders indicated that they knew a word, the word was added to the list.

Higher scores indicate that the text is harder, but the resulting score does not correspond to grade levels, nor is it appropriate for text aimed at children below 4th grade. In our analysis, we obtained the difficult words in two ways:

1. By constructing a list of 'easy' words and considering every word not on the list as difficult. The list of easy words is described in Section 4.2..
2. By considering words with more than seven characters as difficult.

Spache readability formula (Spache, 1953) is calculated as:

$$\text{SRF} = 0.141\frac{\text{words}}{\text{sentences}} + 8.6\frac{\text{unique difficult words}}{\text{unique words}} + 0.839.$$

Difficult words are defined as words that do not appear in the list of commonly used words, which is the same as the one used in the Dale–Chall readability formula. This method was specifically designed for texts targeting children up to the fourth grade, and was not designed to perform well on harder text. The obtained score corresponds to grade levels.

[1] As there exists no established automatic method for counting syllables of Slovene words, we used a rule-based approach designed for English.

Automated readability index (Senter and Smith, 1967) is calculated as:

$$\text{ARI} = 4.71\frac{\text{characters}}{\text{words}} + 0.5\frac{\text{words}}{\text{sentences}} - 21.43.$$

The formula was designed so that it could be automatically captured in the times when texts were written on typewriters and therefore does not use information relating to syllables or difficult words. The obtained score corresponds to grade levels.

SMOG (Simple Measure of Gobbledygook) (Mc Laughlin, 1969) can be calculated as:

$$\text{SMOG} = 1.043\sqrt{\text{difficult words}\frac{30}{\text{sentences}}} + 3.1291,$$

where difficult words are defined as words with three or more syllables. The score corresponds to grade levels.

LIX (Björnsson, 1968) is calculated as:

$$\text{LIX} = \frac{\text{words}}{\text{sentences}} + 100\frac{\text{long words}}{\text{words}},$$

where long words are defined as words consisting of more than six characters. LIX is the only measure we used that was not designed specifically for English but for a variety of languages. Because of this, it does not use syllables or a list of unique words. The score does not correspond to grade levels.

RIX (Anderson, 1983) is a simplification of LIX, and is calculated as:

$$\text{RIX} = \frac{\text{long words}}{\text{sentences}}.$$

Coleman-Liau index (Coleman and Liau, 1975) is calculated as:

$$\text{CLI} = 0.0588L - 0.296S - 15.8,$$

where L is the average number of letters per 100 words and S is the average number of sentences per 100 words. The obtained score corresponds to grade levels.

3.2. Additional readability criteria

As mentioned in Section 1. the readability formulas mentioned in Section 3.1. are simple and use a low number of common criteria, such as the number of syllables in words or the number of words in a sentence. In our analysis, we also analyzed Slovene texts using the following additional statistics:

- percentage of long words,
- percentage of difficult words,
- percentage of verbs,
- percentage of adjectives,
- percentage of unique words,
- average sentence length.

Most of these (percentage of long words, difficult words, unique words, and average sentence length) are used as features in the readability measures described above. We evaluate them individually to determine how important each of them is for Slovene texts. The percentage of verbs is used because a higher number of verbs can indicate more complex sentences with multiple clauses. The percentage of adjectives was chosen because we assumed a higher percentage of adjectives could indicate longer, more descriptive sentences that are harder to understand. To take into account richer morphology of Slovene and a less fixed word order compared to English, we computed two additional criteria:

Context of difficult words, which is the average number of difficult words that appear in a context (i. e. the three words before or after the word) of a difficult words. The difficult words are defined as words that do not appear on the list of common words. The intuition behind this metric is that a difficult word that appears in the context of easy words is easier to understand than if it was surrounded by other difficult words.

Average morphological difficulty. To calculate this, we use Sloleks (Arhar Holdt, 2009) to assign a morphological richness score to each word. Sloleks contains frequency information for morphological variants of over 100 000 lemmas, and we use the relative frequency of a variant compared to other variants of the same lemma as the morphological difficulty score.

We also collected the number of words in each document. In our case, this was not a useful criterion for determining readability since it was largely determined by the type of document (e.g., the documents belonging to the subcorpus of newspapers contained individual articles and were therefore short, while computer magazines contained the entire magazine and were longer).

4. Analysis of Slovene texts

In this section, we describe the methodology used for our analysis. In Section 4.1. we describe the datasets on which we conducted our analysis and in Section 4.2. we describe how we constructed the list of easy words used in some of the readability measures.

4.1. Datasets

For the analysis we have created a set of subcorpora from the Gigafida reference corpus of written Slovene (Logar et al., 2012). Gigafida contains 39 427 Slovene texts released from 1990 to 2011, for a total of 1 187 002 502 words. We focused on texts published in magazines, newspapers, and books while ignoring texts collected from the internet. The texts in the Gigafida corpus are tokenized, segmented into sentences and paragraphs, and part-of-speech tagged using the Obeliks tagger (Grcar et al., 2012). To determine the performance of readability measures we grouped them based on the intended audience, obtaining the following subcorpora.

Children's magazines include magazines aimed at younger children (to be read from by their parents), namely Cicido and Ciciban.

Pop magazines contain magazines aimed at the general public, namely Lisa, Gloss, and Stop.

Newspapers contain general adult population newspapers, namely Delo and Dolenjski list.

Computer magazines include magazines focusing on technical topics relating to computers, namely Monitor, Računalniške novice, PC & Mediji, and Moj Mikro.

National Assembly includes transcriptions of sessions of the National Assembly of Slovenia.

In Table 1 we show the number of documents in each subcorpus and the average number of words per document. The subcorpus of newspapers contained the largest number of documents, while the subcorpus of text sourced from the National Assembly of Slovenia contained the fewest.

Subcorpus	#docs	Avg. #words / doc
Children's magazines	125	5,488
Pop magazines	247	33,967
Newspapers	14,011	12,881
Computer magazines	163	110,875
National Assembly	35	58,841

Table 1: The number of documents and the average number of words per document for each subcorpus.

Our hypothesis is that the readability measures will be able to distinguish texts from different subcorpora. We assume children's magazines will be easily distinguishable from other genres that are addressing adult population. We also suppose that general magazines are less complex than specialized magazines. The National Assembly transcripts were included as they differ from other texts in two major ways: a.) they are transcripts of spoken language and b.) they relate to a highly technical subject matter. Because of this we were interested in how readability measures would grade them. To test our hypothesis, and to determine how well each readability measure works, we analyzed texts from each subcorpus to obtain score distribution for each measure. The scores were calculated separately for each source text (e.g., one magazine article, a newspaper, or one assembly session).

4.2. List of common words

For designing the list of common words, we took a corpus-based approach. Note that the methodology to create a list of common words from language corpora was already tested for other languages, see e.g., (Kilgarriff et al., 2014). From the corpora Kres, Janes, Gos and Šolar, we extracted the most common words and defined common words as the ones which appear in all four corpora (and are therefore not specific to a certain text type). With four corpora we aimed at an inclusion of corpus texts that primarily reflect language production by different language users (GOS, JANES, Šolar), as well as corpus texts that primarily reflect the language community's every-day language reception (Kres). We aimed at covering younger speakers (e.g. Šolar) and adult production. For some corpora, we could have assigned words to different age levels (e.g. using pupils' grade levels in Šolar or using the age groups available in GOS metadata), but these corpora are very specific and the resulting word groups would mainly reflect the genre instead of age levels. Because of this we opted for the approach of crossing the word lists to obtain a single list. The overlap of the most common words in four corpora eliminates frequent words which are reflecting only one of the corpora (e.g. administrative language in Kres, spoken language markers in GOS, Twitter-specific usage in Janes and literary references from essays in Šolar).

More details on the four corpora used as a source of information for commonly used words, are provided below.

Šolar (Kosem et al., 2011) contains 2703 texts written by pupils in Slovenia from grades 6 to 13 (grade 6 to 9 in primary school, and grade 1 to 4 in secondary school). The texts include essays, summaries, and answers to examination questions.

GOS (Verdonik et al., 2011) contains around 120 hours of recordings of spoken Slovene (1 035 101 words), as well as transcriptions of the recordings. The recordings are collected from a variety of sources, including conversations, television, radio, and phone calls. Around 10% of the corpus consists of recorded lessons in primary and secondary schools.

JANES (Fišer et al., 2014) contains Slovene texts from various internet sources, such as tweets, forum posts, blogs, comments, and Wikipedia talk pages.

Kres (Logar Berginc and Šuster, 2009) is a sub-corpus of Gigafida that is balanced with respect to the source (e.g. newspaper, magazine or internet).

From each corpus, we extracted the top 10 000 most frequent word lemmas and part-of-speech tuples. In order to construct a list of common words representative of Slovene language, we selected the word lemmas that occurred in the most frequent word lists of all the four corpora. We obtained a list of 2562 common words which we used in readability measures. In this paper, we are using the automatically assembled version of the list as described above. In future work, the list will be linguistically analyzed, refined, and made publicly available for further use.

5. Results

For each text in each subcorpus, we calculated readability scores using all readability measures described in Section 3. In Figure 1 we present a few examples of obtained score distributions. We show distributions for three text subcorpora (children's magazines, newspapers, and technical magazines) and three readability scores (Goobledygook, Coleman-Liau, and average sentence length).

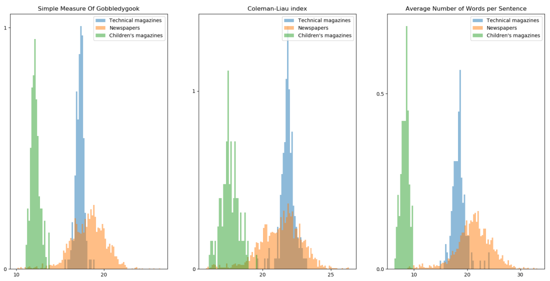


Figure 1: The score distributions for three text subcorpora and three readability measures. The distributions show that technical magazines readability scores are the most consistent, while newspapers' scores are more diverse. Children's magazines' scores have a strong peak on the left-hand side (easier texts) that is well separated from the other sources.

To show a compact overview of all included readability measures we calculated the median, first, and third quartiles of the distribution for each score and each text subcorpus. The box-and-whiskers plots showing these results are visualized in Figure 2 which shows that most readability measures are able to distinguish between different subcorpora. Additionally, some of the readability measures fit our original hypothesis, i.e. they are able to distinguish children's magazines from other genres that are addressing adult population, and evaluate general magazines as less complex than computer magazines.

Figure 2 also allows for additional interpretation of readability measures. For example, children's magazines vs. general magazines vs. newspapers mean scores show increasing complexity in the following measures: Percentage of long words, Flesh Kincaid Grade Level, Gunning Fog Index, Dale-Chall Readability Formula (based on complexity defined by syllables), Context of Difficult Words, SMOG, LIX, RIX and Automated Readability Index. All these measures consider the length of words and/or sentences. The percentage of adjectives also seems to correlate with the complexity of these three text types, although to a lesser extent. The same holds for Flesh Reading Ease, since higher scores indicate lower complexity. For the majority of these measures, the distinction between newspapers and specialized computer magazines is either less evident or not evident at all, but they do indicate that computer magazines are less readable than general magazines.

Scores using the list of common words do not lead to the same conclusions. Percentage of Difficult Words and Dale-Chall Readability Formula with word list do not reflect the complexity of genres, but to some extent they do distinguish between general and specialized texts (i.e. newspapers and general magazines have lower scores than specialized computer magazines). One of the reasons for the relatively high scores for complexity of children magazines might be in the large proportion of literary language, such as in poems for children with many words not in the list of common words. For example, "KRAH, KRAH, KRAH! MENE NIČ NI STRAH!" has 7 words, out of which 4 are on the list of simple words, while the word KRAH is not on the simple words list. Therefore the proportion of difficult words in this segment is 42.8% (3 occurences of word KRAH out of 7 words in total). On the other hand, the words are short, therefore length-based measures consider them to be simple words.

The readability scores for the National Assembly subcorpus show high variability across the measures, which might also be attributed to the fact that it is a different genre (spoken, but specialized). E.g., in several measures where the readability complexity rises from children's magazines to general magazines and newspapers, the National assembly scores are close to general magazines. Very long words might be used in spoken language with lower probability, even in a political context. Average morphological difficulty and context of difficult words lead to the interpretation that this genre is more complex (less "readable"). The very high score for context of difficult words might be attributed to enumeration of Assembly members (e.g., "Obveščen sem, da so zadržani in se današnje seje ne morejo udeležiti naslednje poslanke in poslanci: Ciril Pucko, Franc Kangler, Vincencij Demšar, Branko Kalalemina, ..."). The relatively high percentage of verbs can also be interpreted from this perspective, e.g., the National assembly text include many performatives, such as "Pričenjam nadaljevanje seje" and "Ugotavljamo prisotnost v dvorani".

In summary, using a list of common words described in Section 4.2. did not improve the separation of the text subcorpora perceived as easy and difficult to read. Both measures that use them (Dale-Chall and Spache readability formulas) are poor separators. A number of simple readability measures worked well, such as the percentage of long words, percentage of verbs/adjectives, and the average morphological difficulty.

We also calculated the sample mean and standard deviation of readability measures for each text subcorpus. The results are shown in Table 2.

Using these results, we calculated the Bhattacharyya distance between the distributions of Children's magazines and newspapers for each score. The Bhattacharyya distance measures the similarity between two statistical distributions. We assumed the scores were distributed normally, as the results shown in Figure 1 show the scores approximately follow a normal distribution, and calculated the distance using the following formula:

$$D_B(p,q) = \frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{\sigma_p^2}{\sigma_q^2}+\frac{\sigma_q^2}{\sigma_p^2}+2\right)\right)+\frac{1}{4}\left(\frac{(\mu_p-\mu_q)^2}{\sigma_p^2+\sigma_p^2}\right).$$

We also show the Bhattacharyya coefficient, which measures the overlap between two statistical distributions and can be calculated as:

$$\mathrm{BC} = e^{-D_B(p,q)}$$

The results are presented in Table 3. These results are similar to the ones shown in Figure 2, with the readability formulas using the list of difficult words showing less dichotomization power. The largest distance is obtained using average sentence lengths.

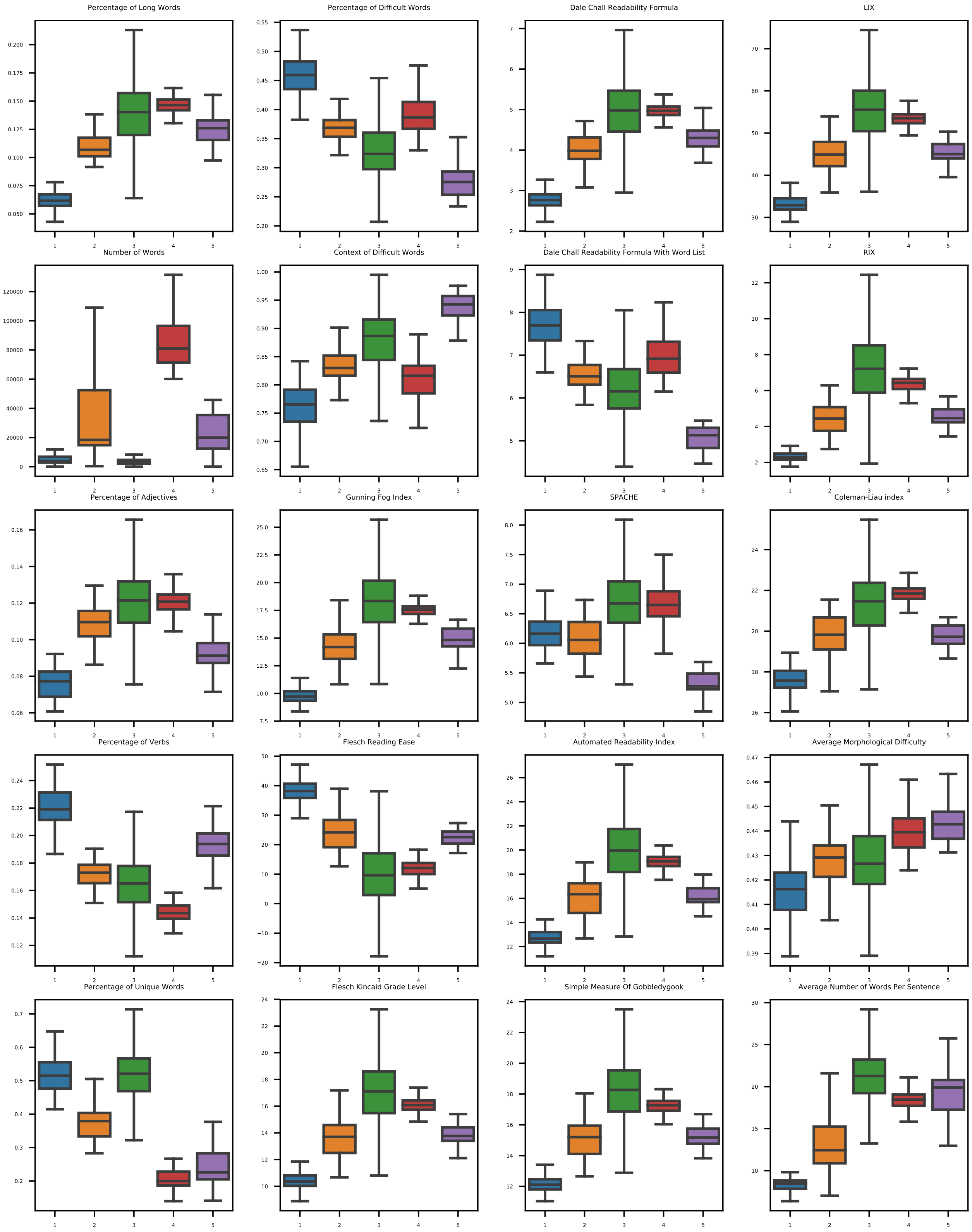


Figure 2: The scores of each readability measure for each subcorpus of texts, represented with box plots. The subcorpora are: 1.) Children's magazines, 2.) General magazines, 3.) Newspapers, 4.) Computer magazines, and 5.) National assembly text. The boxes show the first, second, and third quartile of the distributions while the whiskers extend for 1.5 IQR past the first and third quartile.

Measure	Children's mag.	Magazines	Newspapers	Technical mag.	National assembly
% long words	0.065 (0.015)	0.109 (0.011)	0.137 (0.029)	0.146 (0.010)	0.137 (0.046)
Number of words	5488 (6184)	33966 (34821)	12881 (84708)	110875 (151007)	58841 (106515)
% adjectives	0.078 (0.016)	0.111 (0.013)	0.120 (0.020)	0.120 (0.008)	0.096 (0.022)
% verbs	0.216 (0.026)	0.170 (0.015)	0.161 (0.034)	0.144 (0.013)	0.180 (0.044)
% unique words	0.517 (0.077)	0.375 (0.053)	0.513 (0.114)	0.244 (0.144)	0.277 (0.173)
Context of difficult words	0.756 (0.054)	0.834 (0.027)	0.849 (0.133)	0.808 (0.036)	0.929 (0.044)
% difficult words	0.464 (0.048)	0.369 (0.022)	0.356 (0.122)	0.389 (0.032)	0.280 (0.036)
Gunning Fog Index	9.950 (1.255)	14.272 (1.271)	18.662 (9.319)	17.470 (0.800)	15.901 (3.493)
Flesch reading ease	37.592 (4.989)	23.855 (5.217)	10.002 (24.128)	12.520 (4.340)	19.178 (13.098)
Flesch–Kincaid grade level	10.500 (0.894)	13.596 (1.193)	17.356 (8.959)	15.999 (0.741)	14.523 (2.761)
Dale–Chall	2.845 (0.425)	4.036 (0.306)	4.972 (1.270)	4.941 (0.258)	4.560 (0.971)
Dale–Chall with word list	7.781 (0.720)	6.534 (0.357)	6.643 (2.163)	6.955 (0.484)	5.208 (0.539)
Spache readability formula	6.217 (0.368)	6.079 (0.348)	6.977 (3.499)	6.685 (0.323)	5.482 (0.600)
Automated readability index	12.873 (1.086)	16.117 (1.428)	20.474 (11.456)	19.007 (0.885)	17.014 (3.371)
SMOG	12.206 (0.759)	15.095 (1.066)	18.200 (2.757)	17.194 (0.611)	15.849 (2.500)
LIX	33.676 (3.384)	44.999 (3.282)	56.016 (23.123)	53.260 (2.077)	47.909 (9.073)
RIX	2.381 (0.496)	4.481 (0.781)	7.370 (3.836)	6.354 (0.518)	5.250 (2.574)
Coleman-Liau index	17.785 (1.120)	19.823 (0.861)	21.220 (1.807)	21.762 (0.903)	20.318 (2.170)
Avg. morphological difficulty	0.419 (0.017)	0.428 (0.010)	0.436 (0.044)	0.441 (0.017)	0.445 (0.026)
Avg. sentence length	8.353 (0.820)	13.389 (2.843)	21.120 (4.043)	18.641 (1.960)	19.063 (3.826)

Table 2: The mean and standard deviation for each subcorpus of texts and each readability score.

Measure	Distance	Coefficient
Average sentence length	**2.866**	**0.057**
SMOG	1.433	0.239
% long words	1.350	0.259
RIX	1.101	0.333
Flesch–Kincaid grade level	0.956	0.385
Automated readability index	0.945	0.389
Dale–Chall readability formula	0.885	0.413
Gunning fog index	0.880	0.415
LIX	0.853	0.426
Spache readability formula	0.797	0.451
Flesch reading ease	0.776	0.460
% adjectives	0.719	0.487
Coleman-Liau index	0.708	0.493
% verbs	0.432	0.649
% difficult words	0.365	0.694
Dale–Chall with word list	0.318	0.728
Context of difficult words	0.285	0.752
Avg. morphological difficulty	0.235	0.790
% unique words	0.039	0.961

Table 3: The Bhattacharyya distances and coefficients between the distributions of scores for children's magazines and newspapers for each readability measure. The results are sorted by decreasing distance.

6. Conclusion and Future work

We analyze statistical distributions of well-known readability measures designed for English on Slovene texts. We extract five subcorpora of texts from the Gigafida corpus with commonly perceived different readability levels: children magazines, popular magazines, newspapers, technical magazines, and national assembly texts. We find that the readability formulas are able to distinguish between these subcorpora reasonably well, with the exception of national assembly texts, which are of a different, spoken, genre and the measures were not originally designed to handle it. A number of simple readability statistics, such as the context of difficult words and average sentence length, also dichotomize the different subcorpora of text.

In this work, we only focus on simple readability formulas along with some additional readability criteria. There exists a variety of more complex methods for evaluating the complexity of text, such as the one presented in (Lu, 2009) and (Wiersma et al., 2010). More advanced methods might be more suitable for Slovene texts than the simple methods used in this paper.

Most of the English readability formulas were designed to correlate with school grades and were tested on that domain. For Slovene, there currently does not exist a publicly available dataset where texts are tagged according to the grade level they are appropriate for. This makes analyzing the readability measures from this perspective difficult. In the future work, we plan prepare such a corpus and design several readability scores fit for different purposes. This will also allow us to frame determining readability as a classification problem with the goal of predicting the grade level of a text. A similar approach that is also worth considering would be to have experts annotate texts with readability scores. This would allow us to fit a regression model using the readablilty measures analyzed in this paper.

Another area that we plan to explore is the use of coherence and cohesion measures (Barzilay and Lapata, 2008), (Crossley et al., 2016), which are used to determine if words, sentences, and paragraphs are logically connected. Coherence and cohesion methods usually rely on machine learning approaches that can rely on language specific features and would therefore need to be evaluated on Slovene text. The same applies to readability measures that rely on machine learning (François and Miltsakaki, 2012), which we also plan to analyze in the future.

Acknowledgements

The research was financially supported by the Slovenian Research Agency through project J6-8256 (New grammar of contemporary standard Slovene: sources and methods), project J5-7387 (Influence of formal and informal corporate communications on capital markets), a young researcher grant, research core fundings no. P2-0209 and P2-0103; Republic of Slovenia, Ministry of Education, Science and Sport/European social fund/European fund for regional development/European cohesion fund (project Quality of Slovene textbooks, KaUč).

7. References

Jonathan Anderson. 1983. Lix and Rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Špela Arhar Holdt. 2009. Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo*, 54(3–4):43–56.

Alan Bailin and Ann Grafstein. 2016. *Readability: Text and context*. Springer.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Carl Hugo Björnsson. 1968. *Läsbarhet*. Liber.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4):1227–1237.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Łukasz Debowski, Bartosz Broda, Bartłomiej Nitoń, and Edyta Charzyńska. 2015. Jasnopis–a program to compute readability of texts in polish based on psycholinguistic research. *Natural Language Processing and Cognitive Science*, page 51.

