Zbornik konference

Jezikovne tehnologije in digitalna humanistika

Proceedings of the Conference on

Language Technologies & Digital Humanities

29. september – 1. oktober 2016

Filozofska fakulteta, Univerza v Ljubljani

Ljubljana, Slovenija

September 29th – October 1st, 2016

Faculty of Arts, University of Ljubljana

Ljubljana, Slovenia

Uredila / *Edited by*

Tomaž Erjavec, Darja Fišer

ZBORNIK KONFERENCE
JEZIKOVNE TEHNOLOGIJE IN DIGITALNA HUMANISTIKA

PROCEEDINGS OF THE CONFERENCE ON LANGUAGE TECHNOLOGIES & DIGITAL HUMANITIES

Uredila / *Edited by*: Tomaž Erjavec, Darja Fišer
Tehnični uredniki / *Technical editors*: Jaka Čibej, Katja Zupan

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

Izdal / *Issued by*:
Institut »Jožef Stefan«, Ljubljana */ Jožef Stefan Institute, Ljubljana*
Oddelek za prevajalstvo */ Department of Translation Studies*

Za založbo / *For the publisher*:
Branka Kalenić Ramšak,
dekanja Filozofske fakultete / *Dean od the Faculty of Arts*

Ljubljana, 2016
Prva izdaja / *First edition*

Spletno mesto konference / *Conference web site:* http://www.sdjt.si/jtdh-2016/

Publikacija je brezplačno dostopna na: / *Publication is available free of charge at:*
http://nl.ijs.si/isjt16/proceedings-sl.html / *http://nl.ijs.si/isjt16/proceedings-en.html*

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID=286548224
ISBN 978-961-237-862-2 (pdf)

Predgovor k zborniku konference "Jezikovne tehnologije in digitalna humanistika"

Prva konferenca Jezikovne tehnologije je potekala davnega leta 1998 in se nato redno nadaljevala vsaki dve leti. Ob jubilejni deseti konferenci JT v letu 2016 smo se odločili za programsko širitev na področje digitalne humanistike, ki je kot presek digitalnih tehnologij in humanistike aktualno raziskovalno področje, kjer se digitalne tehnologije na eni strani uporabljajo pri raziskavah v humanistiki za študij jezika, družbe in kulture, na drugi strani pa humanistika spodbuja razvoj novih tehnoloških rešitev. Digitalna humanistika je mednarodno že uveljavljeno področje, kot npr. kažejo revija Journal of Digital Humanities, zveza Alliance of Digital Humanities Organizations s svojimi letnimi konferencami Digital Humanities in mreža NeDiMAH (Network for Digital Methods in the Arts and Humanities). V samem izhodišču je digitalna humanistika izrazito interdisciplinarno in kolaborativno raziskovalno delo, ki bistveno spreminja ustaljene humanistične pristope ter spodbuja razvoj novih analitičnih tehnik in metod, v slovenskem prostoru pa še nima skupnega mesta za predstavitev dosežkov svojega dela in diskusijo med različnimi deležniki na področju. Zato so Slovensko društvo za jezikovne tehnologije (SDJT), Center za jezikovne vire in tehnologije Univerze v Ljubljani (CJVT) ter raziskovalni infrastrukturi CLARIN.SI in DARIAH-SI 29. 9.–1. 10. 2016 na Filozofski fakulteti Univerze v Ljubljani organizirale konferenco Jezikovne tehnologije in digitalna humanistika.

Za razliko od konferenc JT, in glede na specifike DH, tokrat prispevkov nismo omejili na polne članke, temveč smo sprejemali tudi razširjene povzetke in uvedli študensko sekcijo, v spremljevalni konferenčni program pa uvrstili še panel Terminologija v poklicnem vsakdanu: stanje in potrebe in sekcijo s predstavitvijo sponzorjev.

Vse redne in študentske prispevke sta pregledala dva recenzenta, k sodelovanju pa nam je uspelo pritegniti tudi pet vabljenih predavateljev. Zbornik vsebuje 58 sprejetih prispevkov, od tega 5 prispevkov vabljenih predavateljev, 30 rednih polnih prispevkov in 17 povzetkov ter 6 študentskih prispevkov, pri čemer je 41 prispevkov v slovenskem, ostali pa v angleškem, hrvaškem, srbskem in bosanskem jeziku.