Darja Fišer, Tomaž Erjavec, Ana Zwitter Vitez, and Nikola Ljubešić. 2014. Janes se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. *V: Jezikovne tehnologije: zbornik*, 17:56–61.

Thomas François and Eleni Miltsakaki. 2012. Do nlp and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57. Association for Computational Linguistics.

Miha Grcar, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statisticni oblikoskladenjski oznacevalnik in lematizator za slovenski jezik. In *Zbornik Osme konference Jezikovne tehnologije, Ljubljana, Slovenia*.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

J Justin. 2009. *Učbenik kot dejavnik uspešnosti kurikularne prenove: poročilo o rezultatih evalvacijske študije*. Ljubljana: Pedagoški inštitut.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Iztok Kosem, Tadeja Rozman, and M Stritar Kučuk. 2011. How do Slovenian primary and secondary school students write and what their teachers correct: A corpus of student writing. In *Proceedings of Corpus Linguistics Conference 2011, ICC Birmingham*, pages 20–22.

Nataša Logar Berginc and Simon Šuster. 2009. Gradnja novega korpusa slovenščine. *Jezik in slovstvo*, 54(3–4):57–68.

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

Xiaofei Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28.

G Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading*, 12(8):639–646.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, University of Cincinnati, Ohio.

Lucius Adelno Sherman. 1893. *Analytics of literature: A manual for the objective study of English prose and poetry*. Ginn, Boston.

George Spache. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

Darinka Verdonik, Ana Zwitter Vitez, and Hotimir Tivadar. 2011. *Slovenski govorni korpus Gos*. Trojina, zavod za uporabno slovenistiko.

Wybo Wiersma, John Nerbonne, and Timo Lauttamus. 2010. Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing*, 26(1):107–124.

Ana Zwitter Vitez. 2014. Ugotavljanje avtorstva besedil: primer ”trenirkarjev”. In *Language technologies: Proceedings of the 17th International Multiconference Information Society – IS 2014*, pages 131–134, Ljubljana, Slovenia, October. Institut Jožef Stefan.

Exploring Finno-Ugric linguistics through solving IT problems

Tobias Weber,*[*] Jeremy Bradley[‡]

[*]Ludwig Maximilian University of Munich
Institut für Finnougristik / Uralistik
Ludwigstraße 31/III
D-80539 München
weber.tobias@campus.lmu.de

[‡]University of Vienna
Institut EVSL, Abteilung Finno-Ugristik
Campus AAKH, Hof 7-2, Spitalgasse 2-4
A-1090 Wien
jeremy.moss.bradley@univie.ac.at

Abstract

This paper seeks to introduce our approach of integrating computational methods, digital resources, and computer literacy skills into the curriculum of Finno-Ugric (Uralic) linguistics. Our starting point is the class Digital Resources in Linguistics, which we taught at the Institute of Finno-Ugric/Uralic Studies at LMU Munich in 2017; our eventual aim is the compilation of teaching materials (a textbook with supplementary online materials) on this subject matter and their integration into Finno-Ugric curricula. While there are numerous high-quality textbooks on computational linguistics, our endeavour is more tied to the framework of Digital Humanities, stressing the background in humanities and social sciences rather than details of specific technologies, and attempting to be conscientious to the specific needs, interests, and skills of our students. This endeavour is happening within the context of the ongoing internationalization of our research discipline, exemplified by the Erasmus+-strategic partnership INFUSE (Integrating Finno-Ugric Studies in Europe, 2015–2018), which in its next iteration, COPIUS (Community of Practice in Uralic Studies, 2018–2021) will also focus on the development and pooling of teaching materials.

1. Introduction

In order to discuss the challenges of conveying computer literacy skills in classrooms of subjects traditionally associated to the humanities, we need to consider our targeted audience first and explore their access to computers and programming. An overview of existing literature will help shed some light onto issues our students may experience in using these materials. Based on these observations we will outline our approach by highlighting overlaps between the topics which are already covered in our courses and computational methods and tools that can be used in relation to them. Finally, we will present ways of integrating our concept into the curricula of the European institutes for Uralic studies and exemplify them using the curriculum at LMU Munich, where we have taught our pilot course.

1.1. Target audience

To understand our role in this endeavour, let us step back and discuss our professional relationship to our students and consider their needs and interests in acquiring computer literacy skills. As educators at university level in the digital era, we, in spite of teaching in a discipline traditionally seen as part of the humanities, instruct students who have been exposed to advanced technologies throughout their previous educational careers as well as in their social lives. This means that we are not building knowledge from scratch in absence of a preexisting foundation. However, our students are enrolled in linguistic, philological, or ethnographic courses, and are generally aiming to acquire an education within the domain of the humanities, rather than in more technically-oriented subject fields. They generally are users of applications which they find online or through recommendations by teachers or supervisors but have no expertise in developing applications of their own, or the intention of doing so.

From conversations with our students prior to our course, we had gathered that most regard computational methods as too abstract and not relevant enough for their own research, and that they are reluctant to use applications or technologies with which they do not feel confident in their research. At the same time, it was obvious to us that many tasks our students face on a regular basis could be streamlined if they could overcome these reservations. Our approach is informed by the discord between reservations our students - who do not consider themselves “tech people” - feel, and the profit they could garner from basic IT literacy in their work. Our course aims to create a basic understanding of how computers handle language data, and help students develop a critical view of the possibilities and limitations in electronic data processing. We are educating “cross-disciplinary thinkers” (Furman, 2015, p. 4). Whether students go on to take classes in computational linguistics of digital humanities or not, they will see computers in a different light. Even students who do not themselves start working with software tools we cover in our courses will profit from this better understanding, as it will enable better communication with programmers and coders in collaborative efforts.

1.2. Approaches in existing literature

Computer-Assisted Language Learning constitutes one important field of study within language pedagogy for linguists. While this field is only peripherally related to our endeavour, it comprises many insightful essays on the use of computers in classrooms of language and other "soft sciences". Among these papers, there are also critical voices calling for considerate use of computers and technology, a position which is derived from the incorporation of insights from media science and philosophy. Richard Kern (2014) uses the term *pharmakon* to describe the role of technology in a language classroom – it can be both the cure for an ailment and a poison. This point of view calls upon educators and students to deal with computers and technology critically, i.e. to acknowledge pitfalls as well as benefits.

The amount of literature on NLP and computer skills for linguists has increased immensely over the last decade. Some books use the "dummies" approach and assume as little prior knowledge as possible or try to convey practical skills in using computers and particular software. Among those, some authors claim to teach broad skills for the labour market or to "prepare you for success in a modern world full of computers" (Wempen, 2014, p. 1). While the role of computer literacy in private economy cannot be underestimated, it appears that a university course should primarily tackle specific issues within the discipline but also convey skills from which students will profit in their later lives regardless of the career paths they choose. Another subset of literature is directly aimed at students of computational linguistics, e.g. (Jurafsky and Martin, 2009; Carstensen et al., 2009), usually coming from a technology background and with the goal of teaching skills for software applications as well as programming and coding languages. These books are, without doubt, the benchmark for textbooks on computational linguistics and it should be the hope of every educator in digital humanities that their students can go on to read and understand this set of literature or acquire practical knowledge of a programming language, should it match their interests and needs. However, as outlined above, our target audience is not particularly interested in writing programmes and would feel easily intimidated when confronted with theoretical concepts from computer science or mathematics (e.g. discrete mathematics for the description of automata or formal languages). Hence, our objective is rather to foster a general understanding and awareness of how applications relevant to our subject field work.

It should furthermore be the goal of our course and teaching materials to explain these abstract concepts as practical knowledge and formalism as an excursus rather than primary content of the class. This becomes most relevant in assessment, where such knowledge should not be tested explicitly – students should be enabled to understand the concepts but not be tested on formalisms. A few textbooks take this approach, e.g. (Dickinson et al., 2012), i.e. that "[t]he goal of our courses is to show students the capabilities of [NLP] tools, and especially to encourage them to take a reflective and analytic approach to their use." (Dickinson et al., 2012, p. xiii). This textbook, like the teaching materials we aim to create, puts emphasis on the discussion of computational methods, on students' own work, as well as in the academic community – questioning the "engineering mentality" (Popoveniuc, 2010) and the sociocultural aspect of technology in scientific discourse, cf. (Schmidt, 2010).

2. Why does it matter?

The relevance of our project, in spite of the large existing body of textbooks on computational linguistics, stems from the discrepancy between the issues and approaches followed in mainstream textbooks and topics relevant to our target audience: first of all, literature covering NLP issues on Uralic languages is still scarce. This is not to say that there is no scientific output on NLP pertaining to Uralic languages, but that publications are either very specific in their target language (mostly on the three Uralic national languages of Europe, Hungarian, Finnish, and Estonian) or have a strong focus on technical issues (e.g. publications by Giellatekno, the Centre for Saami language technology; publications in the Northern European Journal of Language Technology). We consider our project to act as a bridge between "paper and pencil" linguistics and the research carried out by computational linguists by facilitating students' access to this field of study, or at least educating them about the range of possibilities offered to the study of Uralic languages by NLP.

Furthermore, Uralic linguistics has a long research tradition which has given rise to peculiarities in terminology or practices in handling language data, e.g. the Finno-Ugric Transcription FUT (Setälä, 1901), which predates IPA as a transcription standard. This means that working with Uralic language data requires the researcher to know about these conventions. While transliteration between transcription systems does not pose an obstacle to a computational linguist, cf. (Bradley, 2017), it can dishearten scholars outside the Uralic scientific community to work with our data, cf. (Widmer, 2004). These peculiarities are not only potentially alienating to scholars outside of our discipline, they also give rise to difficulties when using software applications not specifically designed for Uralic languages, for example transcription software or programs used for linguistic annotation. For example, as many values in FUT lack Unicode code points of their own, they can only be represented using combining characters. The appropriate usage of these often poses an insurmountable hurdle to students not trained in appropriately dealing with such issues.

A further argument in favour of our project pertains to the ethical duties of linguistics – as researchers on social subjects, we need to ensure a good reciprocal relationship between us and our informants, cf. (Moran, 2016), and need to prioritise the communities' needs and rights, cf. (Austin, 2010). This means that we, as instructors, need to ensure that all of our students are familiar with best practices in handling language data and using available electronic resources: Many Uralic languages are endangered, have little electronic resources, and thus require more documentary research. Should our students aspire to conduct fieldwork, it is imperative that they know about technological methods in archiving and transcribing, see (Austin, 2006; Gippert, 2006; Dry, 2008). They might also be asked to contribute to revitalisation efforts, which nowadays also include technological contributions such as learning or dictionary applications, cf. (Grenoble and Whaley., 2006; Grenoble et al., 2008; Dauenhauer and Dauenhauer, 1998; Hinton et al., 2018).

3. Contents of the syllabus

While creating the syllabus for our course, we have followed the usual sequence of topics introduced in courses in computational linguistics or basic computer science. This includes overviews of "how computers think", data types and structures, character encoding and digital language data representation, basic concepts of programming, use of corpora, regular expressions, and comments on current topics in computational linguistics like machine translation, machine learning, or OCR. This discussion of current issues is essential for giving our students an idea of the work carried out in neighbouring disciplines: should they require IT assistance with one of their projects, or if they are working in interdisciplinary teams, they will need a basic understanding of the capabilities and limitations of certain computational tools in order to communicate their needs and to give informed comments on project work. It should also help them in planning their work to consider whether computational methods could reduce their workload or make their tasks easier – and to give them an idea of what can be expected of linguistics software, which issues it can help to solve and where its limits lie. If someone asks them why they have or have not used a particular method or software tool, they should be able to give a confident, reasoned answer and not have to say that they did not consider a methodology because they were afraid to use a computer or did not understand the results an application delivered. For the topics covered in the class, we plan to add practical work, not in active programming, but in working with code, to our course and teaching materials: supplying basic code segments and explaining how they work, and how they can be tweaked to deliver the desired results.

An example would be the creation of a frequency list. We would supply a script in a more-or-less human-readable code with comments and highlighted code fragments which have to be replaced in order to, for example, get the output in a desired alphabetical order (NB! The alphabetical order differs between different Uralic alphabets), or to read data from a specific file. All of these examples will come with a sandbox (or equivalent applications which are available online) as well as test files so that students can practise with dummy files before working on actual data for their own research tasks.

4. Linking with the curriculum

During our pilot course, most students were enrolled in later years of undergraduate or graduate programmes. While it is desirable that all students learn about methods in digital humanities and acquire some skills during their studies, it is questionable whether this task should be left for advanced students or rather tied into first-year modules. An important factor in favour of the second option is that an understanding of basics like Unicode can make a student's work easier and prevents them from using obsolete custom solutions, e.g. fonts to encode special characters, copy-pasting often incorrect symbols (e.g. Cyrillic and Latin characters that are visually identical, but not so for language processing tools). It can furthermore open their minds to computational approaches in our discipline and might lead them to consider introductory classes in computational linguistics, digital humanities, or computer science as their electives. It also reduces the likelihood of last-minute efforts to learn a software or a method for a term paper or dissertation. Computers become increasingly involved in our everyday work, and knowledge about how to use them correctly becomes more and more a requirement than an optional skill.

Most curricula in linguistics and philology contain at least one class on research methodology in an early semester (first and/or second). Such courses cover basics of literature search, citing techniques, note taking, or presentational skills – solid skills of information literacy (cf. `informationskompetenz.de`). But – moving into an increasingly digital age where everyday tasks like citations are assisted by software – why should information literacy be taught detached from computer or digital literacy skills? This does not mean that we aim to replace courses in information literacy – on the contrary, students should be able to do all the tasks their computer will do for them, to avoid over-reliance on technology. Therefore, we propose a system of mutual learning, where "paper and pencil" methods are mirrored in computational tasks.

5. Bringing it all together

As outlined above, we are aiming to foster awareness of computational methods among undergraduate students, ideally in their first year. First-year undergraduate students at the Institute of Finno-Ugric/Uralic Studies at LMU Munich must attend: a two-semester introduction to Finno-Ugric studies with a focus on history of the field, development of the languages, and basic typology; modules on phonetics and phonology; an introduction to linguistic theory; practical courses on scientific writing and information literacy; a language class. The possibilities to tie in computational approaches are numerous:

5.1. Introduction to Finno-Ugric studies

Within the introduction to Finno-Ugric studies, students learn about concepts of stem alternations – present in Finnic, most Saamic languages, and Nganasan – and vowel harmony – present in some shape or form in more Uralic languages than not, for example in Hungarian: the dative suffix has two variants used depending on the quality of the vowels of the base words, *-nak* after back vowels (*barát* 'friend' → *barátnak* 'to the friend') and *-nek* after front vowels (*zseb* 'pocket' → *zsebnek* 'to the pocket'). The resulting allomorphy can be used to explain search functions using regular expressions (e.g. implementing a search function that will find both variants of the aforementioned dative suffix) and highlight how computer applications have to be constructed to allow for this allomorphy (e.g. in lemmatising/stemming, morphological analysis).

5.2. Phonetics and Phonology

An integral part of the phonetics and phonology modules is acquiring practical skills in transcribing spoken language into a conventional transcription system, either IPA, FUT, or a standard orthography of the language. While there are numerous possible excursuses from reading a spectrogram and working with applications like Praat to speech recognition and synthesis, these topics are mostly covered by entire modules at the Institute of Phonetics and Speech Processing where our students can take classes as electives. However, in the compulsory class at our institute, students practise writing and reading phonetic transcriptions, and become acquainted with the Cyrillic alphabet used by the majority of Uralic languages spoken in the Russian Federation (e.g. Mari, Udmurt). This enables us to address three issues: firstly, students will have to submit coursework throughout the semester, and handing it in electronically or in type-written form will require the use of special characters on a computer. They should know about Unicode and encodings, as well as tools for writing and transcribing with Unicode characters (e.g. `transcribe.mari-language.com`). Secondly, they will want to use the right characters in writing, so teaching them how to write Cyrillic characters on their keyboards becomes relevant (in addition to special characters found in Uralic languages utilizing the Latin alphabet, e.g. Hungarian <ű>, <ő>). Thirdly, we can present tools for transliteration between orthographies and explain how these simple search-and-replace processes are coded.

5.3. Information Literacy and Scientific Writing

The module on information literacy and scientific writing can be amended by a brief overview of linguistics software, or how to make best use of basic tools like using the regular expression function in Word. Furthermore, information literacy should also include a discussion about research ethics, stressing the importance of using best practice in scientific work. In fact, the information literacy course is most easily converted into an information and digital literacy course and should be seen as a platform for teaching the desired skills mentioned above, and also introducing students to more modern methods of finding and organizing references, e.g. Google Scholar and bibliographic software.

5.4. Language Classes

Lastly, students taking a language class will face the challenge of writing in their target language using all special characters of its orthography (see above). Moreover, they will want or need to use electronic resources like dictionaries, spell-checkers, morphological synthesis and analysis, or even corpora. We feel that we should do more than hand them a list of available resources and referring them to documentation on these but should rather give our students ideas of how to utilise the digital resources appropriately and how to get the desired results out of them. This will also help them with other course work in the following years.

5.5. Summary

As could be seen from the discussion above, the first-year modules in our undergraduate programme already give a basis for weaving in basic IT skills through highlighting the computational side of the (paper and pencil) processes or techniques presented. Within the course of the first year, this approach would cover Unicode, special characters and encodings, transcription and transliteration, regular expressions, morphological analysis and synthesis, electronic resources, and excursuses and discussions of issues in computational linguistics. This covers all the basics which would be covered in the initial chapters of any computational linguistics or computer science textbook. Students could then go on to continue along this pathway by taking classes of computational linguistics, or by using these methods for their own research. Either way, they will have acquired important knowledge about using computers efficiently and can use this knowledge throughout their studies.

6. Evaluation of the pilot

We taught our course “Digital Resources in Linguistics” as a pilot project in May, June, and July 2017 as a seminar, open to undergraduate and graduate students of Uralic studies and neighbouring disciplines as an elective course. Our course consisted of six four-hour sessions with a limited workload worth 3 ECTS points. Six students enrolled for our course. As is common practice in universities, our students were asked for their evaluation of the module at the end of the term. Given the experimental nature of our course, we requested some more detailed information on students’ evaluation of their learning progress, and obtained permission to reproduce their answers in print. This feedback showed that the students felt that they had learned much (“I now feel more confident in handling the more technical side of my research”) and found the workload appropriate. While such an evaluation cannot guarantee the success of an approach, it demonstrates that our idea gets positive feedback (“would greatly recommend”) and that this course, which was held as an optional module, managed to provide interesting insights. The greatest point of criticism raised by students is that it was not offered at the optimal point in their studies: “The only regret I have about the course is that it came too late in my studies, namely in my last semester of my Master’s. Would it have taken place in my second or third semester of my Bachelor’s, I may have considered taking more courses in that direction.” As there are no elective courses in the first two years of our Bachelor’s programme, however, we currently have no possibility of offering this course at an earlier stage of studies. Changes to the curriculum would be necessary for this to happen in future.

Students’ interest also showed in their independent projects which they had to conduct for receiving the credits on this module: brief case studies where computational methods are currently used or could be used in the future to solve problems in their field of interest including a discussion of potential benefits and pitfalls (“As a conversational analyst, I still very much rely on the ‘pen and paper’-method of data analysis; however, I now have a better understanding of how to work with the raw data in the transcription phase").

Students enjoyed learning about computer basics, discussions of software use, and thematic excursuses. These topics inspired conversations with students and prompted individual questions (a student working part-time as a proofreader: "This gain in cross-disciplinary work also proved to be useful in working life"). Teaching principles of programming and working with scripts turned out to be more difficult for our students and might require reworking of our teaching materials. Furthermore, we see the necessity to address commonly used applications (e.g. ELAN, R, LaTeX) more directly, as it appeared that our students had already heard about such software before our class (in some cases, competencies with these software packages were expected from them) but without receiving instruction in their usage and range of capabilities. There were no issues with the thematic progression and the order of topics.

7. Outlook

We hope to teach this course again in future, as budget and student interest allows, both in Munich and at our partner institutions across Europe. For future iterations of this course we are creating a script and reading list to give the course a more solid outline. Eventually, we intend to create a textbook (to be published online in an openly accessible manner) and supplementary online materials on the basis of our course materials.

Our efforts so far have been happening within the framework of the Erasmus+-strategic partnership INFUSE (Integrating Finno-Ugric Studies in Europe, 2015–2018, cf. `www.infuse.finnougristik.uni-muenchen.de`), which is administered by the Institute of Finno-Ugric/Uralic Studies at LMU Munich, and consists of eight European Finno-Ugric departments (Hamburg, Helsinki, Munich, Szeged, Tartu, Turku, Uppsala, Vienna). Funding has been guaranteed for a continuation of this strategic partnership, COPIUS (Community of Practice in Uralic Studies, 2018–2021, cf. `www.finnougristik.uni-muenchen.de/aktuelles/nachrichten/copius`; Budapest has now joined our consortium). In COPIUS, our focus will lie on the development and pooling of teaching materials. We have committed ourselves to creating an openly accessible integrated online learning platform for our subject field, including a general introduction to Finno-Ugric studies, and a number of expansion modules (e.g. on fieldwork methods, etymology, individual Finno-Ugric languages). Our teaching materials can constitute an additional module in this learning platform.

We would like to reiterate that this does not mean that we are trying to replace traditional computer science or computational linguistics courses. We are rather aiming to close the gap between students with prior knowledge of computer science and students without exposure to principles of computing. This enables the latter to make best use of available tools and resources and gives all students guidelines for basic NLP tasks pertaining to Uralic languages, while simultaneously presenting an apparatus of computational methodologies.

We hope that our contribution will help students to see the "mysticism" of computers and computational methods in a different light, and thereby help to bridge the digital divide.

8. References

Peter K. Austin. 2006. Data and language documentation. In Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, editors, *Essentials of Language Documentation*, pages 87–112. de Gruyter, Berlin.

Peter K. Austin. 2010. Communities, ethics and rights in language documentation. *Language Documentation and Description*, 7:34–54.

Jeremy Bradley. 2017. Transcribe.mari-language.com: Automatic transcriptions and transliterations for Mari, Tatar, Russian, and more. *Acta Linguistica Academica*, 64(3):369–382.

Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne Jekat, Hagen Langer, and Ralf Klabunde, editors. 2009. *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Spektrum, Heidelberg, third edition.

Nora Marks Dauenhauer and Richard Dauenhauer. 1998. Technical, emotional, and ideological issues in reversing language shift: examples from southeast alaska. In Lenore A. Grenoble and Lindsay J. Whaley, editors, *Endangered languages: Language loss and community response*, pages 57–98. Cambridge University Press, Cambridge.

Markus Dickinson, Chris Brew, and Detmar Meurers. 2012. *Language and Computers*. John Wiley & Sons, Chichester.

Helen Aristar Dry. 2008. Preserving digital language materials: Some considerations for community initiatives. In Wayne Harbert, Sally McConnell-Ginet, Amanda Miller, and John Whitman, editors, *Language and Poverty*, pages 202–222. Channel View Publications, Bristol.

Robert L. Furman. 2015. *Technology, Reading, and Digital Literacy. Strategies to Engage the Reluctant Reader*. International Society for Technology in Education, Eugene, OR.

Jost Gippert. 2006. Linguistic documentation and the encoding of textual materials. In Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, editors, *Essentials of Language Documentation*, pages 337–361. de Gruyter, Berlin.

Lenore A. Grenoble and Lindsay J. Whaley. 2006. *Saving languages: an introduction to language revitalization*. Cambridge University Press, Cambridge.

Lenore A. Grenoble, Keren D. Rice, and Norvin Richards. 2008. The role of the linguist in language maintenance and revitalization: Documentation, training and materials development. In Wayne Harbert, Sally McConnell-Ginet, Amanda Miller, and John Whitman, editors, *Language and Poverty*, pages 183–201. Channel View Publications, Bristol.

Leanne Hinton, Leena Huss, and Gerald Roche, editors. 2018. *The Routledge Handbook of Language Revitalization*. Routledge, New York - London.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language* *processing, computational linguistics, and speech recognition*. Pearson Education International, Prentice Hall, Upper Saddle River, NJ.

Richard Kern. 2014. Technology as Pharmakon: The Promise and Perils of the Internet for Foreign Language Education. *The Modern Language Journal*, 98(1):340–357.

Mary H. Moran. 2016. The digital divide revisited: Local and global manifestations. In Roger Sanjek and Susan W. Tratner, editors, *eFieldnotes: The Makings of Anthropology in the Digital World*, pages 65–77. University of Pennsylvania Press, Philadelphia.

Bogdan Popoveniuc. 2010. What is a technological mentality? In Viorel Guliciuc and Emilia Guliciuc, editors, *Philosophy of Engineering and Artifact in the Digital Age*, pages 125–135. Cambridge Scholars Publishing, Cambridge.

Colin T. A. Schmidt. 2010. Cognitive life re-engineered. In Viorel Guliciuc and Emilia Guliciuc, editors, *Philosophy of Engineering and Artifact in the Digital Age*, pages 67–80. Cambridge Scholars Publishing, Cambridge.

Eemil Nestor Setälä. 1901. Über die transskription der finnisch-ugrischen sprachen. *Finnisch-ugrische Forschungen*, 1:15–52.

Faithe Wempen. 2014. *Computing Fundamentals: Digital Literacy Edition*. John Wiley & Sons, Chichester.

Anna Widmer. 2004. Reconnecting and Reconsidering: Remarks on the Final Discussion of the International Linguistic Symposium “Reconnecting Finnic”. *Linguistica Uralica*, 2004(3):197–212.

Teaching women writers with NEWW Virtual Research Environment

Katja Mihurko Poniž,[1] Narvika Bovcon,[2] Marie Nedregotten Sørbø,[3] Viola Parente-Čapková,[4] Amelia Sanz,[5] Suzan van Dijk,[6] Aleš Vaupotič[7]

[1]School of Humanities, University of Nova Gorica, Vipavska 13, SI 5000 Nova Gorica
katja.mihurko.poniz@ung.si

[2]Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana
narvika.bovcon@fri-uni.lj.si

[3]Volda University College, Post box 500. NO-6101 Volda
mns@hivolda.no

[4]Department of Finnish Literature, University of Turku
FI-20014 University of Turku
viocap@utu.fi

[5]Faculty of Philology, Complutense University, 28040 Madrid
amsanz@filol.ucm.es

[6] Huygens ING KNAW, Postbus 10855, 1001 EW Amsterdam
suzan.van.dijk@huygens.knaw.nl

[7]School of Humanities, University of Nova Gorica, Vipavska 13, SI 5000 Nova Gorica
ales.vaupotic@ung.si

1.Introduction

The underrepresentation of women in cultural historiography has challenged a number of feminist responses in the form of supplementary female canons since the 1970s. The DARIAH Working Group *Women Writers in History* (https://www.dariah.eu/activities/working-groups/women-writers-in-history/) takes this task a step further, and investigates historical sources until 1930 to find out whether female authors were read in the past. The objective of the DARIAH Working Group WWIH is: to carry out research about female authorship in history, the international reception of women's writing and the connections between women authors. Evidence of readership, translations and commentary is contained in the digital repository NEWW VRE (Virtual Research Environment) http://resources.huygens.knaw.nl/womenwriters, which serves as a collaborative research tool for the above mentioned working group.

This tool aims to facilitate research about women's authorship in Europe from the Middle ages until the early 20th century. Data are entered when any proof of reception (comments in press, private letters, translations, adaptations etc.) is found: in other words, when it becomes clear that a woman writer and/or her works were received and read. Although the main focus lies on European women authors, the NEWW VRE also includes information on works and reception created in European colonies, Canada and the United States, due to the mutual cultural exchanges between these regions and Europe. NEWW VRE provides information which is not always easy to find elsewhere; in particular it is presented in a structured way and within a meaningful context. The larger part of the information is open for consultation by everybody: data about women authors, their works and the reception of these works are accessible without a password. The systematic scrutiny of reception data from large-scale sources (library and booksellers' catalogues, the periodical press) forms the basis for the study of women's participation in this process. This includes hyperlinks to online biographies, texts and testimonies of reception (for instance, in the periodical press).

The present version of the NEWW VRE was (and is being) developed by the IT department of the Huygens Institute for the History of the Netherlands thanks to the European HERA project *Travelling TexTs 1790-1914* (2013-2016) http://travellingtexts.huygens.knaw.nl/. The HERA project had been prepared by a series of earlier digitizing projects (NWO, SURF), projects in international networking (NWO, COST), and in IT development (CLARIN-NL). Within the COST-collaboration (Action IS0901 *Women Writers in History*) several of the participants succeeded also in having connected projects funded at the national level.

It is the ultimate goal of the NEWW VRE to contribute to the inclusion of women in European literary historiography in order to do justice to the roles they played in their own time: to understand that they actually had "something to say" – which was recognized by contemporary readers.

2. About the project Teaching Women Writers: exploring the possibilities of VRE Women Writers

Teaching covers a wide range of activities, from using outcomes of the research to contributing to it. We believe that indeed the students could be given a more important role than the rather passive one they are allowed to have now. Besides, given the important number (over 6000) of European women authors before the early 20th c. discovered up to now and accumulated in our NEWW VRE, we have to enlarge the group of active collaborators. And most importantly, it now seems possible to have the collaboration of students – in the classroom or as trainees – and also seniors as "citizen scientists". They can much more than before, participate in scholarly enterprises like ours, considering that so much of textual material (including those by and about women authors; including also the periodical press) is now available online.

In this way, the research tool NEWW VRE can also become important within the process of teaching women's presence in the literary field – at different levels: MA students can study individual cases (within the large context provided by NEWW); Bachelor students can become familiar with women's too little-known contribution to literature, and even grammar school students can find information and access the writings.

A number of the WWIH-members are interested in exploring the usability of the tool to enhance, more generally, students' digital competences.

The discussions and training sessions held at the international conference and workshop *Teaching Women Writers – exploring "NEWW Virtual Research Environment" possibilities* (Ljubljana, 16-18 November 2017) have been continued by individual work in the VRE at home, with checking and support given by a qualified assistant hired especially for this and funded by DARIAH. Such training and support has contributed to easiness of data entry, to the quality of data, and to further development of the online tool.

3. References

Stephen Brier. Where's the Pedagogy? The Role of Teaching and Learning in the Digital Humanities http://dhdebates.gc.cuny.edu/debates/text/8, 12 April 2018.

DH teaching material. open-source, high quality, multilingual teaching materials for the digital arts and humanities. https://teach.dariah.eu/, 12 April 2018.

Katja Mihurko Poniž (ed.). 2017. *Teaching women writers: exploring "NEWW virtual research environment" possibilities: [programme and abstracts]: international conference and workshop, Ljubljana, 16th and 17th November 2017*. Nova Gorica: School of humanities, 2017.

Aleš Vaupotič, Narvika, Bovcon. 2017, Visualization of the WomenWriters database: interdisciplinary collaboration experiments 2012-2015. In: Mihurko Poniž, Katja (ed.). *Reception of foreign women writers in the Slovenian literary system of the long 19th century*. Nova Gorica: University of Nova Gorica Press.

Odnosi do jezika v slovenski, hrvaški in srbski računalniško posredovani komunikaciji

Damjan Popič,* Darja Fišer*†

* Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
damjan.popic@ff.uni-lj.si
† Odsek za tehnologije znanja, Institut Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
darja.fiser@ff.uni-lj.si

1. Uvod

V prispevku predstavimo odnose do jezika v slovenski, hrvaški in srbski računalniško posredovani komunikaciji (RPK). To izvedemo tako, da za vsak jezik na podlagi izbranih ključnih besed izoliramo tvite, vezane na jezikovno tematiko, in v njih poskušamo določiti prevladujoči odnos avtorjev. Tovrstna primerjava se nam zdi zanimiva zaradi sociolingvističnih in v veliki meri tudi političnih razmer po razpadu nekdanje Jugoslavije, med katerimi je prišlo do razhajanja standardov in poudarjanja nacionalnih jezikov (gl. Požgaj Hadži in Balažic Bulc, 2015). Raziskava temelji na Popič in Fišer (2017), kjer smo kategorizirali odnose do jezika in jezikovnih vprašanj v slovenski RPK, ki jo razširjamo na hrvaški in srbski prostor. S korpusno podprto sociolingvistično analizo ključne besede *pravopis* tako preverimo, v kolikšni meri se restandardizacija nacionalnih jezikov Hrvaške in Srbije odraža v RPK ter ali so odnosi do pravopisa primerljivi s tistimi v slovenskem okolju.

2. Sociolingvistična izhodišča

V pričujočem razdelku predstavimo sociolingvistična izhodišča za raziskavo odnosov v slovenskem, hrvaškem in srbskem jezikovnem okolju, na katerih temlji izbor ključnih besed za analizo. Čeprav gre za podobna jezikovna okolja, za analizo ne moremo uporabiti povsem prekrivnega nabora ključnih besed, saj so nekatere specifične samo za eno od okolij (npr. vejica za slovensko in cirilica za srbsko govorno okolje). V nadaljevanju izpostavimo nekatere specifike posameznih okolij, na podlagi katerih smo izbirali ključne besede za izvedbo analize.

2.1. Slovensko govorno okolje

Kot ugotavljamo v prispevku Popič in Fišer (2017), je slovensko jezikovno okolje izrazito direktivno naravnano (gl. Škiljan, 1999), jezikovna raba oz. obvladovanje jezikovnega standarda pa je v veliki meri nosilka družbenega prestiža. Zaradi prestižne in normativne narave jezikovne regulacije, v kombinaciji s pestro zgodovino poskusov slabljenja slovenskega jezika s strani močnejših jezikov, zlasti nemščine in srbohrvaščine, je jezik postal temeljni slovenski simbol razlikovalnosti od drugih narodov (gl. Popič, 2014), argument jezikovnega (ne)znanja pa je postal prestižni argument moči, kar pomeni, da se domnevno nepoznavanje jezika pri nekom zelo pogosto izrablja kot argument tipa *ad hominem*.

Seveda je določena mera družbene stratifikacije, ki jo nosi jezikovna raba, v jeziku povsem inherentna (gl. npr. Labov 1966), vendarle pa je slovensko jezikovno okolje specifično v tem oziru, da je slovenski pogled na jezik oblikoval prav odklonilni odnos do tujega, ki ga je pojil predvsem strah pred agresorji, to pa je tudi eden glavnih razlogov za poudarjanje jezikovne pravilnosti in *dobrega* jezika. Kljub izrazito ugodnemu trenutnemu položaju slovenščine kot evropskega jezika pa tovrstni strah še vedno obstaja v obliki purističnih teženj znatnega dela slovenske normativistike kot tudi nestrokovne javnosti (Popič in Fišer, 2017).

V predhodnih raziskavah tako ugotavljamo, da je odnos slovenskih uporabnikov RPK izrazito normativno usmerjen in da se argument jezikovnega (ne)znanja pogosto uporablja kot argument moči v javni diskusiji (Popič in Fišer, 2017). Te odnose v končnem prispevku primerjamo s tistimi v hrvaškem in srbskem kulturnem okolju.

2.2. Hrvaško govorno okolje

Kot izpostavljata Požgaj Hadži in Balažic Bulc (2015: 71), je »[m]ed jeziki, nastalimi na novoštokavski narečni osnovi, […] v 90. letih največ sprememb doživel hrvaški standardni jezik«. Pri tem je bistvenega pomena to, da je imel nabor izbrane leksike na Hrvaškem velik simbolni naboj, hkrati pa je utrjeval hrvaško nacionalno identiteto (Kordić, 2010). »Tako so bili govorci, ki so uporabljali ustrezno leksiko, označeni kot domoljubi, govorci, ki te leksike niso uporabljali, pa nenadoma niso govorili »čistega« hrvaškega jezika in so zato dobili različne »politične etikete« (ibid.). Zato so se govorci v svojem lastnem jeziku počutili nelagodno in se jim je celo zatikalo ob iskanju »prave« besede (Badurina, 2015: 69).

Obenem se je »[h]rvaška jezikovna politika v 90. letih lotila tudi tujih besed, še posebej srbizmov« (Požgaj Hadži in Balažic Bulc, 2015: 73), vse to pa so podpirale vladajoče politične elite (Požgaj Hadži in Balažic Bulc, 2015: 72), »[z]ato je to obdobje idealno za objavljanje razlikovalnih slovarjev hrvaškega in srbskega jezika, slogovnih priročnikov in slovarjev odvečnih besed v hrvaščini, ki vneto opisujejo do tedaj 'prepovedane' leksikalne razlike med hrvaščino in srbščino (Požgaj Hadži in Balažic Bulc, 2015: 73). Na podlagi sociolingvističnih razlik se zato pri analizi odnosov do hrvaškega jezika osredotočamo predvsem na jezik, tujke, slovarje, slovnico in pravopis.

2.3. Srbsko govorno okolje

Za razliko od številnih sprememb, ki jih je doživel hrvaški jezik, se srbski jezikovni standard na zunaj v politično napetih 90. letih ni spremenil (Požgaj Hadži in Balažic Bulc 2015: 75): »Za razliko od Hrvaške, kjer so vladajoče politične elite nacionalizirale jezik s pomočjo različnih institucij (akademije, inštituta za hrvaški jezik, sveta za hrvaški jezik ipd.), so za srbski jezik in srbsko nacijo v Srbiji 'skrbele' neformalne skupine. Vendar daleč od tega, da se v srbščini ni dogajala nacionalizacija jezika, gre le za drugačno vrsto nacionalizma.« Po besedah Bugarskega (2012: 52) gre v Srbiji za »reduktivni tip nacionalizma, ki želi predvsem zagraditi in zavarovati lastno ozemlje«.

Pri srbski jezikovni situaciji nas bo med analizo zanimala predvsem diada cirilica : latinica, saj je v primerjavi s hrvaškim in slovenskim okoljem povsem specifična. Po tem, ko je v drugi polovici 20. stoletja latinica skorajda izpodrinila cirilico, so mnogi to doživeli kot napad na srbsko samobitnost. Odgovor na tovrstne napade je bila sprememba 10. člena srbske ustave (leta 2006), s katerim je država latinici odvzela status alternativne pisave, »kar pa ni bilo v skladu z realnostjo« (Požgaj Hadži in Balažic Bulc, 2015: 77).[1]

3. Analiza odnosov do jezika v RPK

3.1. Gradivo

Za raziskavo odnosov do jezika v RPK v obravnavanih jezikovnih skupnostih uporabimo izolirane tvite iz treh različnih korpusov, in sicer iz korpusa Janes (Erjavec et al., 2018) za slovenščino, iz 25-milijonskega korpusa hrvaških tvitov Tweet-hr[2] za hrvaški jezik ter iz 205-milijonskega korpusa Tweet-sr[3] za srbski jezik. Za vsak jezik izberemo relevanten nabor ključnih samostalnikov, za katere obstaja velika verjetnost, da so uporabljene v tvitih, ki izražajo odnos do jezika. Ključne besede, ki se med jeziki delno prekrivajo (npr. *pravopis*, *slovnica*/*gramatika*, *jezik*), delno pa so o prilagojene kulturnim specifikam obravnavanih področij (npr. *latinica*, *ćirilica*, *tuđica*), so podane v Tabeli 1.

Slovenščina		Hrvaščina		Srbščina	
Ključna beseda	Frekvenca	Ključna beseda	Frekvenca	Ključna beseda	Frekvenca
jezik	38.234 (142,40/mio)	gramatika	139 (5,40/mio)	ćirilica	3.967 (19,28/mio)
pravopis	1.465 (5,50/mio)	hrvatski	19.668 (763,50/mio)	gramatika	2.310 (11,23/mio)
slovenščina	12.041 (44,90/mio)	jezik	3,581 (139.01/mio)	jezik	28.182 (136,90/mio)
slovnica	2.097 (7,81/mio)	pravopis	209 (8,11/mio)	latinica	1.473 (7,16/mio)
vejica	4.003 (14,91/mio)	riječnik/rječnik	196 (7,61/mio)	pravopis	2.350 (11,42/mio)
/	/	tuđica	8 (0,31/mio)	srpski	74.525 (362,10/mio)

Tabela 1: Ključne besede za slovenščino, hrvaščino in srbščino.

Ob zavedanju, da so nekatere ključne besede lahko uporabljene tudi v pomenih in kontekstih, ki za pričujočo raziskavo niso relevantni, ni presenetljivo, da sta izraza *hrvaški* in *srbski* toliko pogostejša od vseh ostalih

[1] Zelo indikativne so tudi nedavne pobude za kaznovanje rabe latinice v Srbiji, ki naj bi ob uvedbi sveta za srbski jezik uvedla tudi kazni za uporabo latinice v uradnih kontekstih, kazni pa naj bi znašale do milijon srbskih dinarjev

[2] https://www.clarin.si/noske/run.cgi/corp_info?corpname=tweet_hr&struct_attr_stats=1&subcorpora=1

[3] https://www.clarin.si/noske/run.cgi/corp_info?corpname=tweet_sr&struct_attr_stats=1&subcorpora=1

ključnih besed na seznamu. Izstopa pa dejstvo, da je relativna frekvenca za ključno besedo *jezik* v vseh treh jezikih primerljiva. Glede na relativno frekvenco je *pravopis* najpogostejši v srbščini, sledi mu hrvaščina, šele nato pa slovenščina, v kateri se pojavlja dvakrat redkeje kot v srbščini. Tudi *slovnica* je najpogostejša v srbščini, sledi ji slovenščina, na zadnjem mestu pa je hrvaščina, v kateri je dvakrat redkejša kot v srbščini. V srbščini po pogostosti izrazito izstopa *ćirilica*, ki je še pogostejša od slovenske vejice. Hrvaške *tuđice* pa zaradi izjemno nizke frekvence v nadaljnje analize nismo vključili.

3.2. Razvrščanje

Tvite v vseh treh jezikih, vključenih v raziskavo, ki vsebujejo izbrane ključne besede, nato razvrstimo v kategorije glede na odnos, ki ga do jezika avtor tvita izkazuje. Za to uporabimo tipologijo in smernice za označevanje, razvite za slovenščino (Popič in Fišer, 2017). Za to različico prispevka smo za vse tri jezike za vsako kategorijo označili pojavitve ključne besede *pravopis*, ki je v vseh obravnavanih jezikih enaka, s čemer smo želeli preveriti ustreznost tipologije tudi za druge jezike in dobiti prvi vpogled v podobnosti in razlike v odnosu govorcev do pravopisa. V nadaljnjih analizah bomo z njo označili po 100 naključno vzorčenih tvitov za vsako ključno besedo za vse tri jezike. Na podlagi teh oznak bomo nato opravili primerjalno kvantitativno analizo odnosa do jezika v slovenskem, hrvaškem in srbskem prostoru.

Odnos	Slovenščina	Hrvaščina	Srbščina
vprašalni	Veste, da se po Slovenskem pravopisu imena praznikov datumov, razen tistih, ki so izpeljana iz priimkov, pišejo z malo začetnico?	Pazite li na pravopis i gramatiku na društvenim mrežama?	Da li ima neko knjigu pravopisa, priručnik neki?
informativni	Nekaj napotkov je menda v slovenskem pravopisu .	Relevantni pravopis je online i jednostavan je za upotrebu http://t.co/oFvpzNyDci pa nema više izlike za nepismenost:)	Đ se po novom pravopisu, u zvaničnoj upotrebi, piše Dj.
pritoževalni	to tudi meni živec potegne, ker ne poznajo niti osnov pravopisa , pa nastanejo taka skropucala, da je groza	Meni hrvatski pravopis je sve gori. Užas	O jebem ti državu kad se pravopis promeni minimum 4 puta samo za mog školovanja
šaljivi	Pravopis je zarota levosučnih akademikov.	Inace ne volim viceve, ali novom pravopisu u Hrvata, kao laik, povremeno moras priznati komicni momenat.	Ana, mislim da će da me ubace u pravopis neki ako nastavim ovako
dismisivni	Siromak si. V pameti in znanju pravopisa.	Ne znam prema kojim to kriterijima se danas zapošljavaju novinari koji fejlaju već na pravopisu .	ŠUPČINOO Pre svega, ne znaš pravopis.
defenzivni	Aja, a zdej smo pa pri pravopisu. A je argumentov zmanjkal.	Ne vjeruj ženi s lošim pravopisom .	Kad baba pocne da ti kenja o srpskom pravopisu i nacinu izrazavanja danasnjih tinejdzera...
opravičujoči	Sram me je za " moj" pravopis ...	nemojte mi o pravopisu na rano jutro, nisam nepismena samo lijena	Zbog tvitera cu iz pravopisa da imam keca
idiomatični	Kdor se še nikoli ni zatipkal, naj vrže pravopis vame ☺	" Pravopis sa zvijezdama " bih gledao.	Godine prolaze, greške u pravopisu ostaji.
nacionalistični	Če se imate za velikega Slovenca, se najprej naučite pravopisa .	Spasimo spomenik Ljudevitu Gaju jer je bio ustaša i kao ustaša stvorio je ustaški pravopis 1830	Ako si hejter i pljuvač nauči barem svoj maternji jezik i pravopis .
puristični	Pravopis deklica, pravopis :) ;)	Zašto nitko ne provjeri pravopis prije tiska? Sramočenje	Kažu da je pravopis na tviteru nebitan. Ja baš izbegavam pravopisne greške. Mora biti savršeno napisan svaki tvit.

Tabela 2: Odnosi do jezika v slovenščini, hrvaščini in srbščini na primeru pravopisa.

Za vse tri jezike je značilen pogost dismisiven in nacionalističen ton, ki opravljanje vidnejših funkcij (npr. v politiki) pogojuje s poznavanjem pravopisa, prav tako pa diskreditira strokovnost vseh, ki pravopisa ne upoštevajo. V vseh treh jezikih je prav tako precej razširjeno opravičevanje za pravopisno neustrezne tvite, v slovenskem korpusu pa je bistveno več vprašanj o pravilni rabi jezika. V slovenskem in hrvaškem korpusu naletimo na račune, katerih osnovni namen je tvitanje o pravopisnih vprašanjih, medtem ko v srbskem korpusu nanje nismo naleteli. Zanimivo je, da sta bila glede na zadetke tako v srbskem kot hrvaškem korpusu v istem obdobju objavljena nova pravopisa za ta dva jezika, saj se številni tviti nanašajo na izid novega pravopisa, ki je večinoma sprejet z neodobravanjem. V obeh korpusih je pogosto pritoževanje o prevelikem številu obstoječih pravopisov, na kar v slovenskem korpusu nismo naleteli. V hrvaškem korpusu je v povezavi s pravopisom še najmanj nacionalističnih izjav, prisotno pa je – v našem naboru raziskovanih okolij – izrazito singularno dojemanje pravopisa kot jeziko(slov)nega dela, ne zgolj kot nabora občeveljavnih pravil v okviru jezikovnega predpisa, kar je značilno za slovensko in srbsko jezikovno okolje.

4. Sklep

V prispevku smo predstavili zasnovo primerjalne raziskave odnosov do jezika v treh sorodnih jezikovnih okoljih, kot se izkazuje v slovenskih, hrvaških in srbskih tvitih. Z njo bomo v nadaljnjih raziskavah ugotavljali, v kolikšni meri so odnosi do jezika v teh skupnostih prekrivni in v katerih točkah se razlikujejo, obenem pa bomo poskušali za morebitna razhajanja poiskati v sociolingvističnih okvirih divergentnega razvoja vseh treh jezikov v zadnjih dveh desetletjih in pol.

Literatura

Lara Badurina. 2015. Standardizacija ili restandardizacija hrvatskoga jezika u 90-im godinama 20. stoljeća. V: T. Pišković, T. Vuković (ur.): *Jezične, kulturne i književne politike*, str. 57–79, Zagreb: Zagrebačka slavistička škola.

Ranko Bugarski. 2012. *Portret jednog jezika*. Beograd, Biblioteka XX vek.

Tomaž Erjavec, Nikola Ljubešić in Darja Fišer. 2018. Korpus slovenskih spletnih uporabniških vsebin Janes. V: D. Fišer, ur., *Viri, orodja in metode za analizo spletne slovenščine*, str. 16–43. Ljubljana, Znanstvena založba Filozofske fakultete. Ljubljana, Znanstvena založba Filozofske fakultete.

Snežana Kordić. 2010. *Jezik i nacionalizam*. Zagreb, Durieux.

William Labov. 1966. *The Social Stratification of English in New York City*. Washington, Center for Applied Linguistics.

Damjan Popič. 2014. *Korpusnojezikoslovna analiza vplivov na slovenska prevodna besedila*. Doktorsko delo, Univerza v Ljubljani.

Damjan Popič in Darja Fišer. 2017. Fear and loathing on Twitter : attitudes towards language. V: STEMLE, Egon W. (ur.), WIGHAM, Ciara R. (ur.). *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)*, str. 61–64, Bolzano.

Vesna Požgaj Hadži in Tatjana Balažic Bulc. 2015. (Re)standardizacija v primežu nacionalne identitete: primer hrvaškega, srbskega, bosanskega in črnogorskega jezika. *Slovenščina 2.0*, 3(2):67–94.