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

Ljubljana, september 2016

Tomaž Erjavec, Darja Fišer

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

The first Language Technologies Conference in Slovenia took place in the long-past year 1998 and from then on regularly every second year. On the occasion of the anniversary tenth conference in 2016 we decided to widen its scope to the field of Digital Humanities. DH is, as the intersection of digital technologies and the humanities, a highly relevant research field, where, on the one hand, digital technologies are used in the study of language, society and culture, and, on the other hand, humanities research paves the way for the development of new digital technologies. Digital Humanities is internationally already an established field, as is evidenced by the Journal of Digital Humanities, the Alliance of Digital Humanities Organizations with its annual conferences Digital Humanities and the Network for Digital Methods in the Arts and Humanities (NeDiMAH). DH is, by definition, highly interdisciplinary and collaborative, radically changing the accepted practices in the humanities research and encouraging the development of new analytical techniques and methods but has so far lacked a national or regional event to present its results and encourage discussion.
To this end, the Slovenian Language Technologies Society (SDJT), the Centre for Language Resources and Technologies at the University of Ljubljana (CJVT), and the Slovenian research infrastructures CLARIN.SI and DARIAH-SI organised the conference Language Technologies and Digital Humanities which took place September 29th to October 1st 2016 at the Arts Faculty of the University of Ljubljana.

In contrast to the previous editions of the Language Technologies conferences, and taking into account the specifics of the Digital Humanities field, the 2016 submission were not limited to full papers, but extended abstracts were accepted as well. Additionally, a student session, a round table on terminology and a session with the presentation of the sponsors was organised.
All regular and student contributions were reviewed by two members of the programme committee, with the conference also presenting the talks of five invited speakers. The proceedings contain 58 accepted papers, 5 of which are from invited speakers, 30 regular full papers, 17 abstracts, and 6 student contributions, with 41 contributions in Slovenian, with the others in English, Croatian, Serbian, and Bosnian languages.

The editors would like to thank all the people who contributed to the success of the conference: the invited speakers, authors of the papers and abstracts, the programme committee for exemplary reviewing of the contributions, the organising committee for the smooth functioning of the event, sessions chairs who made sure the talks proceeded smoothly, the technical editors for the preparation of the proceedings and the sponsors for supporting the event.

Ljubljana, September 2016

Tomaž Erjavec, Darja Fišer

Organizacijski odbor / Organising committee

Darja Fišer, predsednica / Chair (SDJT, FF, IJS)

Jerneja Fridl (DARIAH-SI, ZRC SAZU)

Vojko Gorjanc (CJVT, FF)

Simon Krek (CJVT, CLARIN.SI, IJS)

Mojca Šorn (DARIAH-SI, INZ)

Kaja Dobrovoljc (študentska sekcija / student session, SDJT)

Programski odbor / Programme committee

Predsedstvo programskega odbora / Steering committee

Tomaž Erjavec**,** predsednik / Chair

Institut "Jožef Stefan" / Jožef Stefan Institute (CLARIN.SI)

Darja Fišer

Filozofska fakulteta, Univerza v Ljubljani / Faculty of Arts, University of Ljubljana (SDJT)
Institut "Jožef Stefan" / Jožef Stefan Institute

Vojko Gorjanc

Filozofska fakulteta, Univerza v Ljubljani / Faculty of Arts, University of Ljubljana (CJVT)

Andrej Pančur

Inštitut za novejšo zgodovino / Institute of Contemporary History (DARIAH-SI)

Vodje tematskih področij / Area chairs

Govorne tehnologije / Speech technologies:

Jerneja Žganec Gros, Alpineon d.o.o.