Online database in Research of Correspondence of Franjo Ksaver Kuhač (1834-1911)

Sara Ries

Croatian Academy of Sciences and Arts
mailto:
sararies5@gmail.com

1. Introduction

As archival documentation is an objective source for research, the preserved correspondence of an author is a valuable source for interpretation. Franjo Ksaver Kuhač, who was, due to his tireless work which greatly contributed to Croatian music historiography, recognised as the first Croatian musicologist, left a comprehensive collection of his letters. The correspondence is collected in thirteen books, so-called *Briefcopirbücher*, as Kuhač named them, and contain the total of 3 thousand letters.[1] The letters are written by Kuhač himself and are mainly copies and concepts of his sent letters. He kept those letters as a kind of a memo of what he had written to someone. Unfortunately, the letters addressed to him are not preserved. As the majority of the letters are written in Gothic script, also known as *Kurrentschrift*, in German language, the research of the correspondence includes transliteration and translation of the letters.

The correspondence covers the time span from 1860 to 1911, the period of the growing awareness of national culture and heritage in order to promote new national identity between the neo-absolutist period and the First WW. Those letters are an important source of information about cultural, political and musical events and they provide insights about Kuhač's life, work and numerous activities. They are also evidence on the period marked by important political and cultural changes, as well as Kuhač's strivings to collect financial and moral supports for his researches and endeavours in collecting South Slavic folk-songs.

Furthermore, one can acknowledge not only information about Kuhač's attitudes, ideas and relations between him and his colleagues, but also his views and attitudes about his work and musical, national and cultural affairs in Croatia and Europe. He persistently wrote letters to many eminent politicians and public persons, as well as his colleagues not only in Croatia but in Hungary, Austria, Germany, Czechia, Italy, Serbia, Montenegro, Bosnia, Slovenia, Bulgaria, France, etc. with aim to present himself, his ideas regarding national music, music theory and music historiography, his work, and to gain new information on musical culture, as well as financial and moral support for publishing his opus and, if possible, to further educate himself.

2. Goal of the paper

Kuhač's relations with composers, musicographers and melographers from the central and south-eastern Europe will be investigated through his correspondence. A comprehensive research on Kuhač's contacts will result with insights which will offer a deeper understanding of the relations between political and cultural circumstances not only in Croatia, but also in the Austrian, i.e. Austro-Hungarian Empire. The research is a part of a nationally financed project Networking through Music: Changes of Paradigms in the "Long 19th Century" – From Luka Sorkočević to Franjo Ks. Kuhač. One of the planned results of the project is the creation of an on-line accessible database, which will provide basic information about the documents, terms and contain all accessible biographical data of the addressees and persons mentioned in

[1] See: Marija Janaček Buljan. 1984. Kuhačeva korespondencija. In: *Zbornik radova sa znanstvenog skupa održanog u povodu rođenja Franje Ksavera Kuhača (1834-1911)* Zagreb, 20.-21. studenoga 1984, pages 463-472. JAZU, Zagreb.

the letters.[1] It is a relational type of database with a MySQL management system. The database is a driven application with PHP. The said technologies are chosen because the data structure is suitable for this type of database – the data is organized in tables. In addition, only the relational data model is used, excluding other metadata or other standards. The database has several searchable topics: Entries, Works, Notions, Persons, Sources. One can search by the certain date or range of dates, terms, compositions or other kind of works of art (librettos or texts), persons, addresses (places), key words (such as employment, biographical data or title) within three thematic areas: 1) the diary of the Dubrovnik nobleman, diplomat and composer Luka Sorkočević (1734-1789) documenting his sojourn in Vienna; 2) performances of foreign itinerant opera companies and soloists in Zagreb in the mid-19th century; 3) Franjo Ksaver Kuhač's correspondence. The list as well as all the biographical data of all addressees mentioned in the second and the third book of the correspondence will be of a great significance in preparing critical editions (transliteration in Latin script, translation into Croatian, and comments supplementing each letter) according to the already published first book[2] (in 1989 and 1992) of the correspondence.[3] The aforementioned will also be of a great value for making a comprehensive Kuhač's biography, bibliography and zeitgeist. The database is still work in progress – the names and surnames of all the addressees from second and third volume have been entered, together with all the places (hometowns of the addressees) and some of the notions and sources. Regarding cities and places, all their name variants are specified. The brief biographies of some of the persons have been entered too, however the complete and accurate biographies still need to be researched.

The aim of this paper is a brief presentation of all the important features of the database, search options and the work which is so far done.

3. References

Marija Janaček Buljan. 1984. Kuhačeva korespondencija. In: *Zbornik radova sa znanstvenog skupa održanog u povodu rođenja Franje Ksavera Kuhača (1834-1911)* Zagreb, 20-21 November 1984, pages 463-472. JAZU, Zagreb

Vjera Katalinić – Stanislav Tuksar (ed.). 2013. *Franjo Ksaver Kuhač (1834.-1911.): Glazbena historiografija i identitet, Zbornik radova međunarodnog muzikološkog skupa*. HMD, Zagreb.

Ladislav Šaban, Koraljka Kos (ed.). 1989. *Kuhač, Franjo Ksaver: Korespondencija I/1 (1860-1862)*. JAZU, Zagreb.

Ladislav Šaban, Koraljka Kos (ed.). 1992. *Kuhač, Franjo Ksaver: Korespondencija I/2 (1863)*. HAZU, Zagreb.

[1] The link to the database: http://hmd-music.org/netmus19/index.php.

[2] Ladislav Šaban, Koraljka Kos (ed.). 1989. *Kuhač, Franjo Ksaver: Korespondencija I/1 (1860-1862)*. JAZU, Zagreb, and Ladislav Šaban, Koraljka Kos (ed.). 1992. *Kuhač, Franjo Ksaver: Korespondencija I/2 (1863)*. HAZU, Zagreb.

[3] For now, it is not planned to publish the correspondence as a digital critical edition.

Distant Reading for European Literary History. A COST Action

Christof Schöch,[1] Maciej Eder[2], Carolin Odebrecht[3], Mike Kestemont[4], Antonija Primorac[5], Justin Tonra[6], Katja Mihurko Poniž[7], Catherine Kanellopoulou[8]

[1] Department for Computational Linguistics and Digital Humanities, University of Trier,
54286 Trier
schoech@uni-trier.de

[2] Institute of Polish Language, Polish Academy of Sciences, al. Mickiewicza 31, 31-120
Krakow, Poland
maciej.eder@ijp.pan.pl

[3] Faculty of Language, Literature and Humanities, Humboldt University of Berlin,
Dorotheenstraße 24, D- 10117 Berlin
carolin.odebrecht@hu-berlin.de

[4] University of Antwerpen, Department of Literature, Grote Kauwenberg 18, S.D.115, 2000
Antwerpen, Belgium
mike.kestemont@uantwerpen.be

[5] Faculty of Humanities and Social Sciences, University of Rijeka, Sveučilišna avenija 4,
HR-51000 Rijeka, Croatia.
Antonija.Primorac@ffri.uniri.hr

[6] School of Humanities, National University of Ireland Galway, University Road, Galway,
H91 TK33, Ireland
justin.tonra@nuigalway.ie

[7] School of Humanities, University of Nova Gorica, Vipavska 13, 5000 Nova Gorica
katja.mihurko.poniz@ung.si

[8] Ionion Universtity Korfu, Department of Audio & Visual Arts 7, Tsirigoti Square, 49100
Corfu
catherine.kanellopolou@gmail.com

1. Introduction

This poster aims to stimulate awareness of the existence of the newly-established COST Action on “Distant Reading for European Literary History” (2017-2021). In the context of this networking project, “distant reading” is understood as an umbrella term for recent computational, and particularly quantitative, approaches to the study of large collections of texts. This paradigm is here applied to the multilingual literary traditions of Europe in the long nineteenth century.

2. What is a COST Action?

COST (www.cost.eu) stands for 'European Cooperation in Science and Technology': COST Actions are essentially networking initiatives focused on a particular, timely and innovative research topic, aiming to bring together a critical mass of researchers from Europe and beyond. COST Actions coordinate their activities through working group meetings and offer Training Schools and opportunities for scientific exchange. Examples of previous COST Actions in Digital Humanities include Interedition (http://www.interedition.eu/, 2008-2012) and e-Lexicography (http://www.elexicography.eu/, 2013-2017).

2.1. Aims of the "Distant Reading" Action

The contribution of the Distant Reading paradigm to Literary Studies continues to be a matter of intense debate. In our view, recent, quantitative approaches clearly provide an important

methodological perspective that usefully complements, and at times challenges, more established approaches to literary history and theory in areas like authorship attribution, genre analysis, periodization, canonization and intertextuality.

We aim to create a vibrant and diverse network of researchers jointly developing the resources and methods necessary to change the way European literary history is written. Fostering insight into crossnational, large-scale patterns and evolutions across European literary traditions, we will facilitate the creation of a broader, more inclusive and better-grounded account of European literary history and cultural identity. We will foster distributed research, the systematic exchange of expertise, and the visibility of all participants, activities and resources.

In terms of scientific objectives, we will coordinate the creation of a multilingual European Literary Text Collection (ELTeC). We will use the ELTeC to establish best practices and develop innovative methods of Distant Reading for the multiple European literary traditions. Furthermore, we will engage in an investigation into the theoretical consequences of Distant Reading approaches for literary history and literary theory. We also aim to foster the acquisition of state-of-the-art methods related to data curation, standards, best practices and quantitative analysis in workshops and training schools. Last but not least, we aim to address the current gender imbalance among practitioners of Distant Reading research.

2.2. The network

Our network of members is currently comprised of researchers in Corpus Linguistics, Computational Linguistics, (Digital) Literary History and Literary Theory from 26 different countries and more than 40 cities across Europe and beyond (Figure 1).

Figure 1: Map of Europe with the locations of Action members, created using the DARIAH GeoBrowser. Interactive version:
https://geobrowser.de.dariah.eu/?csv1=https://geobrowser.de.dariah.eu/storage/515798.

2.3. Our key deliverable: the ELTeC

Our key deliverable is the European Literary Text Collection (ELTeC) that brings together comparable sets of nineteenth-century novels from at least 10 different European languages. Each set will comprise 100 different novels published in the late nineteenth century, with extensions covering the early nineteenth century or adding additional novels from the late nineteenth century. The purpose of the ELTeC is to serve as a benchmark corpus for the

evaluation and development of annotation tools and distant reading methods across languages and as the basis for investigations into patterns and trends in literary history in multiple literary traditions. Text curation will happen on GitHub (with distributed access, issue tracking and version control). Experience with similar, national initiatives shows the importance of openness, standards and technical sustainability. As the Action progresses, linguistic annotation will be added to the texts (at least, concerning lemmata, part of speech and named entities). A shared format for the representation of document-level metadata and linguistic annotation across subcollections will be used, based on best practices in the field as recommended e.g. by DARIAH (Digital Research Infrastructure for the Arts and Humanities) and CLARIN (Common Language Resources and Technology Infrastructure).
The ELTeC will contain linguistic annotation in a cross-linguistically compatible manner by mapping each language-specific tagset onto a coarse-grained, shared tagset.

2.4. Research strands

Our activities are divided into three main research strands (organized in working groups):

- “Scholarly Resources”, focused on structuring, annotating and publishing the ELTeC;
- “Methods and Tools”, concerned with using, evaluating and developing methods for distant reading analysis;
- “Literary History and Theory”, dedicated to the theoretical consequences of distant reading methods for literary history and theory.

In addition, a working group on “Dissemination” provides infrastructure services, enables communication within the Action and gives visibility to the Action’s activities and results.

3. Learn more, learn how to join

To learn more, see the Action’s profile page http://www.cost.eu/COST_Actions/ca/CA16204 and the full proposal linked there (“Memorandum of Understanding”). The Action’s website is available at https://www.distant-reading.net/. Researchers from Computational Linguistics, (Digital) Literary Studies as well as Computer Scientists and Librarians are welcome to get involved!

4. References

Andrew Goldstone. 2017. The Doxa of Reading. *PMLA*, 3(132):636–42.

Franco Moretti. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. New York: Verso.

Andrew Piper. 2017. Think Small: On Literary Modeling. PMLA, 3(132):651–58.

Richard Jean So. 2017. ‘All Models Are Wrong’. PMLA, 3(132):668–673.

Ted, Underwood. 2017. A Genealogy of Distant Reading. Digital Humanities Quarterly, 11, 2
http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

Korpus in baza Gos Videolectures

Darinka Verdonik

* Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Koroška 46, 2000 Maribor
darinka.verdonik@um.si

1. Uvod

Leta 2011 je bil dokončan prvi sklop referenčnega govornega korpusa Gos. Zajetih je bilo 120 ur posnetkov govora oz. 1 mio. besed. S tem obsegom se korpus Gos uvršča na spodnjo mejo primerljivih referenčnih govornih korpusov za ostale evropske jezike, ki se v zadnjih letih v vse več jezikih bližajo obsegu 10 mio. besed.

Z namenom razširitve obstoječih gradiv govornega korpusa, hkrati pa tudi izdelave dodanega avdio-tekstovnega gradiva za razvoj avtomatskega razpoznavanja tekočega govora, se je leta 2016 začel projekt izdelave dodatnega korpusa in avdio baze h korpusu Gos, poimenovanega Gos Videolectures. Projekt[1] izvaja Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Zaključil se je leta 2018, celotna predvidena baza obsega ca. 14 ur govora oz. 130.000 besed.

2. Pregled stanja

Poleg referenčnega govornega korpusa Gos obstaja za slovenski jezik še nekaj večinoma manjših ali specifičnih govornih zbirk oz. korpusov.

V repozitoriju Clarin.si sta pod kategorijo »speech database« poleg Gos Videolectures na voljo še bazi SNABI (posnetih 132 govorcev, ki so prebrali vsak po 200 povedi, skupaj ca. 15.000 posnetkov) in SOFES 1.0 (transkribirani in segmentirani avdio posnetki poizvedovanj po letalskih informacijah v skupnem obsegu 10 ur). Prek ELRE je distribuirana baza BNSI Broadcast News (Žgank et al., 2004), ki zajema segmentirane avdio posnetke in transkripcije 36 ur govora informativnih TV-oddaj. Ostale govorne baze oz. korpusi niso dostopni neposredno oz. ne kot podatkovna baza. Sorodna bazi BNSI Broadcast news je baza SiBN Broadcast News (Žibert in Mihelič, 2004), ki zajema 29 ur govora informativnih TV-oddaj. Baza Sloparl (Žgank et al., 2006) vključuje 100 ur govora s transkripcijami v obliki obdelanih magnetogramov parlamentarnih razprav – gre za področno zelo specifično in zelo grobo transkribirano govorno bazo. Govorna baza projekta Translectures (Golik et al., 2013) zajema 33 ur govora za slovenščino. Narečni korpus vasi Kopriva GOKO (60 minut posnetkov) (Šumenjak, 2013) zajema posnetke narečnega govora vasi Kopriva in je dostopen prek iskalnika, ne pa kot podatkovna baza. Podobno velja za narečni korpus GOSP, ki vključuje govor vasi Osp v Slovenski Istri.

3. Korpus in avdio baza Gos Videolectures

3.1. Izbor gradiv

Gos Videolectures skuša ustrezno upoštevati tako potrebe jezikoslovja kot potrebe govornih tehnologij po jezikovnih virih. Zajema več kot 14 ur (130.000 besed transkripciji) izbranih posnetkov javnih predavanj s portala Videolectures.net. Posnetki so izbrani tako, da zastopajo različna strokovna področja in različne skupine govorce (predavateljev) glede na spol, starost in regijo.

Razlog za širitev korpusa Gos na področje javnih predavanj je aktualnost tega področja tako za jezikoslovne raziskave kot avtomatsko razpoznavanje govora. Z jezikoslovnega vidika imamo pri tem opraviti z akademskim jezikom. Gre za jezik javnega diskurza, ki ima po eni strani velik vpliv na oblikovanje slovenskega govorjenega standardnega (zbornega) jezika, po drugi strani pa skozi procese šolanja vpliva tudi na vsakdanji govorjeni jezik v visoko šolstvo vključene populacije, ki je vse večja. Ob tem ta jezik pomembno vpliva na razvoj strokovne terminologije in (bolj ali manj) uspešno širjenje slovenščine na eno najbolj zahtevnih in terminološko najbolj hitro razvijajočih se področij – znanost. Z jezikoslovnega vidika je jezik predavanj zato izredno aktualno področje.

Tudi za razvoj razpoznavanja govora so javna predavanja eno najbolj aktualnih področij za aplikacijo tehnologije: avtomatsko razpoznavanje govora v javnih predavanjih je prvi korak k avtomatskemu strojnemu prevajanju javnih predavanj, kar bi omogočilo boljšo dostopnost vsebin neslovensko govorečim (referenčni projekt s tega področja je bil Translectures – https://www.translectures.eu/). Za domače uporabnike bi z avtomatskim razpoznavanjem govora v javnih predavanjih omogočili naprednejše metode avtomatskega

[1] Projekt sta sofinancirala Ministrstvo za kulturo RS in Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna v okviru raziskovalnega programa P2-0069.

indeksiranja ter boljše avtomatsko iskanje in premikanje po vsebinah predavanj v bazah, kot je Videolectures. Izreden pomen bi imela integracija tehnologije avtomatskega razpoznavanja govora na področje javnih predavanj za gluhe in naglušne.

Tabela 1 predstavlja razporeditev izbranih posnetkov v Gos Videolectures glede na strokovno področje. Pri definiranju osnovnih področij smo izhajali iz šifrantov ved, področij in podpodročij, ki veljajo v raziskovalni dejavnosti in jih beleži ARRS (https://www.arrs.gov.si/sl/gradivo/sifranti/). ARRS deli vede na naslednje velike sklope: naravoslovje, tehnika, medicina, biotehnika, družboslovje, humanistika, interdisciplinarne raziskave. Vsaka od teh ved ima več področij in podpodročij. Precej podoben je tudi evropski šifrant raziskovalne dejavnosti CERIF, ki loči humanistične vede, družboslovje, naravoslovno-matematične vede, biomedicinske vede in tehnološke vede. OECD in EUROSTAT pa upoštevata šifrant FOS, ki loči naravoslovne vede, tehniške in tehnološke vede, medicinske in zdravstvene vede, družbene vede in humanistične vede. Na podlagi tega smo se odločili deliti izbrana predavanja na 5 področij: humanistika, družboslovje, medicina, naravoslovje/matematika ter tehnika. Pri tem smo si prizadevali kolikor mogoče enakomerno razporediti obseg gradiva po teh področjih. V končni različici v bazi vseeno nekoliko prevladuje področje humanistike, saj smo morali zaradi končnega manjšega števila besed v transkripciji, kot smo predvidevali, v prvotno enakomerno uravnotežen izbor vključiti dodatne razpoložljive posnetke. Čeprav v splošnem velja formula, da je pri govorjenju v eni minuti izgovorjenih 150 besed, je pri javnih predavanjih ta številka okrog 140 besed na minuto.

Zaradi potrebe po uravnoteženju gradiva smo se omejili tudi pri dolžini izbranih posnetkov, in sicer smo pretežno zajemali posnetke, krajše od 45 minut, kar 16 od skupno 37 posnetkov je celo krajših od 20 minut.

Področje	Št. posnetkov	Dolžina	Št. besed
Humanistika	7	4:31:03	39.871
Družboslovje	9	3:29:55	28.840
Medicina	7	2:13:59	17.721
Naravoslovje/matematika	7	2:46:31	24.202
Tehnika	7	2:31:44	21.713
Skupaj	37	15:33:12	132.347

Tabela 1: Razporeditev gradiv Gos Videolectures glede na strokovno področje.

Zajete podatke smo skušali kolikor mogoče uravnotežiti tudi po demografskih kriterijih. Korpus Gos, ki je bil pri tem izhodišče za razmislek, zajema demografske kriterije: spol, starost, dosežena izobrazba in regijski izvor, pa tudi prvi jezik (tuji govorci slovenščine) in državo bivanja (kar se nanaša na slovenske manjšine v sosednjih državah). Za podkorpus Gos Videolectures je treba te kriterije prilagoditi specifikam področja. Zajemanje govorcev iz slovenskih narodnostnih manjših in tujih govorcev slovenščine ni prednostni cilj baze. Prav tako ni smiseln kriterij o izobrazbi, saj kot predavatelji/ce večinoma nastopajo osebe z visoko izobrazbo. Kot ključni demografski kriteriji, ki jih skušamo po najboljših močeh upoštevati pri zajemanju posnetkov, tako ostanejo spol, starost in regijska pripadnost. Vendar – z izjemo spola – uravnotežanje pri tem ni bilo mogoče v celoti, saj smo lahko o starosti in regijski pripadnosti govorcev sodili le na podlagi videza, slušnega vtisa, kraja izvajanja dogodka oz. dostopnih javnih podatkov. Zanesljiv demografski podatek v bazi je zato samo podatek o spolu govorcev, kot je prikazano v tabeli 2.

Govorci	%
Moški	57 %
Ženske	43 %

Tabela 2: Razporeditev gradiv Gos Videolectures glede na spol govorcev.

Čeprav je bil izhodiščni cilj, da je število govorcev po spolu enakomerno razporejeno, je na koncu vendarle v prid moških govorcev, delno zaradi dodajanja gradiv ob koncu zaradi manjšega števila besed v transkripciji, kot je bilo predvidevano, delno pa tudi zato, ker se je pokazalo, da moški govorci v celotni bazi Videolectures.net toliko prevladujejo, da je ob upoštevanju področnih in drugih demografskih kriterijev težko popolnoma uravnovesiti obseg gradiv glede na ta kriterij.

3.2. Transkripcije

Pri transkribiranju gradiv za Gos Videolectures smo izhajali iz specifikacij transkribiranja, definiranih v korpusu Gos (Verdonik in Zwitter Vitez, 2011), in prav tako vključuje zapis na dveh nivojih: pogovornem, kjer zapišemo besede ortografsko (ne fonetično), vendar tako, kot so izgovorjene, ter standardiziranem, kjer različnim variantam neke besedne oblike pripišemo krovno standardno obliko.

Zaradi cilja, da bazo bolje prilagodimo tudi potrebam govornih tehnologij, pa vendarle vključimo nekaj manjših sprememb. O teh smo že obširno razmišljali in jih specificirali v objavah Žgank et al. (2014b) in v Verdonik (2014), zato zainteresiranega bralca napotujemo na ta vira. V splošnem gre za nekoliko natančnejše označevanje akustičnega ozadja in akustičnih dogodkov ter za nekatere posebnosti zapisovanja govora (zapisovanje dvoustničnega 'U' in člena 'ta' v pogovornem zapisu, zapisovanje neverbalnih in polverbalnih glasov, standardizacija nestandardnih polnopomenskih izrazov pa v gradivu Gos Videolectures ni bila aktualna problematika, saj se tovrstni izrazi ne pojavljajo).

Ob transkribiranju smo zabeležili tudi osnovne metapodatke, med drugim podatke o predavanju (regijo, kraj in čas, kdaj je potekalo, ter kratek opis, za kakšno predavanje gre). Za potrebe avdio procesiranja je na podlagi slušnega vtisa dodana tudi (subjektivna) informacija o kakovosti zvočnega posnetka z lestvico od 1 do 12. Za govorce smo zabeležili podatek o spolu, na podlagi dostopnih informacij pa ocenili tudi podatek o starosti (do 35 let ali nad 35 let) ter regionalni pripadnosti (jugozahodna ali severovzhodna).

Transkribiranje v pogovornem zapisu smo izvajali v orodju Transcriber 1.5.1 (Barras et al., 2000). Čeprav je na voljo novejša različica istega orodja, Transcriber AG (http://transag.sourceforge.net/), se je pri njenem testiranju pokazalo, da je nestabilna in ima preveč hroščev, starejša različica, ki teh težav nima, pa hkrati ponuja tudi vse potrebne funkcionalnosti.

Zaradi konsistentnosti zapisov je vse transkripcije pogovornega zapisa izvedel en izkušen zapisovalec. Standardizirani zapis je bil v prvem koraku avtomatsko izdelan ter nato ročno popravljen v orodju Transcriber 1.5.1, pri čemer smo hkrati izvedli tudi dodatno ročno kontrolo pogovornega zapisa in odpravili odkrite napake v zapisu.