Jezikovne tehnologije / Language technologies:

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

Prevodoslovje / Translation studies:

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

Leksikologija in leksikografija / Lexicology and lexicography:

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

Jezikovna standardizacija / Language standardisation:

Helena Dobrovoljc, Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU / Fran Ramovš Institute of Slovenian Language, Scientific Research Centre of the Slovenian Academy of Sciences and Arts (ZRC SAZU)

Jezikovna didaktika in opismenjevanje / Didactics and literacy pedagogy:

Špela Arhar Holdt, Trojina, zavod za uporabno slovenistiko / Trojina, Institute for Applied Slovene Studies and Filozofska fakulteta, Univerza v Ljubljani / Faculty of Arts, University of Ljubljana

Izobraževanje / Education:

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

Literarne vede / Literary studies:

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

Zgodovinopisje / History studies:

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

Tehnološke rešitve za raziskave in predstavitve kulturne dediščine / Technology solutions for research and presentation of cultural heritage:

Jerneja Fridl, ZRC SAZU / Scientific Research Centre of the Slovenian Academy of Sciences and Arts

Digitalne umetnosti in prakse / Digital arts and practices:

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

Diseminacija in arhiviranje digitalnih vsebin / Dissemination and archiving of digital resources:

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

Člani programskega odbora in recenzenti / Programme committee members and reviewers

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

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

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

Vlado Delić, Fakulteta za tehnične vede, Univerza v Novem Sadu / Faculty of Technical Sciences, University of Novi Sad

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

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

Tatjana Hajtnik, Arhiv Republike Slovenije / Archive of the Republic of Slovenia

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

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

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

Iztok Kosem, Trojina, zavod za uporabno slovenistiko / Trojina, Institute for Applied Slovene Studies in/and Filozofska fakulteta, Univerza v Ljubljani / Faculty of Arts, University of Ljubljana

Žiga Kokalj, Inštitut za antropološke in prostorske študije ZRC SAZU / Institute of Anthropological and Spatial Studies, ZRC SAZU

Mojca Kotar, Univerza v Ljubljani / University of Ljubljana

Cvetana Krstev, Filozofska fakulteta, Univerza v Beogradu / Faculty of Arts, University of Beograd

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

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

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

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

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

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

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

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

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

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 and 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 in/and Information Science, University of Ljubljana

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

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

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

Benjamin Štular, Inštitut za arheologijo ZRC SAZU / Institute of Archaeology, ZRC SAZU

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

Organizatorji / Organizers

Sponzorji / Sponsors

Urnik / Timetable

Četrtek, 29. 9. 2016 / Thursday 29-9-2016

8.30-9.00	**Registracija / Registration**
9.00-9.30	**Otvoritev / Opening** (predavalnica / room: 34)
9.30-10.30	**Vabljeno predavanje / Invited lecture** (predavalnica / room: 34) Daniel Zeman, *Universal Dependencies for Slavic Languages*
10.30-11.00	**Odmor za kavo / Coffee break**
11.00-12.30	**Paralelna sekcija 1A / Parallel session 1A** (predavalnica / room: 34)
11.00 - 11.20	Mihael Arčan, Maja Popović, Paul Buitelaar: *Asistent – A Machine Translation System for Slovene, Serbian and Croatian*
11.20 - 11.40	Shaun Azzopardi, Albert Gatt, Gordon Pace: *Integrating Natural Language and Formal Analysis for Legal Documents*
11.40 - 12.00	Nikola Ljubešić, Tomaž Erjavec, Darja Fišer, Tanja Samardžić, Maja Miličević, Filip Klubička, Filip Petkovski: *Easily Accessible Language Technologies for Slovene, Croatian and Serbian*
12.00 - 12.15	Nikola Ljubešić, Tanja Samardžić, Maja Miličević: *Analysing Spatial Distribution of Linguistic Variables in Geoencoded Tweets from Croatia, Bosnia, Montenegro and Serbia*
12.15 - 12.30	Marijana Janjić, Sara Librenjak, Kristina Kocijan: *Asian Language Teaching and Learning - the Influence of Technology on Students' Skills in SL Classroom*
11.00-12.30	**Paralelna sekcija 1B / Parallel session 1B** (predavalnica / room: 18)
11.00 - 11.20	Slobodan Mandić: *The First World War on the Web - The Case of Serbia*
11.20 - 11.40	Aleš Vaupotič, Marco Buziol, Narvika Bovcon: *Digital Video in Digital Humanities Methodology: A Case Study*
11.40 - 12.00	Narvika Bovcon, Jure Demšar, Aleš Vaupotič: *Organiziranje projekta vizualizacije podatkov*
12.00 - 12.15	Gregor Strle, Matija Marolt: *Language Technologies in Humanities: Computational Semantic Analysis in Folkloristics*
12.15 - 12.30	Mario Hibert: *What Is Critical in Digital Humanities?*
12.30-13.30	**Odmor za kosilo / Lunch break**
13.30-14.30	**Vabljeno predavanje / Invited lecture** (predavalnica / room: 34) Niels-Oliver Walkowski, *The Landscape of Digital Annotations and Its Meaning*
14.30-15.00	**Odmor za kavo / Coffee break**