4. Zaključek

Korpus in avdio posnetki Gos Videlectures so na voljo za raziskave prek konkordančnika NoSketchEngine pri IJS. Prav tako so dostopne izvorne datoteke, ki jih lahko uporabniki snamejo v repozitoriju CLARIN.SI. Tam so dostopni avdio posnetki v formatu wav, katerih uporaba je vezana na izvorne licence pri Videolectures.net in niso na voljo za komercialno rabo (licenca CC BY-NC-ND 4.0). Transkripcije so na voljo v treh formatih: kot TEI xml, kot vertikalna tabela Sketch Engina in kot izvorne transkripcijske datoteke, ki se lahko odprejo s programom Transcriber 1.5.1 ali drugim podobnim, ki podpira format .trs, oz. v tekstovnem urejevalniku. Dostopne so pod licenco CC-BY 4.0.

Kot smo nakazali v uvodu, v slovenščini še vedno močno zaostajamo v razpoložljivih virih za govorjeni jezik ne samo v primerjavi z velikimi evropskimi jeziki, ampak tudi v primerjavi s takimi, kjer je število govorcev podobno majhno kot pri slovenščini (npr. nizozemski, slovaški, danski, češki). Nekaj dodatnega govorjenega avdio gradiva s transkripcijami lahko v naslednjih letih pričakujemo v okviru projekta Slovenščina na dlani (http://projekt.slo-na-dlani.si/sl/), vendar bo to spet v majhnem obsegu in prilagojeno potrebam projekta, ki je usmerjen v izdelavo sodobnih učnih e-pripomočkov za učenje slovenščine v osnovnih in srednjih šolah. Potreba po novem projektu, ki bi bil usmerjen (tudi) v izgradnjo obsežnejšega kvalitetnega govornega vira, tako ostaja aktualna.

Literatura

Claude Barras, Edouard Geoffrois, Zhibiao Wu, Mark Liberman. 2000. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication, special issue on Speech Annotation and Corpus Tools*, 33(1–2):5–22.

Pavel Golik, Zoltan Tüske, Ralf Schlüter, Hermann Ney. 2013. Development of the RWTH transcription system for Slovenian. V: *Zbornik konference Interspeech 2013*, str. 3107-3111, Lyon, Francija.

Klara Šumenjak. 2013. Priprava gradiva in standardizacija nivojev zapisa za potrebe dialektološkega korpusa GOKO. V: A. Žele, ur., *Družbena funkcijskost jezika (vidiki, merila, opredelitve).* Obdobja 32, str. 443–449. Ljubljana, Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete. http://www.centerslo.net/files/file/simpozij/simp32/zbornik/Sumenjak.pdf.

Darinka Verdonik. 2014. Vprašanja zapisovanja govora v govornem korpusu Gos. V: T. Erjavec, J. Žganec Gros, ur., *Jezikovne tehnologije: zbornik 17. mednarodne multikonference Informacijska družba - IS 2014,* str. 151-156. Ljubljana, Institut Jožef Stefan. http://library.ijs.si/Stacks/Proceedings/InformationSociety/2014/2014_IS_CP_Volume-G_(LT).pdf.

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

Andrej Žgank, Tomaž Rotovnik, Mirjam Sepesy Maučec, Darinka Verdonik, Jani Kitak, Damjan Vlaj, Vladimir Hozjan, Zdravko Kačič, Bogomir Horvat. 2004. Acquisition and annotation of Slovenian Broadcast News database. V: *Zbornik konference LREC 2004*, str. 2103–2106. Lizbona, Portugalska..

Andrej Žgank, Tomaž Rotovnik, Matej Grašič, Marko Kos, Damjan Vlaj, Zdravko Kačič. 2006. SloParl - Slovenian parlamentary speech and text corpus for large vocabulary continuous speech recognition. V: *Zbornik konference Interspeech 2006*, str. 197-200. Pittsburgh, Pensylvania, ZDA.

Andrej Žgank, Gregor Donaj, Mirjam Sepesy Maučec. 2014. Razpoznavalnik tekočega govora UMB Broadcast news 2014: Kakšno vlogo igra velikost učnih virov? V: T. Erjavec, J. Žganec Gros, ur., *Jezikovne tehnologije – IS 2014*.

Andrej Žgank, Ana Zwitter Vitez, Darinka Verdonik. 2014. The Slovene BNSI broadcast news database and reference speech corpus GOS: towards the uniform guidelines for future work. V: *Ninth International Conference on Language Resources and Evaluation*, str. 2644-2647. Reykjavik, Islandija. http://www.lrec-conf.org/proceedings/lrec2014/index.html.

Janez Žibert, France Mihelič. 2004. Development of Slovenian broadcast news speech database. V: *Zbornik konference LREC 2004.* Lizbona, Portugalska.

Korpus tvitov slovenskih politikov Janes TwePo

Urška Bratoš*

* Maribor

Povzetek

V prispevku predstavljamo izdelavo korpusa tvitov slovenskih politikov Janes TwePo. Pojasnjujemo izbiro predmeta raziskovanja, proces pridobivanja virov in oblikovanje končnega seznama ter osnovne metapodatke. S pomočjo konkordančnika Sketch Engine smo nato na podlagi korpusnih metapodatkov izvedli še kratko analizo tviterskega jezika politikov. Analiza teh podatkov je med drugim pokazala, da politiki na Twitterju uporabljajo standardno slovenščino, vendar manj kot pričakovano.

1. Uvod

Družbeno omrežje Twitter je bilo ustanovljeno leta 2006 in je danes s 335 milijonov mesečno aktivnih uporabnikov (statista.com) eno izmed priljubnejših spletnih platform. Med njegovimi uporabniki so tudi politiki, ki Twitter uporabljajo predvsem za širjenje političnih sporočil in samopromocijo v političnih kampanjah (Golbeck et al., 2010). Ker predstavljajo institucionalni glas ljudstva in zasedajo najpomembnejše funkcije v državi, se od politikov pričakuje določena mera profesionalnosti tudi na družbenih omrežjih, katerih prvotni namen je komuniciranje zasebne narave. Eden izmed temeljnih pokazateljev profesionalnosti politikov je uporaba jezika. Dosedanje raziskave kažejo, da se jezik na družbenih omrežjih, kot že prej npr. v SMS-ih (Kalin Golob, 2009), v marsičem razlikuje od pisnega standarda (Erjavec in drugi 2015). Še pred pojavom družbenih omrežij je internetni jezik analiziral Crystal (2001) in ugotovil, da ima značilnosti tako govorjenega kot zapisanega jezika. V njem so pogoste krajšave, uporaba emocionalne ikonografije, spletni neologizmi, neobičajna raba ločil in simbolov ter igriva ponavljanja in grafološke inovacije. Stavki jezika družbenih omrežij so enostavni in zgoščeni, pojavlja se slogovna heterogenost in jezikovna inovativnost (Strehovec, 2003). Za jezik družbeno izpostavljenih posameznikov, ki so izvoljeni predstavniki ljudstva, se zdi, da tovrstnih odstopanj ni pričakovati, vendar poglobljene analize jezika javnih osebnosti ali politikov na tem omrežju na slovenskem gradivu še niso bile narejene.

V prispevku predstavljamo izdelavo korpusa tvitov slovenskih politikov Janes TwePo. Kot kaže že ime, smo besedila pridobili iz gradiva korpusa Janes (Fišer et al., 2016). S pomočjo konkordančnika Sketch Engine (Kilgarriff et al., 2004) smo nato izvedli še korpusno analizo, ki jo omogočajo metapodatki, in sicer z naslednjim ciljem: ugotoviti standardnost, sentiment in ključne besede tviterskega jezika politikov.

2. Gradnja korpusa

V korpus Janes TwePo smo iz korpusa Janes (Fišer et al., 2016) zajeli vse tvite slovenskih politikov, parlamentarnih strank, ministrstev in vlade, napisane v obdobju od 1. junija 2013 do 5. januarja 2016, nato pa zaradi obsega raziskavo omejili le na analizo tvitov posameznih politikov. Ustvarili smo podkorpus Janes TwePo-Posamezniki in vanj vključili le tiste politično delujoče fizične osebe, ki so v izbranem časovnem okvirju opravljale funkcijo predsednika države, predsednika vlade, ministra, poslanca, evroposlanca ali župana. Tvite pravnih oseb, torej institucij, kot so vlada, parlamentarne stranke in ministrstva smo iz analize izključili.

Podatki o predsedniku države, predsednikih vlade, ministrih, poslancih, evropskih poslancih in evropskih komisarjih so dostopni na uradnih spletnih straneh. Pri oblikovanju seznama poslancev in strank smo si pomagali tudi s poročilom o delu državnega zbora v dveh obdobjih (2011–2014 in 2014–2018, dz-rs.si), medtem ko smo podatke o županih pridobili pri pristojni osebi z Ministrstva za javno upravo RS.

Ko smo glede na časovni okvir in politično funkcijo pripravili abecedno urejen seznam politikov, nas je zanimalo dvoje: ali ima politik na Twitterju ustvarjen uporabniški račun in ali je ta profil zaveden v korpusu Janes. Sledil je ročni vnos podatkov za vsakega politika posebej. Velja še omeniti, da smo v korpus vključili tudi politike, katerih računi niso več aktivni, so pa zajeti v korpus Janes, medtem ko v korpus nismo vključili računov s potencialno relevantnimi uporabniškimi imeni, ki pa jih nismo mogli zanesljivo identificirati (so brez slike in profila), prav tako nismo vključili tvitov politikov, ki sicer imajo račun na Twitterju, vendar so objavili manj kot 50 tvitov.

Metapodatki, ki spremljajo vsak tvit v Janes TwePo, so naslednji:

- **uporabniško ime na Twitterju**
- **ime in priimek politika**
- **spol politika**
- **raven politične funkcije:** lokalna, državna ali evropska
- **politična funkcija:** predsednik države, predsednik vlade, minister, poslanec, evropski poslanec, župan
- **politična stranka, ki ji politik pripada** (pri županih stranka, ki ga podpira)

V kategorijah raven *politične funkcije, politična funkcija* in *stranka* je bilo lahko izbranih več vrednosti (npr. v primerih, ko so poslanci Državnega zbora RS postali evropski poslanci, ali ko so politiki z daljšim stažem opravljali več kot eno funkcijo in ko so politiki zamenjali stranko).

Da bi zmanjšali možne napake, ki se ob ročnem preverjanju lahko zgodijo, smo postopek razvrščanja in preverjanja vseh podatkov ponovili 5-krat. Nato smo iz korpusa Janes zajeli vso relevantno gradivo in izdelali korpus. Pri tem smo prevzeli vse obstoječe Janesove metapodatke in oblikoskladenjske oznake ter pripisali še nove metapodatke, naštete zgoraj. Korpus smo naložili na konkordančnik Sketch Engine, po predstavitvi prispevka na konferenci pa ga bomo objavili tudi v repozitoriju CLARIN.SI.[1]

3. Zgradba korpusa in prve analize

3.1. Velikost korpusa

Celoten korpus Janes TwePo vsebuje nekaj več kot 104 tisoč tvitov, ki obsegajo skoraj 1,8 milijona pojavnic (dobrih 1,3 milijone besed) in 75 tisoč lem (Tabela 1). Podkorpus Janes TwePo-Posamezniki vsebuje (ki ne vključuje vlade, ministrstev in parlamentarnih strank), pa vsebuje nekaj več kot 77 tisoč tvitov, ki obsegajo približno milijon besed in 65 tisoč lem, kar predstavlja slabe tri četrtine celotnega korpusa tvitov politikov.

	TwePo	TwePo-Posamezniki
Št. tvitov	104.369	77.643
Št. pojavnic	1.791.166	1.284.212
Št. besed	1.354.241	1.056.858
Št. lem	74.251	64.845

Tabela 1: Velikost korpusa Janes TwePo in Janes TwePo-Posamezniki.

3.2. Spol avtorja tvita

Podkorpus TwePo-Posamezniki vsebuje tvite skupno 78 slovenskih politikov[3] od tega 50 (64 %) politikov moškega spola in 28 (36 %) političark, pri čemer so od skupno 77.643 tvitov, ki so zajeti v naš podkorpus, 62.745 tvitov (81 %) napisali moški, 14.898 (19 %) pa ženske (Tabela 2). V povprečju to pomeni 1255 tvitov na posameznega politika in 532 tvitov na posamezno političarko – povedano drugače: v obravnavanem obdobju je bilo na Twitterju aktivnih za tretjino več politikov kot političark, ki so v skupni obseg tvitov prispevali štirikrat toliko sporočil kot političarke.

Spol	Št. tvitov	Delež v %
Moški	62.745	81
Ženski	14.898	19
Skupaj	**77.643**	**100**

Tabela 2: Število tvitov in delež v korpusu Janes TwePo glede na spol njihovih avtorjev oz. avtoric.

3.3. Raven politične funkcije avtorja tvita

Podatki kažejo, da so največji odstotek tvitov objavili politiki, aktivni na državni ravni, sledijo tviti evropski politikov, nato pa tviti lokalnih politikov (Tabela 3). To seveda ni presenetljivo, saj je tudi število politikov na evropski in lokalni ravni znatno manjše kot na državni. Ustreznejšo primerjavo zato kaže relativni obseg: relativno gledano, "vodijo" župani, ki so objavili 1,6-krat toliko sporočil kot državni poslanci.

Raven politične funkcije	Frekvenca	Delež v %
Slovenska	61.989	78
Evropska	12.055	15
Lokalna	5.094	7
Skupaj	**79.138**	**100**

Tabela 3: Število in delež tvitov v korpusu Janes TwePo glede na raven politične funkcije njihovih avtorjev oz. avtoric.

3.4. Politična funkcija avtorja tvita

Največ tvitov so napisali poslanci Državnega zbora RS, in sicer 54.950, sledijo evropski poslanci z 12.055 tviti. Ministri so objavili 5.686 tvitov, 5.094 tvitov so napisali župani, predsednik države je napisal 3.113 tvitov,[2] 1.465 sta jih napisala premiera (Tabela 4).

Politična funkcija	Št. tvitov	Delež v %
Poslanec	54.950	67
Evropski poslanec	12.055	15
Minister	5.686	7
Župan	5.094	6
Predsednik države	3.113	4
Predsednik vlade	1.465	2
Skupaj	**82.363**	**100**

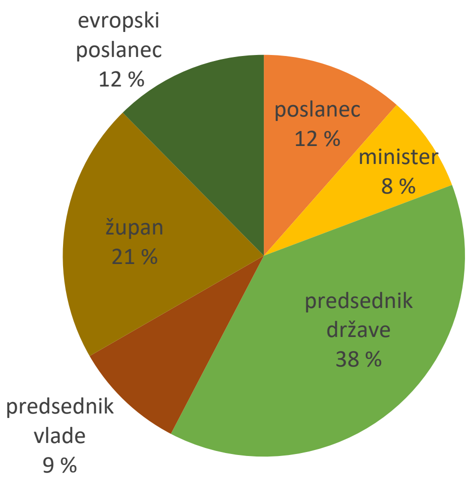
Tabela 4: Število in delež tvitov v korpusu Janes TwePo glede na politično funkcijo njihovih avtorjev oz. avtoric.

Če podatke delimo s številom politikov, ki opravljajo določeno funkcijo, dobimo podatek o tvitersko najaktivnejši politični funkciji: največ tvitov je napisal predsednik države, sledijo župani, predsednika vlade, poslanci in evropski poslanci ter ministri (Slika 1).

[1] http://www.clarin.si/info/o-projektu/

[2] Na profilu predsednika države tvite objavi predsednik sam ali njegova ekipa (označeno s PRS). V analizi so zajeti vsi tviti tega profila.

[3] Celoten korpus Janes TwePo zajema tvite skupno 78 politikov, 9 parlamentarnih strank (SDS, NS, SLS, SD, PS, ZL, ZAB, SMC, DL), 3 ministrstev (Ministrstvo za izobraževanje, znanost in šport RS, Ministrstvo za kulturo RS in Ministrstvo za obrambo RS) in 1 vlade.



Slika 1: Delež tvitov v korpusu Janes TwePo glede na politično funkcijo njihovih avtorjev oz. avtoric.

3.5. Politična stranka, ki ji avtor tvita pripada

V podkorpusu analizirani politiki pripadajo skupno 23 različnim strankam. Podatki v Tabeli 5 kažejo, da so na Twitterju najaktivnejši člani stranke Slovenska demokratska stranka (SDS) z napisanimi nekaj manj kot 28 tisoč tviti. S skoraj polovico manj tvitov, 16 tisoč, sledi Stranka modernega centra (SMC, prej Stranka Mira Cerarja), s skoraj 10 tisoč tviti je na tretjem mestu Državljanska Lista (DL), na četrtem pa so Socialni demokrati (SD) z dobrimi 9 tisoč tviti. Najmanj aktivni so bili v obdobju, ki ga zajema naš podkorpus, nestrankarski kandidati za župane, Združena levica (ZL), stranka Zares in Demokratična stranka upokojencev (Desus).

Politična stranka, ki ji politik pripada	Št. tvitov	Delež v %
SDS	27.604	28
SMC	16.027	16
DL	9.932	10
SD	9.269	9
SLS	4.890	5
PS	4.496	5
Zveza za primorsko ZZP	3.461	4
SMS STRANKA MLADIH ZELENI SLOVENIJE	3.461	4
MAŠA KLAVORA IN SKUPINA VOLIVCEV	3.461	4
NSI IN SLS	2.450	2
NP	2.156	2
SD IN STRANKA EVROPSKIH SOCIALISTOV	2.098	2
ZAAB	1.853	2
NSI	1.781	2
ZBOR ČLANIC IN ČLANOV ZAAB KOMEN	1.594	2
LDS	1.398	1
LISTA DR. IGORJA ŠOLTESA	941	1
DESUS	598	1
ZARES	565	1
ZL	177	0
LEANA TOMIČ IN SKUPINA VOLIVCEV	33	0
TJAŠA MÖDERNDORFER S SKUPINO VOLILCEV	6	0
JOŽE MERMAL IN SKUPINA VOLIVCEV	6	0
Skupaj	**98.257**	**100**

Tabela 5: Število in delež tvitov v korpusu Janes TwePo glede na strankarsko pripadnost njihovih avtorjev oz. avtoric.

3.6. Število tvitov po posameznih avtorjih

Podatki iz podkorpusa kažejo, da je vseh 78 analiziranih politikov objavilo skupno 77.643 tvitov, kar v povprečju pomeni 995,4 tvita na posameznega avtorja.

Največ tvitov je v analiziranem obdobju objavil Marko Pavlišič (poslanec stranke DL, obdobje 3 let in 6 mesecev), in sicer 9.115 (Tabela 6), kar pomeni, da je v povprečju dnevno objavil nekaj več kot 7 tvitov. Drugi na Twitterju najaktivnejši politik je bil Kamal Shaker (poslanec SMC), tretji pa Bojan Krajnc (poslanec SMC). Na petem mestu je bil poslanec SDS Tomaž Lisec, na šestem pa predsednik države Borut Pahor. Opaziti je, da so na tviterju od vseh političnih funkcij najaktivnejši poslanci državnega zbora, in to ne glede na starostno generacijo, in da ne gre za predsednike ali predsednice strank. Izstopa tudi podatek, da so največ tvitov napisali člani stranke SDS, vendar pa so se prav med najaktivnejše politike na Twitterju uvrstili le trije člani te stranke, kar pomeni, da je tvitanje članov SDS "razpršeno". Janez Janša je glede na število objavljenih tvitov zasedel 12. mesto.

Politik	Št. tvitov	Delež v %
Marko Pavlišič (DL)	9.115	23
Kamal Shaker (SMC)	6.355	16
Bojan Krajnc (SMC)	5.956	15
Uroš Brežan (SLS)	3.461	9
Tomaž Lisec (SDS)	3.297	8
Borut Pahor (SD)	3.113	8
Jožef Jerovšek (SDS)	2.596	6
Roman Jakič (PS)	2.340	6
Andrej Čuš (SDS)	2.106	5
Tanja Fajon (SD)	2.098	5
Skupaj	**40.437**	**100**

Tabela 6: Politiki, ki so v korpus Janes TwePo prispevali največ tvitov (prvih 10 mest), ter število in delež njihovih tvitov.

Najmanj tvitov je v analiziranem obdobju objavila ekipa Zorana Jankovića (poslanec PS). Druge gl. v Tabeli 7.

Politik	Št. tvitov	Delež v %
Zoran Janković	6	2
Ljubo Žnidar	17	6
Dragutin Mate	18	7
Danijel Krivec	21	8
Zvonko Lah	26	10
Aljoša Jerič	31	12
Peter Vilfan	33	12
Jasna Gabrič	33	12
Polonca Komar	39	15
Irena Tavčar	42	16
Skupaj	**266**	**100**

Tabela 7: Politiki, ki so v korpus Janes TwePo prispevali najmanj tvitov (zadnjih 10 mest), ter š tevilo in delež njihovih tvitov.

3.7. Tviti politikov: sentiment in standardnost

Iz metapodatkov, vključenih v korpus Janes TwePo (in pridobljenih že iz korpusa Janes, Fišer in Erjavec, 2016; Ljubešić et. al., 2015), je bilo mogoče avtomatsko pridobiti tudi podatke o tem, v katerem jeziku politiki pišejo tvite, kateri sentiment v njih prevladuje in kakšna je raven njihove skladnosti s standardno slovenščino.

V korpusu Janes TwePo je večina tvitov napisana v slovenščini (93 %), manj kot 5 % tvitov je napisanih v angleščini. Slednjo uporabljajo tako poslanci Državnega zbora RS kot evropski poslanci, saj občasno nagovarjajo tuje naslovnike. Ostala 2 % tvitov sta zapisana v nemškem, italijanskem, hrvaškem ali drugem jeziku.

Nadalje nas je zanimala analiza sentimenta besedila, ki pokaže, ali je avtor oz. avtorica tvita temi, o kateri tvita, naklonjena ali ne. Sentiment je lahko označen kot negativen, pozitiven ali nevtralen (Fišer in drugi, 2016). Slaba polovica tvitov, ki so jih objavili politiki, ima v našem korpusu po metapodatkih nevtralen sentiment (45 %), 30 % jih ima negativen sentiment, 25 % pa pozitivnega (Tabela 8).

Sentiment	Št. tvitov	Delež v %
Nevtralen	35.348	45,5
Negativen	23.066	29,7
Pozitiven	19.229	24,8
Skupaj	**77.643**	**100**

Tabela 8: Število in delež tvitov v korpusu Janes TwePo glede na njihov sentiment.

Sentiment smo nato povezali s podatki o spolu avtorjev tvita in ugotovili, da so političarke napisale približno enako število tvitov z negativnim in pozitivnim sentimentom, moški politiki pa so napisali več tvitov z negativnim odnosom do vsebine (Tabela 9).

Sentiment	Spol	Št. tvitov	Delež v %
Negativen	Moški	18.861	82
	Ženski	4.202	18
Skupaj		**23.063**	**100**
Pozitiven	Moški	14.711	77
	Ženski	4.518	23
Skupaj		**19.229**	**100**

Tabela 9: Število in delež tvitov v korpusu Janes TwePo glede na sentiment in spol avtorja oz. avtorice.

O standardnosti jezika tvitov je mogoče ugotoviti naslednje: politiki so v svojih tvitih uporabljali pretežno standardno slovenščino (68 % tvitov, Tabela 10), zelo nestandardnih zapisov je bilo malo (6 %), npr.:

- a) zamenjava šumnikov s sičniki:
Dragi zupani! Ce ne ze prej, ste danes spoznali ...
(Anja Bah Žibert, 23. 10. 2013)

- b) nepravilna raba ločil, velikih in malih začetnic:
juhuhu se bomo tu... tudi v letu 2015:)
(Anja Bah Žibert, 31. 12. 2014)

- c) napačno pisanja skupaj in narazen ali tipkarske napake:
osebnovkljucit humanitarno, se vam ne zdi???
(Igor Šoltes, 11. 9. 2015)

- č) uporaba pogovornega jezika:
Buh jim je zaplosku. #snežet #potres
(Uroš Brežan, 29. 8. 2015)

Standardnost	Št. tvitov	Delež v %
L1	52.427	68
L2	20.419	26
L3	4.797	6
Skupaj	**77.643**	**100**

Tabela 10: Število in delež tvitov v korpusu Janes TwePo glede na njihovo standardnost.

4. Zaključek

Twitter je v zadnjih letih postal pomemben kanal za politično komuniciranje, zato smo se odločili, da bomo za analizo jezika politikov zgradili korpus. Pri tem nam je bil v pomoč že pripravljeni korpus spletne slovenščine Janes, katerega kar obsežen del so tudi tviti. Kot podkorpus "izločeni" Janes TwePo, zajet v tukajšnjo analizo, tako vsebuje kar 1,2 milijona pojavnic in je tako relevantno dobra osnova za različne jezikovne analize.

V nadaljevanju želimo raziskavo nadgraditi z natančnejšo analizo sentimenta, določiti, kako je sentiment izražen z jezikovnimi sredstvi, in analizirati ključne besede ter podrobneje proučiti na eni strani pravopisna in slovnična, na drugi pa besedna odstopanja od normativnih pravil in nevtralnega stila.

5. Zahvala

Raziskava, opisana v prispevku, je bila opravljena v okviru nacionalnega temeljnega projekta *Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine* (J6-6842, 2014-2017), ki ga financira ARRS. Za tehnično pomoč pri izdelavi, označevanju in objavi korpusa se zahvaljujem prof. dr. Tomažu Erjavcu. Prispevek je del magistrske naloge z naslovom *Jezik politikov na družbenem omrežju Twitter in tviti politikov kot vir novinarskega sporočanja*, ki nastaja na Fakulteti za družbene vede pod mentorstvom doc. dr. Nataše Logar in doc. dr. Darje Fišer. Mentoricama se lepo zahvaljujem za branje in popravljanje prispevka.

6. Literatura

David Crystal. 2011. *Internet Linguistic*. London in New York: Routledge.

Darja Fišer in Tomaž Erjavec. Analysis of sentiment labeling of Slovene user-generated content. V: Darja Fišer (ur.), in Michael Beisswenger (ur.). *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, 27-28 September 2016, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*. 1st ed. Ljubljana: Znanstvena založba Filozofske fakultete. 2016, str. 22-25, ilustr. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Fiser_Erjavec_Analysis-of-Sentiment-Labeling.pdf.

Darja Fišer, Jennifer Golbeck, Justin M. Grimes in Anthony Rogers. 2010. Twitter Use by the U.S. Congress. *Journal of the American Society for Information Science and Technology* 61, 8, 1612–1621.

Monika Kalin Golob. 2009. Med pisnim in govornim ali zgolj po svoje: SMS-sporočila. V: Tanja Oblak Črnič in Breda Luthar (ur.): *Mobilni telefon in transformacija vsakdana.* Ljubljana: FDV. 81–95.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz in David Tugwell. 2004. The Sketch Engine.V: *Proceedings of EURALEX 2004*, str.105–116, Lorient.

Nikola Ljubešič, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak, Iza Škrjanec. Predicting the level of text standardness in user-generated content. V: *Proceedings*, International conference Recent Advances in Natural Language Processing, Hissar, Bulgaria, 7-9 September, 2015. Hissar: [s.n.]. 2015, str. 371-378, ilustr. http://lml.bas.bg/ranlp2015/docs/RANLP_main.pdf.

Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2018 (in millions). *https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/* (2. september 2018).

Janez Strehovec. 2003. *Umetnost interneta. Umetniško delo in besedilo v času medmrežja.* Ljubljana: Študentska založba.

Nada Šabec. 2014. Raba slovenščine in angleščine v fizičnem in virtualnem prostoru. *Slavistična revija*, letnik 62/2014.

You, thou and thee: A statistical analysis of Shakespeare's use of pronominal address terms

Isolde van Dorst

ESRC Centre for Corpus Approaches to Social Science, Lancaster University (UK)
Faculty of ICT, University of Malta (MT)
Faculty of Arts, University of Groningen (NL)

This study creates a prediction model to identify which linguistic and extra-linguistic features influence pronoun choices in the plays of Shakespeare. In the English of Shakespeare's time, the now-archaic distinction between YOU and THOU persisted, and is usually reported as being determined by relative social status and personal closeness of speaker and addressee. But it remains to be determined whether statistical machine learning will support this traditional explanation. 23 features are investigated, having been selected from multiple linguistic areas, such as pragmatics, sociolinguistics and conversation analysis. The three algorithms used, Naive Bayes, decision tree and support vector machine, are selected as illustrative of a range of possible models in light of their contrasting assumptions and learning biases. Two predictions are performed, firstly on a binary (YOU/THOU) distinction and then on a trinary (*you*/*thou*/*thee*) distinction. Of the three algorithms, the support vector machine models score best. The features identified as the best predictors of pronoun choice are the words in the direct linguistic context. Several other features are also shown to influence the pronoun prediction, including the names of the speaker and addressee, the status differential, and positive and negative sentiment.

1. Introduction

For several decades much research has been undertaken on the use of *you*, *thou* and *thee* in Shakespeare's works. However, the results so far have yet to arrive at an exact and conclusive answer regarding how these pronouns were used.

This study combines the strengths of multiple research fields in an effort to determine via hitherto unused methods which linguistic and extra-linguistic features influence the choice of second person singular pronoun (*you* versus *thou* or *thee*) in the plays of William Shakespeare. Prior findings in literary and linguistic studies are utilised to find which features could be relevant in this choice, and tools and applications created for corpus linguistics and computer science are exploited to analyse the data in a more exact way than has so far been accomplished. Through these techniques, I hope to identify which features can contribute to a more accurate prediction of pronoun choice, in a model to mimic the pronoun use of Shakespeare.

It is worth observing at this point that it has not yet been determined whether it is even possible to predict the pronoun based on linguistic features. Part of the aim of this paper is to make a determination on this point. In other words, is it possible to create a computational model that can predict which pronoun will be used based on a set of linguistic and extra-linguistic features taken from the text itself and selected on the basis of knowledge that we have of English in the late 1500s and early 1600s? To accomplish this, all occurrences of *you*, *thou* and *thee* are extracted from Shakespeare's plays, and every instance is manually coded for 23 linguistic and extra-linguistic features, creating data which will serve to ascertain the answer to this primary question. A second question to be addressed is whether some features perform better as predictors of the pronoun choice than others. Thirdly, the issue of whether the use of different algorithms affects the prediction outcomes will be considered.

Throughout this paper, italicised *you*, *thou* and *thee* refer to specific pronoun forms. However, whereas *you* – in Early Modern English as in contemporary English – does not exhibit any formal variation for pronoun case, *thou* is strictly a nominative form with *thee* as its accusative/dative form. *Thou* and *thee* are therefore related inflectional forms of a single pronoun lemma; *you* exists in variation with both. Small capitals are used to indicate the pronoun lemmas, thus: YOU and THOU, where THOU includes both *thou* and *thee*. Whenever discussing pronouns in this paper, I am strictly referring to the singular second-person pronouns *you*, *thou* and *thee* that are examined in this study.

2. Background

2.1. Digital Humanities

Over the past few years, computational research has branched out into other research fields that are not necessarily closely connected to computer science. Digital Humanities (DH) is an umbrella term for all research that is computational but approaches the datasets investigated within, and/or addresses questions or problems that are of importance to, the disciplines of the humanities.

The popularity of Digital Humanities, a cross-domain field of study, is attributable to the fact that it does not diminish the differences between fields but rather operationalises this difference to solve difficulties that could not be dealt with within a single discipline. The role of computational methods in the humanities can be considered as that of a supporting character; in any DH computer modelling research, it should be kept in mind that the interpretation is as important at the suitability of a computational model and its outcomes..

2.2. Early Modern English and YOU/THOU

In Early Modern English (EModE), two different second person singular pronouns were used, namely the formally singular THOU and the formally plural (but pragmatically also respectful-singular) YOU, with only the latter surviving the EModE period (Taavitsainen and Jucker, 2003). The difference between the uses of these two pronouns is evident from multiple literary studies that have addressed Shakespeare's work, work of his contemporaries, and other documents from this era, such as Walker (2003) and Busse (2002). These studies suggest that unwritten social rules governed the use of these pronouns, abiding by which rules was necessary in order to speak according to society's standards. The use of the two different pronouns acted as a sign of relative status: YOU would be used to superiors and THOU towards inferiors. The choice of pronoun can thus also operate as a subtle means of showing respect or disrespect; using the pronouns in this way would have been natural and easy to English native speakers of the period.

Shakespeare lived during the Early Modern English period, and thus used both YOU and THOU in his writing. His work was written less than 100 years before *thou* and *thee* disappeared from the standard language (surviving in dialects and archaicised registers, such as pious addresses to the divinity). Thus we may straightforwardly posit that the disappearance of THOU was likely already in progress around his time. Though obviously heightened in its use of emotional and dramatic language and style to accommodate to the genre of the play script, the language of Shakespeare – including the usage of the two second-person pronouns – can be assumed to be a reasonably good representation of the language used generally in social interaction and conversation at that time (Calvo, 1992).

2.3. Prior studies on YOU/THOU

Most studies of Shakespeare's use of YOU and THOU so far have been literary and nonnumeric studies (Brown and Gilman, 1960; Quirk, 1974; Calvo, 1992); the relative few to have used data-based or quantitative techniques did not implement any method beyond directly comparing raw frequency counts (Busse, 2003; Mazzon, 2003; Stein, 2003). Moreover, these studies did not look at all the extant Shakespeare plays, but instead chose a few plays to focus on. Nonetheless, these studies have demonstrated some patterns in the use of YOU and THOU and thus provide a workable foundation for a more in-depth study of the usage of those two pronouns.

These prior studies support in the overall conclusion that the pronouns YOU and THOU appear to be used to support the explicit expression of respect, social status, and familiarity. Quirk (1974) and Mazzon (2003) characterise the role of the pronoun as a linguistic marker, whose usage can be seen as either marked or unmarked. In other words, the use of a particular pronoun can be seen as marked when it is used unexpectedly, for example when YOU is expected based on social status, but THOU is used. Thus, in contrast to earlier studies (Brown and Gilman, 1960), they do not perceive YOU and THOU to be in direct contrast, and to have a more variable interpretation than was assumed until then, based on the context it occurs in. Calvo (1992) and Stein (2003) expand on this by concluding that markedness of the pronoun is dependent on the context and the situation, in addition to the pronoun choice depending on stable factors such as the social statuses of, and the level of familiarity between, the characters in Shakespeare's plays; the speakers and addressees in this study – rather than *just* the latter factors (Brown and Gilman, 1960). The emotive effect of the utterances within which the YOU/THOU distinction is utilised is of importance as well; feelings such as anger and love for another character may find expression through pronoun choice. This is connected to the notion of respect, as, in an angry remark, marked pronouns can be used to disrespect the addressee based on their social status. (Stein, 2003).

As Stein (2003) and Busse (2006) already stressed in their studies, a study of YOU and THOU in Shakespeare cannot and should not be limited to a single research discipline. Rather, what is needed is a combination of literature, sociolinguistics, pragmatics and conversation analysis, which are all useful in capturing the complexity of pronominal address and the social constrictions that may have underpinned the choice of one honorific pronoun-form over the other.

3. Methodology

As has already been mentioned, this is a strictly empirical study which attempts to verify the findings of earlier research through a computational approach. The use of a computational, statistical method is motivated by the goal of creating a more objective representation of Shakespeare's use of YOU and THOU in his plays than has been accomplished so far, since it does not require analysis of meaning-in-context by a human being, but rather proceeds directly from quantitative measurements.

3.1. Hypotheses

Three hypotheses were formulated on the basis of the literature:

1. No single model will be able to predict the pronominal address term solely based on linguistic and extra-linguistic features.

This, being a null-hypothesis, is exactly what this study aims to falsify by developing such a model. It is not likely that a single model will be able to predict Shakespeare's original choice of YOU or THOU based on linguistic and extra-linguistic features, because this choice is dependent on so many factors. However, the combined application of literature, sociolinguistics, pragmatics and conversation analysis all combined into a computational model will be able to successfully predict the pronoun choice as it includes all the factors that might influence the choice for either YOU or THOU.

2. The features of social status, age and sentiment will be better predictors of the pronoun choice than other features.

A hierarchy will be established according to which the linguistic and extra-linguistic features are in the best performing model. It may be inferred from the literature that social status, age and sentiment are highly likely to be at the top of this hierarchy, among the most influential features; these three features have shown up most reliably in prior research.

3. The best performing algorithm will combine features both dependent and independently.

The different learning biases and assumptions of the three algorithms applied in this study will reveal how the features interact with one another. The first algorithm, Naive Bayes, assumes all features are independent of one another, while the decision tree algorithm assumes that the features are all dependent on each other. Lastly, the support vector machine works with both dependent and independent features. I expect the features to be a combination of dependent and independent of one another and therefore the support vector machine models to perform best. The three algorithms will be discussed in more detail later in 3.3.

3.2. Data

The data for this study comes from the *Encyclopaedia of Shakespeare's Language* project[1], which is a research project of Lancaster University (UK). The project corpus consists of 38 of Shakespeare's plays, which includes all 36 plays from the First Folio with the addition of *The Two Noble Kinsmen* and *Pericles: Prince of Tyre*. A broadly annotated version of the full Shakespeare corpus can be found online[2]. Some of the annotation and all of the abbreviations used for the titles of the plays follow *The Arden Shakespeare*.

3.2.1. Linguistic and extra-linguistic features

Feature	Acronym	Annotation
Genre	Genre	Pre-annotated
Play name	Play	Pre-annotated
Play, act, scene	Scene	Pre-annotated
Speaker ID	S_ID	Pre-annotated
Speaker gender	S_Gender	Pre-annotated
Speaker status	S_Status	Pre-annotated
Production date	Prod_Date	Pre-annotated
N-gram	LW1-3, RW1-3	Automatic
Positive sentiment	Pos_Sent	Automatic
Negative sentiment	Neg_Sent	Automatic
Speaker age	S_Age	Manual
Location	Location	Manual
Addressee ID	A_ID	Automatic
Addressee gender	A_Gender	Pre-annotated
Addressee status	A_Status	Pre-annotated
Addressee age	A_Age	Manual
Status differential	Stat_Diff	Automatic
No. of people addressed	A_Number	Pre-annotated

Table 1: List of all features used in this study

The Encyclopaedia of Shakespeare's Language corpus is richly annotated. However, some additional annotation was necessary to perform a full analysis of what extra-linguistic features could be predictors of the pronominal address term. The full set of features used in this study can be found in Table 1. The added features are briefly described here.

As a referent (such as a second person singular pronoun) is dependent on context, the adjacent part of the utterance is used as a feature to test the effect of co-text. Six co-textual words are included, i.e. a 7-gram altogether. "LW" labels the words occurring on the left of the pronoun, and "RW" the words on the right of the pronoun. Each of these words are numbered based on their distance from the pronoun, e.g., LW3 is the third word on the left of the pronoun. In corpus linguistics, collocations are often examined within a three-word-window, meaning there are three words on either side of the word of interest. While I am not necessarily looking at specific collocations of YOU and THOU, the LW/RW features will look at similarities and

[1] More information on this project, which is funded by the Arts and Humanities Research Council (AH/N002415/1), can be found on http://wp.lancs.ac.uk/shakespearelang/.

[2] CQPweb: http://cqpweb.lancs.ac.uk

differences in co-textual words to see if they can predict the pronoun choice.

Another feature mentioned as critical by prior studies is sentiment, that is the use of the pronoun to convey positivity or negativity. Sentiment was annotated with the use of the 7-gram described above. *SentiStrength* is a lexicon-based sentiment analysis program that scores phrases with a score for positivity and negativity (Thelwall et al., 2010). Since *SentiStrength* was developed to work with online comments rather than complete sentences as in formal written English, it works well with n-grams too. The scores for positivity and negativity are kept as separate variables.

The corpus already included metadata on the speakers; however, I wanted to include age as well. The age of a character is often not given except for when it is an important attribute of that character, making this difficult to annotate. Therefore, Quennell and Johnson's (2002) character descriptions were used. The characters were sorted into a trinary classification, with 'adult' as the default category. Any deviations towards 'younger' or 'older' were based on textual references or the character's name, such as for 'Old Man' in *King Lear*. Older characters were occasionally classified as such based on the fact they had adult children with prominent roles in the plays.

A more global feature is the location where the scene is set. This was difficult to annotate, due to the often unreliable stage directions. Instead of a nominal description for each scene location, I used a binary annotation of 'public' and 'private'. The text itself was examined to determine the location based on what characters said about their location, but in addition Bate and Rasmussen's (2007) annotation and Greenblatt et al.'s (1997) annotations were consulted. The use of these three resources enabled the binary manual annotation of location for every scene.

Besides the information about the speaker and the scene, information regarding the addressee is essential when analysing character interaction from a conversation analysis perspective. As a manual annotation for addressee would be incredibly time consuming, I instead used an automatic method which identifies the previous speaker as the addressee of any given utterance. This is in line with the last-as-next bias used in conversation analysis (Mazeland, 2003). This means that, even in larger group conversations, it is often expected that the last speaker before the current speaker will also be the next speaker, thus making it likely that the current speaker is addressing the last speaker. If the utterances were interrupted by the start of a new scene or other stage directions (e.g., someone walking into the scene), the annotated addressee would be the *next* speaker rather than the previous speaker for the first utterance after the interruption.

Using the data for the social status of the speaker and the addressee, I also created a status differential. As the status category labels are numeric and ordered, this can be done by taking a difference. For example, a king (status = 0) and a servant (status = 6) are distant in status, and thus will have a high status differential (here: 6). Between a king and a prince (status = 1), the difference is a lot smaller (here: 1). This absolute feature was automatically generated from the already annotated features.

A feature that had to be excluded is familiarity between characters (social distance). This data was not already available, and it was beyond the scope of this study to annotate this for all relevant character pairs. The literature has shown this to be a relevant feature. However, through the use of sentiment analysis, I have attempted to cover the complimentary and insulting aspects that could arise from high familiarity, and any lack thereof arising from low familiarity. Obviously, this does not cover all aspects of familiarity, but it means that this feature is not totally neglected.

3.3. Classification based on three algorithms

Three different algorithms are used for the classification task, namely Naive Bayes, decision trees and support vector machines. Whereas it would be ideal to achieve a high precision and recall score, the main goal of this research is to see whether it is even possible to predict the second person singular pronoun choice through a computational application *at all*. If this is indeed the case, what features contribute to this prediction? It is thus more important to verify which features influence the choice and to what extent they do so.

The reason for using three algorithms, and in particular these three, is their differences in learning biases and assumptions. Naive Bayes assumes all features are independent of one another, whereas decision tree attempts to create a dependent, hierarchical structure in the features. Support vector machine (SVM) is more complex and is able to combine both dependent and independent features. The addition of the latter algorithm will be particularly useful if the difference between the two simpler algorithm's models is small.

As well as applying three algorithms, I will also look at the difference between keeping *thou* and *thee* separate and combining them into the one category THOU. For this, I will run both a binary (YOU and THOU) and a trinary (*you*, *thou* and *thee*) classification, to see whether this affects the scores or changes which features are included in the best models.

3.4. Overview of implementation

I ran the three algorithms using the Waikato Environment for Knowledge Analysis (Weka[3]) software[4] with the default settings. The algorithms were run using a 10-fold cross-validation to ensure the best model based on training and testing of all folds combined.

The number of relevant instances of *you/thou/thee* extracted from the dataset is 22,932, which makes up 99.5% of the total number of such pronouns in the dataset. The pronouns were extracted using a Python script with simple heuristics. About 0.5% was missed due to noise in the dataset. The number of instances of *you/thou/thee* that were extracted from each play range from 363 (in *Macbeth*) to 811 (in *Coriolanus*).

I attempted to improve or maintain the scores while making the model simpler by excluding features, that is, through feature ablation. When there were conflicting changes in the scores, the scores of precision and F-measure were prioritised. I hoped to identify which features truly help predict the pronoun by building the simplest but best performing model. The baseline that the models were compared to is derived from the distribution of the pronouns in the dataset, thus 62.6% of YOU and 37.4% THOU.

I first took out groups of features that are related, rather than one feature at a time. Among the 23 features, I created six different groups. The first group related to the wider linguistic and social context (play, production date, genre, scene, location), while the second group was the closer linguistic co-text (n-gram). Information on the speaker (name, status, gender, age) and the addressee (name, status, gender, age, number of people) were groups 3 and 4. I kept status differential on its own, because it relates to multiple groups. Finally, the last group was sentiment (positive and negative). After the group ablation, I went back over the features to see if individual feature exclusions would improve the model further. This ensured the simplest and best model for each algorithm. The scores and the features included in each model are given in Tables 2, 3 and 4.

Algorithm		Precision	Recall	F-measure	Accuracy
Baseline	Weighted Avg.	0.392	0.626	0.483	62.6417%
	you	0.626	1.000	0.770	
	thou	0.000	0.000	0.000	
	thee	0.000	0.000	0.000	
Naive Bayes	Weighted Avg.	0.826	0.826	0.826	82.64%
	you	0.880	0.885	0.882	
	thou	0.865	0.850	0.857	
	thee	0.509	0.510	0.510	
Decision Tree	Weighted Avg.	0.732	0.752	0.712	75.2093%
	you	0.738	0.960	0.835	
	thou	0.896	0.574	0.700	
	thee	0.408	0.097	0.157	
Support Vector Machine	Weighted Avg.	0.854	0.857	**0.854**	85.675%
	you	0.871	0.927	0.898	
	thou	0.919	0.836	0.876	
	thee	0.659	0.566	0.609	

Table 2: Scores for precision, recall, F-measure and accuracy for trinary pronoun prediction

4. Results

4.1. Trinary classification scores

Table 2 shows the results of the trinary classification. As can be seen, each model performed significantly better than the baseline model, on all scores. The F-measure of the best model, the support vector machine model, is highlighted in bold.

4.2. Binary classification scores

Table 3 shows the results of the best models for the binary classification. The F-measure of the best model, again the support vector machine model, is highlighted in bold. This is also the best scoring model out of all models presented in this paper.

[3] http://www.cs.waikato.ac.nz/ml/weka/.

[4] In Weka, Naive Bayes is identified as NaiveBayesMultinominal, decision tree as J48, and support vector machine as SMO.

Algorithm		Precision	Recall	F-measure	Accuracy
Baseline	Weighted Avg.	0.392	0.626	0.483	62.6417%
	YOU	0.626	1.000	0.770	
	THOU	0.000	0.000	0.000	
Naive Bayes	Weighted Avg.	0.868	0.868	0.867	86.8306%
	YOU	0.876	0.920	0.897	
	THOU	0.853	0.782	0.816	
Decision Tree	Weighted Avg.	0.818	0.818	0.818	81.8376%
	YOU	0.849	0.863	0.856	
	THOU	0.764	0.744	0.754	
Support Vector Machine	Weighted Avg.	0.872	0.873	**0.872**	87.2798%
	YOU	0.886	0.914	0.900	
	THOU	0.848	0.803	0.825	

Table 3: Scores for precision, recall, F-measure and accuracy for binary pronoun prediction

4.3. Feature comparison of the models

Overall, the final models contain similar sets of features. The exact compositions are given in Table 4. What is surprising is that the binary classification model for the decision tree is very different from the other models: it does not contain any of the words from the n-gram as a predictor, whereas the others did.

Algorithm	Type	Features included
Naive Bayes	Trinary	LW1, LW2, RW1, RW2, S_ID
	Binary	LW1, LW2, LW3, RW1, RW2, RW3, A_ID
Decision Tree	Trinary	LW1, LW2, RW1, RW2, S_ID, Stat_Diff, Neg_Sent
	Binary	Scene, S_ID, S_Gender, A_ID, A_Status, A_Age, Stat_Diff, Pos_Sent
Support Vector Machine	Trinary	LW1, RW1, S_ID, S_Age, A_ID, A_Age, A_Number, Stat_Diff, Pos_Sent, Neg_Sent
	Binary	LW1, RW1, S_ID, S_Age, A_ID, A_Age, A_Number, Stat_Diff, Pos_Sent, Neg_Sent

Table 4: Features included in the best model of each algorithm[5]

5. Discussion

This study has given some new insights into the analysis of pronominal address terms. Looking at the second person singular pronoun choice as a binary and a trinary classification problem resulted in slightly different outcomes. Even though the highest scores were achieved in the binary classification, one might still wonder whether this is the best method for addressing the second person singular pronoun choice. Looking back at prior studies on pronoun interpretation and comparing them to the features used in this study, we can conclude that *thee* and *thou* are equal in their opposition to *you*, with the main difference being their grammatical role. From the model comparison, we have seen that the co-text is most important when predicting the pronoun. This is evidence of the purely grammatical difference between *thou* and *thee* and their overall similarity in other aspects. Therefore, both linguistically and computationally, it makes more sense to perform a binary classification.

Differences between the algorithms were observed, but all three algorithms easily outperformed the baseline. The support vector machine models performed best, but the scores for the Naive Bayes models were quite similar to those for the SVM models. A choice between these approaches could be based solely on the scores for accuracy, precision, recall and F-measure, or also by taking into account the complexity, which is significantly higher for the support vector machine models. The more nuanced models that the support vector machine creates, which include more features than the models of the other algorithms, may suggest that the extra complexity of SVM models is indeed beneficial.