15.00-16.30	**Paralelna sekcija 2A / Parallel session 2A** (predavalnica / room: 34)
15.00 - 15.20	Darja Fišer, Jasmina Smailović, Tomaž Erjavec, Miha Grčar, Igor Mozetič: *Sentiment Annotation of the Janes Corpus of Slovene User-Generated Content*
15.20 - 15.40	Natalia Korchagina: *Building a Gold Standard for Temporal Entity Extraction from Medieval German Texts*
15.40 - 16.00	Ivana Lalli Paćelat: *Priprema usporedivih korpusa za usporedbu*
16.00 - 16.15	Stevan Ostrogonac, Branislav Popović, Milan Sečujski: *The Use of Semantic Word Classes in Document Classification*
16.15 - 16.30	Tadeja Rozman, Špela Arhar Holdt, Senja Pollak, Iztok Kosem: *Luščenje in jezikoslovna analiza kolokacij iz korpusa Šolar*
15.00-16.30	**Paralelna sekcija 2B / Parallel session 2B** (predavalnica / room: 18)
15.00 - 15.20	Anja Ragolič: *Eagle - Medomrežje Europeana antične grške in latinske epigrafike. Digitalni dostop do antičnih napisnih spomenikov*
15.20 - 15.40	Maja Vičič Krabonja: *Digitalna humanistika v šoli*
15.40 - 16.00	Andrej Pančur: *Popisi prebivalstva Slovenije 1830–1931: Orodje za transkribiranje historičnih demografskih podatkov*
16.00 - 16.15	Aleš Lazar, Sonja Ifko: *Trirazsežno dokumentiranje v službi varovanja nepremične kulturne dediščine*
16.15 - 16.30	Andrej Pančur, Bogomir Rožman: *Dolgotrajno ohranjanje raziskovalnih podatkov v manjših raziskovalnih infrastrukturah: Uporaba odprtokodne aplikacije Archivematica*
16.30-17.30	**Sponzorska sekcija / Presentations of sponsors** (predavalnica / room: 18)

Petek, 30. 9. 2016 / Friday 30-9-2016

9.00-9.30	**Registracija / Registration**
9.30-10.30	**Vabljeno predavanje / Invited lecture** (predavalnica / room: 34) Laurent Romary, *The Text Encoding Initiative: 30 Years of Accumulated Wisdom and Its Potential for a Bright Future*
10.30-11.00	**Odmor za kavo / Coffee break**
11.00-12.30	**Paralelna sekcija 3A / Parallel session 3A** (predavalnica / room: 34)
11.00 - 11.20	Jaka Čibej, Špela Arhar Holdt, Tomaž Erjavec, Darja Fišer: *Razvoj učne množice za izboljšano označevanje spletnih besedil*
11.20 - 11.40	Polona Gantar, Iza Škrjanec, Darja Fišer, Tomaž Erjavec: *Slovar tviterščine*