The best predicting features were the LW and RW features, which supports the importance of the direct linguistic co-text. In particular RW1 appeared as the most important feature in predicting the second person singular pronominal address term. Other important features were the speaker's name, addressee's name, status differential, positive sentiment and negative sentiment, with additional support from the speaker's gender, addressee's status, addressee's age, speaker's age, and number of people addressed. Only six features were not included in any of the models: genre, play, production date, location, speaker's status and addressee's gender.

I am, then, now able to falsify the null-hypothesis that it is not possible to build a reliable prediction model based on linguistic and extra-linguistic features. All six models demonstrate that

[5] Acronyms used as laid out in Table 1.

linguistic and extra-linguistic features substantially improve the prediction of the pronominal address term, as all six outperform the baseline.

The second hypothesis, about which features would be good predictors, was partially correct in predicting that social status, age and sentiment would be included in the best models. However, none of these features were the main predictor of pronoun choice; that was the immediate co-text.

With regard to the final hypothesis, it has been revealed that the features are indeed both dependent on and independent of each other. However, since the Naive Bayes models perform almost identically to the support vector machine models, we can say that the features are, for the most part, independent of one another.

6. Conclusion

The primary finding of this study is that it is indeed possible to build a prediction model for the use of YOU versus THOU with a singular referent in the plays of Shakespeare that is based on linguistic and extra-linguistic features. Moreover, in particular, the direct linguistic co-text of the second person singular pronoun is important. Other important features include the speaker's and addressee's names, status differential and both positive and negative sentiment. All in all this suggests that the pronoun choice is influenced by several linguistic and extra-linguistic features.

The best scoring algorithm and model was the support vector machine with 87.3% accuracy through its binary classification model.

For future research, I would recommend an exploration of other algorithms and features that were left out of this study, such as morphology, word embeddings and POS-tags. This will help us gain more information about the linguistic co-text directly surrounding the second person singular pronoun, which will likely give more insight into why this direct co-text is so important in deciding the choice of YOU or THOU. Moreover, including familiarity between characters (social distance) as a feature would be beneficial, as this has been mentioned multiple times in prior research as an influential factor, but was beyond the scope of this study.

7. References

Jonathan Bate and Eric Rasmussen (eds.). 2007. *William Shakespeare: Complete works*. The Royal Shakespeare Company, London.

Roger W. Brown and Albert Gilman. 1960. The pronouns of power and solidarity. In: T.A. Sebeok, ed., *Style in language*, pages 253–276. MIT Press, Cambridge.

Beatrix Busse. 2006. *Vocative constructions in the language of Shakespeare*. John Benjamins, Amsterdam.

Ulrich Busse. 2003. The co-occurrence of nominal and pronominal address forms in the Shakespeare Corpus: Who says thou or you to whom?. In: I. Taavitsainen and A.H. Jucker, eds., *Diachronic perspectives on address term systems*, pages 193–221. John Benjamins, Amsterdam.

Ulrich Busse. 2002. *The function of linguistic variation in the Shakespeare corpus: A corpus-based study of the morpho-syntactic variability of the address pronouns and their socio-historical and pragmatic implications*. John Benjamins, Amsterdam.

Clara Calvo. 1992. Pronouns of address and social negotiation in As You Like It. In: *Language and Literature*, Vol. 1(1), pages 5–27. Longman Group UK Ltd, London.

Stephen Greenblatt, Walter Cohen, Jean E. Howard, and Katherine E. Maus. 1997. *The Norton Shakespeare: Based on the Oxford edition*. W.W. Norton & Company, New York.

Harrie Mazeland. 2003. *Inleiding in de conversatieanalyse*. Coutinho bv, Bussum.

Gabriella Mazzon. 2003. Pronouns and nominal address in Shakespearean English: A socio-affective marking system in transition. In: I. Taavitsainen and A.H. Jucker, eds., *Diachronic perspectives on address term systems*, pages 223–249. John Benjamins, Amsterdam.

Peter Quennell and Hamish Johnson. 2002. *Who's who in Shakespeare*. Routledge, London.

Randolph Quirk. 1974. Shakespeare and the English language. In: R. Quirk, *The linguist and the English language*, pages 46–64. Edward Arnold, London.

Dieter Stein. 2003. Pronomial usage in Shakespeare: Between sociolinguistics and conversation analysis. In: I. Taavitsainen and A.H. Jucker, eds., *Diachronic perspectives on address term systems*, pages 251–307. John Benjamins, Amsterdam.

Irma Taavitsainen and Andreas H. Jucker. 2003. Introduction. In: I. Taavitsainen and A.H. Jucker, eds., *Diachronic perspectives on address term systems*, pages 1–25. John Benjamins, Amsterdam.

Mike Thelwall, Kevan Buckley, Georgious Paltoglou, Di Cai and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), pages 2544–2558.

Terry Walker. 2003. *You* and *thou* in Early Modern English dialogues: Patterns of usage. In: I. Taavitsainen and A.H. Jucker, eds., *Diachronic perspectives on address term systems*, pages 309–342. John Benjamins, Amsterdam.

Primerjava luščilnikov terminologije Sketch Engine in CollTerm za znanstvena besedila

Klara Eva Kukovičič

Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
klara.eva@gmail.com

Povzetek

Namen članka je primerjati luščilnika terminologije SketchEngine in Collterm na primeru znanstvenih besedil. Natančna terminološka baza ali vsaj enojezični seznami terminov nam lahko pri procesu prevajanja znanstvenih in strokovnih besedil prihranijo veliko časa, zato je pomembno, da njihovo izdelavo vključimo v proces analize besedila oziroma predprevajanja. Samodejni luščilniki terminologije in kolokacij pri tem igrajo pomembno vlogo, saj je gradnja jezikovnih baz z njihovo pomočjo hitrejša in enostavnejša. V prispevku smo raziskovali, kateri izmed luščilnikov terminologije iz izbranih doktoratov iz Korpusa akademske slovenščine izlušči bolj relevantne termine. Pričakujemo, da bosta luščilnika izluščila različno število terminov, prav tako pa bo njihova razporeditev glede na ključnost različna.

Comparison of Sketch Engine and CollTerm extracion tools for scientific texts

The purpose of this article is to compare two tools for automatic term extraction, Sketch Engine and CollTerm for scientific texts. A termbase or monolingual lists of terms can save translators a lot of time in the process of translating scientific texts. For this reason, it is very important to include them in the process of text analysis or pre-translation. Here, automatic term extraction tools play an important role, as their functions help us creating bases faster and simpler. In this article, we will focus on two extraction tools and research which of the tools extract more for translators' relevant terminology from the Slovene corpus KAS.

1. Uvod

Z razvojem interneta se je prevajalski proces drastično spremenil. Tiskane slovarje in slovnice so nadomestili spletni viri, med katerimi so nepogrešljivi spletni slovarji, korpusi, tezavri, razni spletni portali, kot sta Fran[1] in Terminologišče[2] in mnogi drugi. V devetdesetih letih prejšnjega stoletja so se začela razvijati tudi prevajalska namizja, med katerimi velja omeniti vodilne SDL Trados Studio[3], MemoQ[4] in Memsource[5], MateCat[6]. Njihova skupna značilnost je pomnilnik prevodov, običajno pa ta orodja ponujajo še podporo za prevajanje različnih datotečnih formatov, vtičnik za strojno prevajanje, orodje za samodejno preverjanje kakovosti, orodja za vzporejanje dokumentov in seveda upravljanje terminologije.

Prevajalci porabijo približno 30-60 % časa prevajanja zgolj za iskanje ustrezne terminologije (Gornostay et al, 2010). Ravno zato je za prevajalce zelo pomembna terminološka baza, v katero lahko vnaprej ali sproti vpisujejo terminološke vnose. Pomemben korak pri izdelavi terminološke baze predstavlja priprava enojezične baze terminov, ki skrajša čas izdelave terminološke baze (Ponikvar, 2002: 19). Enojezično bazo terminov pa najenostavneje naredimo z orodjem za luščenje terminologije, kot so na primer SketchEngine, CollTerm, LUIZ, Lexterm, SDL MultiTerm Extract.

Različni luščilniki terminologije lahko temeljijo na jezikoslovnem, statističnem ali hibridnem pristopu, ki je kombinacija prvih dveh. Prav tako ima vsak luščilnik vgrajena pravila, po katerih izlušči termine.

V tem prispevku bomo z uporabniškega vidika primerjali dva luščilnika terminologije, in sicer SketchEngine in Collterm, na kratko predstavili raziskovalno področje, potek in rezultate analize ter težave, na katere smo naleteli ob analizi.

2. Namen članka

Namen tega prispevka je na kratko predstaviti Korpus akademske slovenščine (KAS), iz katerega smo črpali gradivo za analizo, področje terminologije, luščenja terminov in orodij za samodejno luščenje terminologije, Sketch Engine in Collterm, ter ugotoviti, katero izmed orodij ponudi boljši oziroma uporabnejši nabor terminov. Naročniki prevodov običajno težijo k čim prejšnji oddaji prevoda, zato smo prevajalci pogosto pod časovnim pritiskom in si ne vzamemo dovolj časa, da bi pregledali nabor vseh terminoloških izpisov, ki nam jih ponudijo luščilniki. Posledično smo se tudi v naši raziskavi odločili, da se osredotočimo zgolj na prvih 100 kandidatov, ki sta jih iz izbranih doktoratov Korpusa akademske slovenščine izluščila Sketch Engine in CollTerm.

3. Korpus akademske slovenščine

Projekt »Slovenska znanstvena besedila: viri in opisi« se je začel leta 2016 kot odgovor na cilje Akcijskega načrta za jezikovno izobraževanje (2015) in Akcijskega načrta za jezikovno opremljenost (2015), ki sta ugotovila, da je potrebno razvijati slovenščino v visokem šolstvu in znanost ter izboljšati položaj slovenščine kot jezika znanosti. V okviru projekta je bil ustvarjen korpus pisnih besedil akademske slovenščine (Erjavec et al.,: 2016).

Korpus akademske slovenščine vsebuje skoraj 1,2 milijarde pojavnic, največji delež korpusa (81,14 %) predstavljajo diplomska dela, sledijo magistrska dela (12,60 %) in doktorska dela (1,38 %) (Erjavec et al., 2016).

[1] https://fran.si/

[2] https://isjfr.zrc-sazu.si/terminologisce#v

[3] https://www.sdltrados.com/

[4] https://www.memoq.com/en/

[5] https://www.memsource.com/

[6] https://www.matecat.com/

V korpusu so najbolj zastopana besedila iz družboslovja (57 %), tehnoloških ved (32 %) ter naravoslovno-matematičnih ved (9 %) (Erjavec et al., 2016).

4. Termini in terminologija

Ključni element znanstvenih in akademskih besedil predstavlja terminologija, s katero se v prvi vrsti ukvarjajo terminologi in lingvisti (Fišer et al., 2016: 35). Po načelih terminološke vede naj bi bili termini enote, ki so pomensko predvidljive in ustaljene. Z razvojem računalniškega pristopa k pridobivanju terminologije pa se je izkazalo, da se termini dinamično spreminjajo v odvisnosti od besedilnih dejavnikov in da niso skladenjsko in oblikoslovno usklajene enote (Vintar, 2009: 347). Termine lahko definiramo tudi kot jezikovne znake, ki označujejo pojem in se po svojih lastnostih razlikujejo od druge leksike (Fajfar et al., 2015).

4.1. Subjektivno dojemanje terminološkosti

O subjektivnosti dojemanja terminologije je bilo do sedaj narejenih že več raziskav. Ena izmed obsežnejših je bila izvedena v okviru doktorske dizertacije, kjer je R. Estopà Bagot analizira, kako različne skupine uporabnikov terminologije dojemajo terminološkost (Vintar, 2008: 47-49). Rezultati so pokazali velika odstopanja med prevajalci, dokumentalisti, strokovnjaki in terminografi (Vintar, 2008: 47-49). Največ enot so za termine označili terminografi (1052), strokovnjaki so označili 938 izrazov, dokumentalisti 486 izrazov, najmanj enot pa so kot termine označili prevajalci (270) (Vintar, 2008: 47-49).

Podobne raziskave pa so bile narejene tudi v slovenskem prostoru (Fajfar et al., 2015: 8). Rezultati raziskave, ki je potekala v okviru projekta TERMIS, so pokazali, da so študenti v analiziranih besedilih označili manj besed za termine, kot so pričakovali avtorji raziskave (Logar Berginc 2013: 248). Avtorice druge raziskave, v katero sta bila vključena dva strokovnjaka s področja odnosov z javnostmi, pa v sklepu prav tako ugotavljajo, da neujemanje rezultatov obeh udeležencev »jasno kaže subjektivnost same definicije terminološkosti« (Logar Berginc et al., 2013: 132-133).

4.2. Orodja za samodejnost luščenje terminologije

Vintar (2009: 345) samodejno luščenje terminologije definira kot postopek, pri katerem program na podlagi statističnih izračunov, jezikoslovnih analiz ali drugih obstoječih podatkov skuša ugotoviti, katere besede in besedne zveze v danem korpusu strokovnih besedil so terminološke. Področje procesiranja naravnih jezikov se že več kot 20 let aktivno ukvarja z luščenjem terminologije, ki lahko temelji na statistični ali jezikovni metodi ali kombinaciji obeh (Pinnis et al., 2012: 193), danes pa se s pridom uporabljajo tudi metode strojnega učenja.

4.2.1. Sketch Engine

Sketch Engine je spletno orodje, ki uporabnikom omogoča brskanje in analiziranje že obstoječih in ustvarjanje lastnih korpusov (Fišer et al., 2016: 136). Orodje prav tako omogoča samodejno luščenje terminologije. Za luščenje terminologije z orodjem Sketch Engine potrebujemo dva korpusa, in sicer prvega, iz katerega želimo izluščiti kandidate, in drugega, tj. referenčnega korpusa, ki mora biti čim večji. Orodje nato primerja besedišče prvega korpusa z besediščem drugega, referenčnega korpusa, in na podlagi pravil izlušči kandidate.

Proces luščenja z orodjem Sketch Engine je dvostopenjski. Prvi korak temelji na pravilih in je odvisen od jezika. V tem koraku se z uporabo tako imenovane slovnice terminov (ang. term grammar) oceni slovnična veljavnost določene zveze v specializiranem korpusu. V drugem koraku pa se kandidate, ki smo jih dobili v prvem koraku, primerja z referenčnim korpusom z uporabo »simplemath« statistike (Fišer et al, 2016: 36). Slovnico terminov za slovenščino so na podlagi češke predloge razvili Fišer et al. (2016). Za slovnico terminov uporabljamo poizvedovalni jezik Corpus Query Language (CQL). Zapis slovnice terminov si lahko ogledamo na naslednjih primerih. Najprej je zapisan primer slovnice za dvobesedne termine, nato pa še za tri- in štiribesedne:

```
1:adj 2:noun & agree(1,2)
*COLLOC "ll_%(1.gender_lemma)_%(2.lemma_lc)-x"
```

V prvi vrstici je pravilo, ki ga orodje uporablja za prepoznavanje vzorca, v drugi pa pravilo, kako naj se prepoznani termin izpiše. Primer ponazarja prepoznavo dvobesednih zvez pridevnika in samostalnika, v katerih se spol pridevnika ujema s spolom samostalnika. Program prepozna vse besedne zveze v korpusu, ki ustrezajo temu pravilu. Nato s pomočjo statistike izvede rangiranje prepoznanih zvez glede na rezultate njihove terminološkosti in jih izpiše v skladu s pravilom v drugi vrstici. Za zgornji primer je navodilo tako, da se samostalnik izpiše kot lema z malo začetnico, pridevnik pa v lemi, ki se po spolu ujema s samostalnikom. Primer izluščenega termina, ki ustreza temu pravilu, je npr. *prosti delec.*

Primer slovnice terminov za tribesedne termine:

```
1:adj 2:noun 3:noun_genitive
*COLLOC
"z_%(1.gender_lemma)_%(2.lemma_lc)_%(3.lc)-x"
```

Zgoraj zapisani primer torej pove, da program v seznamu kandidatov izpiše termine, ki so sestavljeni iz zveze pridevnika, samostalnika in samostalnika v rodilniku, pri čemer se spol pridevnika ujema s spolom samostalnika. Primer takšnega termina je na primer lahko *samodejno prepoznavanje glasu*.

Primer slovnice terminov za štiribesedne termine:

```
1:noun 2:adj_genitive 3:noun_genitive 4:noun_genitive &
agree(2,3)
*COLLOC  "d_%(1.lemma_lc)_%(2.lc)_%(3.lc)_%(4.lc)-
x"
```

Primer štiribesednega termina, ki ustreza zgornjemu pravilu, je npr. *gostota prostih nosilcev naboja*.

4.2.2. CollTerm

O CollTerm je orodje za luščenje kolokacij in terminologije, ki je bilo razvito v okviru projekta ACCURAT in je prosto dostopno. Orodje so razvili Pinnis et al. (2012). Predhodna različica orodja je terminološke kandidate luščila zgolj s pomočjo oblikoskladenjskih vzorcev in z različnimi statistikami za sopojavitev besed oziroma besednih zvez. Danes je orodje nadgrajeno z modulom za nadzorovano učenje. To omogoča, da orodje na koncu poda en sam seznam rangiranih terminoloških kandidatov in ne po en seznam za vsako stopnjo n-gramov kot prej.

CollTerm uporablja enake oblikoskladenjske vzorce kot Sketch Engine, prilagojene za slovenščino. Proces luščenja poteka v štirih korakih, in sicer:

1. morfosintaktično filtriranje,
2. filtriranje za minimalno pogostost
3. statistično rangiranje glede na referenčni korpus in
4. filtriranje na podlagi rezultata za rangiranje oz. uvrstitve na rangiranem seznamu.

Orodje CollTerm tako kot Sketch Enigne omogoča, da shranimo izpis terminoloških kandidatov, kjer sta razvidni tako ključnost kot tudi pogostost pojavitve. Višja kot je ključnost, bolj relevanten je za nas izpisan terminološki kandidat.

Orodje CollTerm za razliko od Sketch Engina pri izpisu terminoloških kandidatov zapiše obliko kandidata v prvem sklonu in v sklonu, v katerem se največkrat pojavi. Če se torej kandidat največkrat pojavlja v četrtem sklonu (npr. umetno inteligenco), ga orodje izpiše tako v tožilniku kot tudi v imenovalniku, pri čemer pa lahko pride do neujemanja spola samostalnika s pridevnikom (npr. umeten inteligenca).

Obe orodji rangirata kandidate glede na »keyness score« oziroma ključnost. Višja kot je ključnost, višje se pojavi kandidat.

5. Zasnova projekta

Delo smo začeli s pregledom doktoratov, vključenih v korpus KAS. Želeli smo izbrati terminološko in tematsko čim bolj podobne doktorate, zato smo se odločili, da se osredotočimo na področje računalništva. S seznama več kot sedemsto doktoratov smo nato izbrali vse tiste, ki so bili napisani na Fakulteti za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru, saj korpus ne vsebuje doktoratov s Fakultete za računalništvo in informatiko, Univerze v Ljubljani. Ker je bil tudi teh doktoratov preveliko, smo jih nato izbrali glede na tematiko, o kateri so pisali. Za temo smo določili umetno inteligenco in dobili precej manjši nabor, iz katerega smo nato določili tri doktorate, ki so si bili vsebinsko najbolj podobni.

Za vsak izbrani doktorat iz KAS je nato sledila priprava podkorpusa, ki smo ga izdelali z orodjem Sketch Engine. Sledil je proces luščenja terminoloških kandidatov, ki je potekal anonimno, saj smo želeli pri analizi ohraniti čim večjo objektivnost. Pri luščenju terminologije smo se obrnili na doc. dr. Darjo Fišer, Filozofska fakulteta Univerze v Ljubljani, in dr. Nikolo Ljubešića, Institut Jožef Štefan, ki sta z orodjem Sketch Engine in CollTerm izluščila kandidate treh izbranih doktoratov. Pri orodju Sketch Engine smo pri možnostih za izpis terminoloških kandidatov iz podkorpusov izbrali možnost »Keywords« oziroma ključne besede, kot referenčni korpus pa smo tako pri obeh orodjih določili celotni korpus KAS. Ker se je izkazalo, da Sketch Engine omogoča luščenje zgolj večbesednih kandidatov, smo tudi v orodju CollTerm omejili izpis na večbesednih kandidatov.

Seznami s terminološkimi kandidati niso vsebovali podatkov o orodju, ki je bilo uporabljeno, s čimer smo pri evalvaciji ohranili objektivnost. Sezname kandidatov smo na začetku označili kot Seznam orodja 1 in Seznam orodja 2. Šele po koncu analize smo preverili, katero orodje je bilo zapisano pod številko 1 in katero pod številko 2.

Naslednji korak je bilo označevanje kandidatov. Pri tem smo uporabljali Navodila za ocenjevanje terminoloških kandidatov, ki so v projektu KAS že služila kot pripomoček in navodilo za označevanje terminov. Smernice vsebujejo pet kategorij, in sicer termin, izvenpodročni termin, nerelevantno, znanstveno pisanje in »ne vem«. Kategorija termin vključuje besede in besedne zveze, ki predstavljajo termine z določenega področja. V kategorijo izvenpodročni termini so bile uvrščene besede in besedne zveze, ki so po oceni označevalca termini z drugega področja. Kategorija znanstveno pisanje zajema izrazje, vezano na pridobivanje, analizo ali predstavitev podatkov v doktorski raziskavi (npr. *tabela*) in besedišče, ki je stalni del doktorskega pisanja ali strukture doktorskega dela (npr. *zaključek*). V kategorijo nerelevantno uvrščamo splošno besedišče (npr. *celota, v bistvu*), zveze, ki so daljši opisi ali ki zahtevajo razpravljalno definicijo oz. so subjektivne narave in besede ali besedne zveze, ki so le del termina ali pa poleg termina vsebujejo še druge elemente (npr. *model govorca* – prekratko, termin je namreč *univerzalni model govorca* in *grozdenje govorcev* – predolgo, termin je namreč zgolj *grozdenje*). Označevalec kandidate po potrebi preverja v kontekstu korpusa KAS s pomočjo konkordanc. Kandidate, ki jih tudi s pomočjo konteksta ni mogoče uvrstiti v eno od kategorij, označimo z oznako »ne vem«.

Z uporabo istih smernic tudi v pričujoči raziskavi smo zagotovili večjo konsistentnost vrednotenja terminov. Izpustili smo zgolj kategorijo »ne vem«, saj bi ta kategorija izkrivila rezultate naše analize. Evalvacijo luščilnikov terminologije sem opravila avtorica prispevka, podiplomska študentka prevajanja.

Kot pomoč pri odločanju, ali je izluščena beseda oz. besedna zveza termin s področja računalništva ali informatike, pa so mi poleg lastnega znanja služili naslednji viri:

a) Islovar: terminološki, razlagalni in informativni slovar, ki strokovno izrazje pomensko in jezikovno opisuje in vrednoti. Zajema izrazje informatike, informacijske tehnologije in telekomunikacij ter drugih računalniških področij. Slovar ne vsebuje besed splošnega izrazja (Islovar).
b) DIS slovarček: slovar računalniških izrazov, verzija 2.1.71, ki je nastal na oddelku za inteligentne sisteme na Inštitutu Jožef Štefan. Slovarček trenutno vsebuje 12.296 gesel (DIS slovarček).
c) English/Slovene dictionary of computer science: spletni slovar računalništva, ki je nastal v okviru Inštituta Jožef Štefan, na voljo je tudi v tiskani verziji (English/Slovene dictionary of computer science).
d) Wikipedia: prosta spletna enciklopedija, ki vsebuje več kot 160.751 člankov (Wikipedia).
e) Prostodostopni zapiski in literatura s predavanj s Fakultete za računalništvo in informatiko.

Če smo na katerem koli viru našli definicijo terminološkega kandidata ali v slovenščini ali v angleščini, smo sklepali, da gre za termin. Glavni namen označevanja kandidata je bil priprava podatkov za analizo, v kateri smo preverjali, kateri luščilnik terminologije je izluščil več kakovostnih terminov oziroma jih je rangiral višje. Že pred začetkom analize smo pričakovali, da bosta luščilnika podala različne kandidate, zato smo se želeli osredotočiti tudi na to, ali katero izmed orodij izlušči kakšen termin, ki ga drugi ne. Končni cilj je bil predvsem preveriti, katero izmed orodij tako profesionalnim prevajalcem kot študentom na prvih stotih mestih ponudi bolj relevantne terminološke kandidate.

6. Analiza

Analiza je sestavljena iz treh delov. V prvem delu nas je zanimala zgolj frekvenca posameznih oznak za posamezen seznam izluščenih kandidatov, v drugem delu smo računali odstotek pozitivnih primerov med prvimi N kandidati, v tretjem pa smo na primeru enega izmed doktoratov preverjali, ali so termini, izluščeni s prvim luščilnikom in ki se nahajajo med prvimi stotimi kandidati prvega seznama, vključeni tudi na seznam kandidatov, izluščenih z drugim luščilnikom.

6.1. Rezultati izluščenih kandidatov

Najprej smo prešteli, koliko izluščenih terminoloških kandidatov smo označili z oznako termin, izvenpodročni termin, znanstveno izrazje ali nerelevantno.

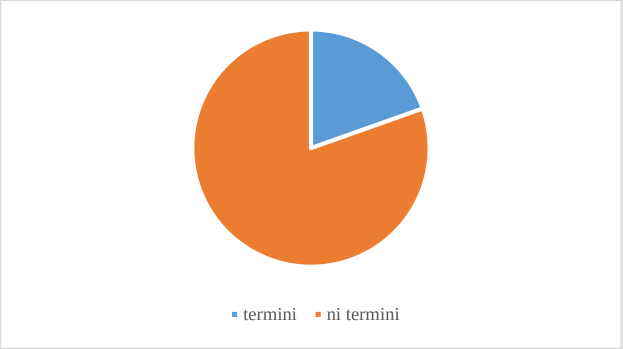
Kot je razvidno iz podatkov analize, je luščilnik CollTerm izluščil več terminov. Pri tretjem doktoratu, kjer smo s Sketch Enginom izluščili več terminov kot s CollTermom, je CollTerm izluščil bistveno več izvenpodročnih terminov. Pri drugih dveh doktoratih sta oba luščilnika izluščila enako število izvenpodročnih terminov, vendar ti niso bili isti.

	Oznaka	CollTerm	Sketch Engine
1	Termin	68	54
2	Izvenpodročni termin	13	3
3	Znanstveno pisanje	106	130
4	Nerelevantno	113	113

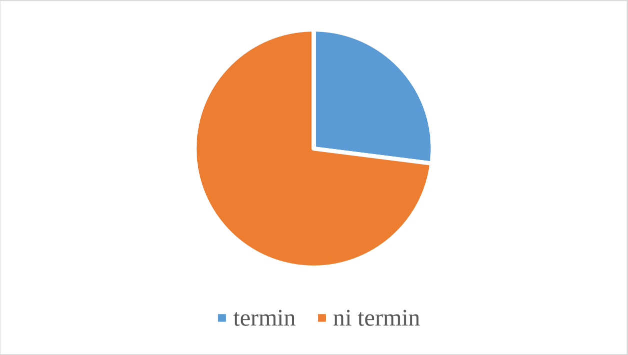
Tabela 1: Frekvenca oznak pri terminoloških kandidatih, izluščenih iz doktoratov

		CollTerm	Sketch Engine
1	Termini	81 (27 %)	57 (19 %)
2	Ne termini	219 (73 %)	243 (81 %)

Tabela 2: Število izluščenih terminov



Graf 1: Luščenje z orodjem Sketch Engine



Graf 2: Luščenje z orodjem CollTerm

Druga tabela prikazuje število in odstotek izluščenih terminov in ostalega besedišča. Vidimo, da je bilo 27 % terminoloških kandidatov, izluščenih s CollTermom, terminov. Nekoliko slabši rezultat smo dobili z orodjem Sketch Engine. Rezultati so predstavljeni tudi na Grafu 1 in 2. Graf 1 prikazuje razmerje med termini in ne termini, ki jih je izluščil Sketch Engine, Graf 2 pa razmerje, ki ga je izluščil CollTerm.

6.2. Odstotek pozitivnih primerov med prvimi N kandidati

V drugem delu analize smo se osredotočili na računanje odstotkov pozitivnih primerov med prvimi N kandidati. Kot pozitivne primere smo označili vse kandidate, ki smo jih pripisali oznako t (termin) ali x (izvenpodročni termin).

Namen drugega dela analize je bil, da ugotovimo, ali se termini na seznamu izluščenih kandidatov gosteje nahajajo na prvih mestih ali so enakomerno razporejeni čez celoten seznam. Ta del analize se nam je zdel ključen predvsem zato, ker se zavedamo, da pri prevajanju pogosto pripravimo zgolj ožji nabor terminologije, pri čemer v terminološko bazo vključimo prvih nekaj kandidatov.

6.2.1. Odstotek pozitivnih primerov med prvimi 20 kandidati

Spodnje tabele (Tabela 3, Tabela 4 in Tabela 5) prikazujejo odstotek terminov med prvimi dvajsetimi izluščenimi kandidati pri vsakem seznamu posebej. Če pogledamo povprečje vseh rezultatov, ki je prikazano v Tabeli 6, vidimo, da je med prvimi 20 kandidati, izluščenimi z orodjem CollTerm, 38 % kandidatov terminov, odstotek pa ni bistveno nižji (35 %) pri kandidatih, izluščenih s Sketch Enginom.

	Seznam 1
CollTerm	25 %
Sketch Engine	15 %

Tabela 3: Pozitivni primeri prvega seznama med prvimi 20 kandidati

	Seznam 2
CollTerm	45 %
Sketch Engine	50 %

Tabela 4: Pozitivni primeri drugega seznama med prvimi 20 kandidati

	Seznam 3
CollTerm	45 %
Sketch Engine	40 %

Tabela 5: Pozitivni primeri prvega seznama med prvimi 20 kandidati

	Povprečje vseh seznamov
CollTerm	38 %
Sketch Engine	35 %

Tabela 6: Povprečje pozitivnih primerov med prvimi 20 kandidati iz vseh seznamov

6.2.2. Odstotek pozitivnih primerov med prvimi 50 kandidati

Spodnje tabele (Tabela 7, Tabela 8 in Tabela 9) prikazujejo odstotek terminov med prvimi petdesetimi izluščenimi kandidati pri vsakem seznamu posebej. Povprečje vseh treh seznamov je prikazano v Tabeli 9. Rezultati kažejo, da se je razlika med odstotkom terminov, izluščenih z orodjema, v primerjavi z naborom 20 kandidatov, povečala. Odstopanje pri naboru prvih 20 kandidatov je bilo zgolj 3-odstotno, razlika pri naboru prvih 50 kandidatov pa se je povečala na 9 %.

	Seznam 1
CollTerm	28 %
Sketch Engine	12 %

Tabela 7: Pozitivni primeri prvega seznama med prvimi 50 kandidati

	Seznam 2
CollTerm	30 %
Sketch Engine	30 %

Tabela 8: Pozitivni primeri drugega seznama med prvimi 50 kandidati

	Seznam 3
CollTerm	36 %
Sketch Engine	24 %

Tabela 9: Pozitivni primeri tretjega seznama med prvimi 50 kandidati

	Povprečje vseh seznamov
CollTerm	31 %
Sketch Engine	22 %

Tabela 10: Povprečje pozitivnih primerov med prvimi 50 kandidati z vseh seznamov

6.3. Rangiranje izluščenih terminov

V tretjem delu analize smo se osredotočili zgolj na en doktorat. Najprej smo izpisali termine in njihovo zaporedno številko na seznamu kandidatov, ki je bil narejen z orodjem CollTerm. Nato smo preverili, pod katero zaporedno številko je isti termin rangiran na seznamu terminoloških kandidatov, ustvarjenim z orodjem Sketch Engine. Nato smo postopek zamenjali in izpisali termine in zaporedne številke terminov iz seznama kandidatov, narejenega z orodjem Sketch Engine in zaporedno številko primerjali z rangiranjem termina na seznamu kandidatov luščilnika CollTerm.

V Tabeli 11 lahko v levi koloni vidimo številko, pod katero je bil na seznamu CollTerm izluščenih kandidatov zapisan termin. Če je na primer zapisana številka 2, to pomeni, da je drugi zaporedni kandidat s seznama, ustvarjenega z orodjem CollTerm, termin. Desni stolpec za prikazuje, na katerem mestu se je isti termin pojavil na seznamu kandidatov, izluščenih z orodjem Sketch Engine. Tabela 12 pa prikazuje ravno obratno situacijo.

Namen zadnjega postopka analize je bil, da preverimo, ali sta luščilnika izluščila iste termine in kje so ti rangirani.

Zaporedna št. v CollT	Zaporedna št. v Sk E
2	30
5	21
7	9
13	205
16	67
17	0
18	135
19	**10**
21	4541
22	0
24	54
31	544
32	546
36	4727
39	40
40	4440
44	4422
50	4681
59	**7**

63	201
70	4669
71	454
72	4106
83	4594
88	0

Tabela 11: Rangiranje terminov

Zaporedna št. v Sk E	Zaporedna št. CollT
2	**185**
7	**59**
9	7
10	**19**
11	**192**
20	**195**
21	5
29	**194**
33	2
34	0
40	39
54	24
66	0
67	16
70	**219**
96	**215**

Tabela 12: Rangiranje terminov

V 11. in 12. tabeli so krepko označeni tisti termini, ki se prej pojavijo na seznamu terminoloških kandidatov, izluščenih s Sketch Enginom kot s CollTermom. Vidimo, da je sicer tudi Sketch Engine izluščil iste termine kot CollTerm, vendar so ti rangirani občutno nižje. Prav tako pa sta obe orodji izluščili nekaj terminov (Sketch Engine 2, CollTerm 3), ki jih drugo orodje ni.

Zahvala

Raziskavo, opisano v prispevku, je podprl projekt ARRS J6-7094 »Slovenska znanstvena besedila: viri in opis«.

7. Zaključek

Rezultati analize so pokazali, da se na seznamu terminoloških kandidatov obeh orodij nahaja precejšnje število terminov. Čeprav sta obe orodji v povprečju izluščili približno tretjino terminov, je orodje CollTerm pokazalo boljše rezultate oziroma je izluščilo več izrazja, relevantnega za prevajalce. CollTerm je v povprečju izluščil več terminov, ne samo med prvimi dvajsetimi in petdesetimi, ampak tudi med prvimi sto kandidati. Ob tem pa seveda moramo upoštevati, da je zaradi subjektivnega dojemanja terminologije verjetno prišlo do manjših odstopanj pri označevanju kandidatov s kategorijami: termin, izvenpodročni term, znanstveno pisanje in nerelevantno, čemur bi se v prihodnjih raziskavah dalo v dobršni meri izogniti z uporabo več označevalcev.

Analiza je prav tako pokazala, da so nekateri termini, ki nam jih ponudi CollTerm, na seznamu kandidatov, ustvarjenim z luščilnikom Sketch Engine, rangirani precej nizko ali pa jih na seznamu sploh ni.

V analizo nismo vključili podatka o tem, koliko terminov sta obe orodji spregledali (recall). Če bi želeli dobiti te podatke, bi morali ročno pregledati in označiti vsaj vzorec besedil, kar sicer presega okvire te raziskave, je pa v naših načrtih za prihodnje delo.

Poznavanje obeh orodij vsekakor nudi odlično podporo pri prevajalskem procesu. Čeprav so rezultati naše raziskave pokazali, da CollTerm prevajalcem ponudi ustreznejše jezikovne elemente, je odločitev, katero orodje bo prevajalec izbral, odvisna tudi od drugih faktorjev, kot sta na primer dostopnost in enostavnost uporabe, kjer ima orodje Sketch Engine v tem trenutku nekaj pomembnih prednosti.

8. Literatura

Akcijski načrt za jezikovno izobraževanje. 2015. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/raziskave-analize/slovenski_jezik/Akcijska_nacrta/ANJI.pdf (Dostop: 25. 3. 2018).

Akcijski načrt za jezikovno opremljenost. 2015. http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/Razpisi/2017/JR-ESS-Ranljive_skupine_govorcev/Akcijski_nacrt_za_jezikovno_opremljenost.pdf (Dostop: 25. 3. 2018).

DIS slovarček http://dis-slovarcek.ijs.si/ (Dostop: 10. 4. 2018).

Darja Fišer, Vit Suchomel, Miloš Jakubiček. 2016. Terminology Extraction for Academic Slovene Using Sketch Engine. RASLAN 2016.

English/Slovene dictionary of computer science https://www.ijs.si/cgi-bin/rac-slovar? (Dostop: 10. 4. 2018).

Islovar. http://www.islovar.org/islovar (Dostop: 10. 4. 2018).

Mārcis Pinnis, Nikola Ljubešić, Dan Ştefănescu, Inguna Skadiņa, Marko Tadić, Tatiana Gornostay. 2012. Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. Proceedings of RASLAN 2016.

Meselina Ponikvar. 2002. Računalniška podpora prevajalskemu in terminološkemu delu na primeru prevajanja v okolju sektorja za prevajanje SVEZ. Diplomsko delo, Ljubljana, Univerza v Ljubljani, Filozofska fakulteta.

Nataša Logar Berginc, Špela Vintar, Špela Arhar Holdt, Terminologija odnosov z javnostmi: korpus – luščenje – terminološka podatkovna zbirka, Slovenščina 2.0 1 (2013), št. 2 = Jezikovne tehnologije, ur. Tomaž Erjavec – Jerneja Žganec Gros, 113-138.

Sketch Engine Documentation https://www.sketchengine.eu/documentation/writing-term-grammar/#file (Dostop: 10. 8. 2018).

Špela Vintar. 2009. Samodejno luščenje terminologije – izkušnje in perspektive. V Terminologija in sodobna terminografija, ur. Nina Ledinek, Mojca Žagar in Marjeta Humar, str. 345-356. Ljubljana, Založba ZRC, ZRC SAZU.

Tanja Fajfar, Mojca Žagar Karer: Strokovnjaki in prepoznavanje terminov v strokovnih besedilih. Jezikoslovni zapiski, 21/1, 2015, str. 7-21.

Tatiana Gornostay, Andrejs Vasiljevs, Signe Rirdance, Roberts Rozins. 2010. Bridging the Gap – EuroTermBank Terminology Delivered to User´s Environment, Proceedings of the 14th Annual Conference of the European Association for Machine Translation.

Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar, Milan Ojsteršek. 2016. Slovenska akademska besedila: prototipni korpus in načrt analiz. Zbornik konference Jezikovne tehnologije in digitalna humanistika https://core.ac.uk/download/pdf/143471580.pdf (Dostop 5. 4. 2018).

Wikipedia https://sl.wikipedia.org/ (Dostop: 10. 4. 2018).

K-means Clustering for POS Tagger Improvement

Gabi Rolih

Department of Linguistics and Philology, Uppsala University
Engelska parken, Thunbergsv. 3H, 751 26 Uppsala
gabirolih@gmail.com

1. Introduction

With the emergence of social media, the internet is turning into an enormous openly accessible corpus with diverse language and domain selection that keeps growing. Using this data would bring many benefits to natural language processing tasks. However, the language use on the internet differs greatly from the standard language due to inconsistent capitalization, omission of diacritics, non-standard spelling and colloquial expressions. The currently available tools in natural language processing (NLP) are not created for dealing with such cases and are therefore not adequate for such texts. Much of the latest research was focusing on creating specialized tools that would be able to deal with non-standard language.

2. Goal of the paper

In this article, we present an implementation of K-means clustering on CMC (computer-mediated communication) data in order to further improve a state-of-the-art tagger for Slovenian. We evaluate the tagger against a human annotated dataset and compare the results with previous work.

In section 2, previous work on this problem is presented. In sections 3, 4 and 5, our datasets, tools and implementation are outlined. In section 6 the performance is evaluated and discussed. In section 7 the article is concluded with future work suggestions.

3. Previous work

Researching CMC data has been in focus in the more recent years.

Eisenstein (2013) analyzes how language is used on the internet and how the NLP community typically deals with these problems. He concludes that there are two standard approaches to NLP tasks for CMC data: normalization and domain adaptation. Normalization is converting the non-standard language into a standard language. This includes changing capitalization, punctuation, spelling and in some cases even changing words into their more formal counterparts. Domain adaptation means adapting already existing tools to a new domain, which in this case is the non-standard language.

Ljubešić et al. (2017) run various experiments to adapt the ReLDI tagger, a state-of-the-art tagger for Slovene, to CMC data. They do this by retraining the tagger on CMC data, using an inflectional lexicon, adding normalization data and using clustering information. One approach that yields significant results is creating Brown clusters (Brown et al., 1992) on CMC data and using that to retrain the tagger. For our research we use the same tools and approaches, except that we use K-means clustering technique instead of Brown clusters.

Brown clusters are a popular choice for word clustering because they are efficient and scale well to large datasets. Turian et al. (2010) evaluate Brown clusters, Collobert and Weston embeddings and hierarchical log-linear (HLBL) embeddings for named-entity recognition (NER) and chunking tasks and confirm that Brown clusters perform best.

Owoputi et al. (2013) successfully use Brown clusters to improve PoS tagging in online conversational texts. They construct a state-of-the-art tagger for Twitter and IRC texts with accurracy above 90.

K-means clustering is a simple and very popular clustering algorithm, but it is not very common in NLP tasks. Lin and Wu (2009) attempt to use K-means on word phrases and use them for NER tasks. Their system achieves the best result for NER systems at the time. They argue that using the cluster on phrases rather than words brings better results.

4. K-means clustering

K-means clustering is a technique proposed by MacQueen et al. (1967). It splits the data into a number of clusters based on the proximity of points to the cluster center. The number of clusters K must be provided as input.

K-means is an iterative algorithm, where each iteration consists of two steps. Step one is assigning clusters to points. For each point, K-means calculates the distances from the point to the cluster centroids. The point is then assigned to the closest cluster centroid. When all points are assigned a cluster, K-means proceeds with step 2. In this step it readjusts the centroids. That is done by averaging all data points belonging to the same cluster. The average is the new cluster centroid. The process is then repeated until conversion.

The main struggle with K-means is choosing the number of clusters before initializing the algorithm. In many cases, it is impossible to know how many clusters we need. However, for this experiment we create the same number of clusters as in Ljubešić et al. (2017), which is 2000 clusters.

5. Word2Vec

In order to use words as input for K-means, they need to be represented by vectors. A very efficient model for that is Word2Vec presented by Mikolov et al. (2013). Word2Vec uses a single layer of a feed-forward neural network. For input, words are encoded as vectors with one-hot representation. Then the context window (which is determined in forehand) is observed and the hidden layer calculates weights that determine the probability of one word co-occuring with some (or more) other words. The intuition behind this is that similar words appear in similar contexts. The output is a feature matrix of words. Word2Vec is typically used to predict the next word, but in this case we use it to create the proper input for K-means clustering.

6. Implementation

6.1. Dataset

The dataset used for clustering is the slWaC v2.0, a corpus of web Slovene (Erjavec et al., 2015), which comprises of 1.2 billion tokens. It also contains lemma and morphosyntactic annotations that were not used for clustering.

For tagger training and testing, the Janes-Tag v1.2 dataset (Fišer, 2016) was used. The training portion consists of 60,367 tokens and the test portion of 7,484 tokens. The development portion was not used for this experiment. The data was tokenized and converted into one sentence per line, which was needed as input.

6.2. Clustering

The first step before clustering is converting words into vectors by Word2Vec. This is done by the Gensim library (Řehůřek and Sojka, 2010). We feed our dataset into Word2Vec continuous bag of words (CBOW) model, where we set some additional parameters. As in (Ljubešić et al., 2017), we only consider words with frequency count above 50. The default window size for English is 5, but we descrease that to 2 to capture more syntactic relation and not semantic. The other parameters take their default values[1]. We then use the Scikit-learn package (Pedregosa et al., 2011) to implement K-means clustering. We create 2000 clusters, which takes roughly 2 days.

6.3. Integration with the tagger

We use the cluster information to improve ReLDI tagger (Ljubešić et al., 2017), but in order to evaluate it correctly we also train it on CMC data. We replace the Brown clusters by K-means clusters. This takes some adaptation, because Brown clustering is hierarchical and the clusters are included together with binary paths for easier search. This is something that K-means clustering does not have because it is not hierarchical. However, the binary paths should not affect performance, only computing time, so we simply add the binary paths to K-means clusters.

The tagger is then trained on these clusters and the training portion of Janes-Tag. Training takes roughly 6 hours.

7. Results

The tagger is evaluated on the test portion of Janes-Tag in two ways: the complete morphosyntactic description (MSD) and only first two labels in description (PoS). We calculate accuracy for both these sets because that is the typical evaluation metric for classification models. Results are compared to the baseline performance (ReLDI retrained on CMC data) and to the Brown clusters model. The results are available in Table 1 below.

[1] Parameter configuration: *size=100, alpha=0.025, window=2, min_count=50, max_vocab_size=None, sample=1e-3, seed=1, workers=3, min_alpha=0.0001, sg=1, hs=0, negative=5, cbow_mean=1, hashfxn=hash, iter=5, null_word=0, trim_rule=None, sorted_vocab=1, batch_words=MAX_WORDS_IN_BATCH.*

	baseline	Brown	K-means
MSD	84.15	85.17	88.32
PoS	89.85	91.12	92.88

Table 1: Comparison of the tagger accuracy between our model (K-means) and models from Ljubešić et al. (2017).

Our model improves the baseline by 4.17% on the MSD set and 3.03% on the PoS set. It also outperforms the Brown clusters slightly. The results are unexpected, because Brown clusters had been analyzed often in research works and proved to be the most efficient. However, Brown clustering works on the intuition that similar words appear in similar contexts, and Word2Vec has that same intuition. This might be the cause of such high performance of K-means.

Especially interesting is the result for the MSD set, because high accuracy seemed hard to achieve in Ljubešić et al. (2017). Our result, although still lower than the PoS result, could be useful for morphologically rich languages that require many tags to describe morphology.

Error analysis shows that our tagger performed well in determining many tags, but might have failed in the final tag or two of the full morphological description. Even though this information is not complete, it would still provide some useful information.

In the PoS set, the most problematic category was nouns. Common nouns account for 20% of the errors, while proper nouns account for 12%. This is not unexpected, as this category is also the most diverse and might contain many words that were not seen in training. Other categories with the greatest error margin were general adverbs with 10% and general adjectives with 11%. There were many words that were annotated as elements from another language, but were mostly recognized as nouns by the tagger (10% of errors). This category is also problematic from another point of view: The words in it belong to two categories simultaneously, as they are elements from a different language, but they also belong to some word class. In these situations K-means cannot be of great help, because it is a hard-clustering technique, which means that a single word only belongs to one cluster. To further improve this, a soft clustering technique would be required.

In the MSD set, the most common errors were common nouns (masculine, singular, nominative case) and general adverbs (positive degree). These groups both account for 6% of errors.

8. Conclusions

This paper presented an implementation of K-means clustering to be used in a standard language tagger for non-standard text analysis. It outperforms the previous attempts to improve the tagger, which could be assigned to Word2Vec. K-means is an easy clustering algorithm to implement and scales well to large datasets. This speaks for the usefulness of this method. Using Word2Vec with clustering should be investigated further and combined with other clustering methods to find the most optimal one.

These results could be improved in several ways. As already mentioned, Ljubešić et al. (2017) presented several experiments where tagger was successfully improved and the final configuration is freely available. K-means together with Word2Vec could be used on that final configuration.

In this paper we created 2000 clusters for K-means, but this should be further investigated. Since there are 960 possible tag combinations for the full morphological description, it would be instightful to create exactly that many clusters. Furthermore are the tags themselves hierarchical, going from wider to more specific categories, so it would be interesting to try with some other hierarchical clustering algorihtms.

Additional improvement possibilities lie in Word2Vec parameters, where we could use a larger window size and increase of decrease the frequency of the words observed or change the default parameters. It might be useful to decrease the frequency, since non-standard language uses more diverse vocabulary and therewith less frequent words.

References

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based *n*-gram Models of Natural Language. *Computational linguistics*, 18(4):467–479, 1992.

Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369, 2013.

Tomaž Erjavec, Nikola Ljubešić, and Nataša Logar. The slWaC Corpus of the Slovene Web. *Informatica*, 39(1):35, 2015.

Darja Fišer. Gold-Standard Datasets for Annotation of Slovene Computer-Mediated Communication. *The 10th Workshop on Recent Advances in Slavonic Natural Languages Processing*, *RASLAN 2016*, page 29, 2016.

Dekang Lin and Xiaoyun Wu. Phrase Clustering for Discriminative Learning. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural*

Language Processing of the AFNLP: Volume 2 - Volume 2, pages 1030–1038. Association for Computational Linguistics, 2009.

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. Adapting a State-of-the-Art Tagger for South Slavic Languages to Non-Standard Text. *In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 60–68, 2017.

James MacQueen et al. Some Methods for Classification and Analysis of Multivariate Observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability,* pages 281-297, 1967.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, an Noah A Smith. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 380-390, 2013.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Van- derplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word Representations: A Simple and General Method for Semi-Supervised Learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.