11.40 - 12.00	Špela Arhar Holdt, Kaja Dobrovoljc, Iztok Kosem: *Predstavitveni portal spletnih jezikovnih virov za slovenščino*
12.00 - 12.15	Kaja Dobrovoljc, Tomaž Erjavec, Simon Krek: *Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino*
12.15 - 12.30	Simon Krek, Polona Gantar, Špela Arhar Holdt, Vojko Gorjanc: *Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres*
11.00-12.30	**Paralelna sekcija 3B / Parallel session 3B** (predavalnica / room: 18)
11.00 - 11.20	Simon Krek, Polona Gantar, Kaja Dobrovoljc, Iza Škrjanec: *Označevanje udeleženskih vlog v učnem korpusu za slovenščino*
11.20 - 11.40	Helena Dobrovoljc: *Povezljivost pravopisnih pravil in slovarja: sanje pravopiscev 20. stoletja*
11.40 - 12.00	Andrej Pančur: *Označevanje zbirke zapiskov sej slovenskega parlamenta s smernicami TEI*
12.00 - 12.15	Bojan Kastelic, Mateja Belak, Andrej Pleterski, Benjamin Štular, Miran Erič: *Zbiva in EWD, spletni orodji za arheološke raziskave*
12.15 - 12.30	-
12.30-13.30	**Odmor za kosilo / Luch break**
13.30-14.30	**Vabljeno predavanje / Invited lecture** (predavalnica / room: 34) Peter Juel Henrichsen, *Speech is Golden - on ASR at the Service of the Danish Public Sector*
14.30-15.00	**Odmor za kavo / Coffee break**
15.00-16.30	**Paralelna sekcija 4A / Parallel session 4A** (predavalnica / room: 34)
15.00 - 15.20	Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Nataša Logar, Milan Ojsteršek: *Slovenska akademska besedila: prototipni korpus in načrt analiz*
15.20 - 15.40	Klemen Kadunc, Marko Robnik-Šikonja: *Analiza mnenj s pomočjo strojnega učenja in slovenskega leksikona sentimenta*
15.40 - 16.00	Katja Zupan, Tomaž Erjavec: *Generiranje kritičnih prepisov s strojnim prevajanjem na ravni znakov*
16.00 - 16.15	Dan Podjed, Saša Babič, Tatiana Bajuk Senčar, Alenka Bezjak Mlakar, Gregor Burger, Jurij Fikfak, Jože Guna, Marko Maver, Matevž Pogačnik, Emilija Stojmenova, Uroš Žolnir: *Razvoj aplikacije za spodbujanje trajnostne mobilnosti*
16.15 - 16.30	Tatjana Veljanovski, Žiga Kokalj: *Slikovna retrospektiva porušenega Breginja in analiza pokrajinskih sprememb*
15.00-16.30	**Paralelna sekcija 4B / Parallel session 4B** (predavalnica / room: 18)
15.00 - 15.20	Andrej Žgank, Darinka Verdonik,Mirjam Sepesy Maučec: *Razpoznavanje tekočega govora v slovenščini z bazo predavanj SI TEDx-UM*

15.20 - 15.40	Jerneja Žganec Gros, Boštjan Vesnicer, Simon Rozman, Peter Holozan, Tomaž Šef: *Sintetizator govora za slovenščino eBralec*
15.40 - 16.00	Simon Dobrišek, David Čefarin, Vitomir Štruc, France Mihelič: *Preizkus Googlovega govornega programskega vmesnika pri samodejnem razpoznavanju govorjene slovenščine*
16.00 - 16.15	Miha Seručnik: *Historična topografija Slovenije*
16.15 - 16.30	Drago Kunej: *Digitalizacija in uporaba digitalnega etnomuzikološkega zvočnega gradiva – izkušnje Glasbenonarodopisnega inštituta ZRC SAZU*
16.30-18.00	**Panel o terminologiji / Round table on terminology** (predavalnica / room: 34)
20.00-22.00	**Večerja / Dinner**

Sobota, 1. 10. 2016 / Saturday 1-10-2016

9.30-9.45	**Registracija / Registration**
9.45-10.30	**Vabljeno predavanje / Invited lecture** (predavalnica / room: 34) Oliver Čulo, *Contrasting Post-editing and Human Translation along the Dimension of Term Translation*
10.30-11.00	**Odmor za kavo / Coffee break**
11.00-12.30	**Paralelna sekcija 5A / Parallel session 5A** (predavalnica / room: 34)
11.00 - 11.20	Benjamin Štular, Franco Niccolucci, Julian Richards: *ARIADNE: povezani odprti podatki (LOD) v praksi*
11.20 - 11.40	Iztok Kosem, Tadeja Rozman, Špela Arhar Holdt, Polonca Kocjančič, Cyprian Laskowski: *Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov*
11.40 - 12.00	Simon Krek, Polona Gantar, Iztok Kosem, Vojko Gorjanc, Cyprian Laskowski: *Baza kolokacijskega slovarja slovenskega jezika*
12.00 - 12.20	Špela Arhar Holdt, Iztok Kosem: *Ohranjanje jezikovne zahtevnosti besedil pri prevajanju testov PISA*
11.00-12.30	**Paralelna sekcija 5B / Parallel session 5B** (predavalnica / room: 18)
11.00 - 11.20	Teja Goli, Eneja Osrajnik, Darja Fišer: *Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter*
11.20 - 11.40	Damjan Popič, Darja Fišer, Katja Zupan, Polona Logar: *Raba vejice v uporabniških spletnih vsebinah*
11.40 - 12.00	Matija Ogrin, Andrejka Žejn: *Strojno podprta kolacija slovenskih rokopisnih besedil. Variantna mesta v luči računalniških algoritmov in vizualizacij*
12.00 - 12.15	Matija Ogrin, Tomaž Erjavec, Jan Jona Javoršek: *Od znanstvenokritične izdaje do repozitorija rokopisov. Dosežki, možnosti in načrti*

12.30-13.30	**Odmor za kosilo / Lunch break**
13.30-14.40	**Študentska sekcija 1 / Student session 1** (predavalnica / room: 34)
13.30 - 13.50	Anja Tavčar, Ines Čeligoj Pregelj, Miha Pompe: *Korpus študentov prevajanja MetaTrans*
13.50 - 14.10	Jure Škerl: *RBP pri strojnih prevajalnikih Amebis Presis, Google Translate in MT@EC*
14.10 - 14.25	Lucija Dačić: *Literature Reloaded: Using Databases to Explore Literary Trends*
14.25 - 14.40	Jernej Rihter: *Digitalna arheologija? Primer uporabe digitalnih orodij za analizo arheološkega najdišča*
14.40-15.00	**Odmor za kavo / Coffee break**
15.00-15.40	**Študentska sekcija 2 / Student session 2** (predavalnica / room: 34)
15.00 - 15.20	Miha Helbl, Žiga Domevšček: *Gradnja in analiza petjezičnega korpusa podnaslovov govorov TED*
15.20 - 15.40	Katerina Pertot, Maja Petrovčič, Nika Strojan: *#Analiza novih komunikacijskih elementov na družbenem omrežju @Twitter*
15.40-16.00	**Zaključek konference s podelitvijo študentske nagrade / Conference closing with best student paper award** (predavalnica / room: 34)
16.00-18.00	**SDJT občni zbor** (predavalnica / room: 34)

Kazalo / Table of Contents

Contrasting post-editing and human translation along the dimension of term and cognate variation

Oliver Čulo

Faculty of Translation Studies, Linguistics and Cultural Studies, University of Mainz
An der Hochschule 2, 76726 Germersheim, Germany
culo@uni-mainz.de

1. Introduction

Post-editing is a rather recent mode of translation production. In the most basic definition, post-editing is the correction of machine translation output, but some definitions include aspects such as quality criteria or the fact that post-editors should be trained translators (see e.g. O'Brien 2011).

Post-editing has been studied from various angles, mainly with respect to the questions of efficiency or quality. In some cases, the process of post-editing has been contrasted to the process of human translation (see O'Brien et al. 2014 for a collection of studies). While differences have been shown to exist between the two processes, e.g. on the level of macro processes (see e.g. Carl et al. 2011), little is known about how the post-edited products differ from human translation (with first insight provided e.g. in Lapshinova-Koltunski 2013; Lapshinova-Koltunski 2015) and which effects this might have on communication.

In this talk, I will first contrast post-editing and human translation along the dimension of term translation within the domain of Languages for Specific Purposes. In the study presented, terminological variation in translations from English into German was measured for both modes of translation. The findings reveal levels of variation on the terminological level in the post-edited texts close, but not identical, to those of the machine translation outcomes. They thus indicate a shining through of the machine translations in the post-editing products, motivating further research into the properties of post-edited texts within corpus-based translation studies. Also, I will present a brief pilot study on how cognates are used differently in machine and human translations: While the machine translation system used was heavily using cognates, in human translation (motivated) variation in the (non-)use of cognates could be observed.

Of course, both findings are very much dependent on the characteristics of the machine translation used. However, the point here is not to make generalizable statements on lexical properties of machine translated texts, but to identify dimensions along which human translations, post-edits and machine translations may differ and may be contrasted. On the basis of these findings, I will discuss in what way machine translation can have an impact on the product of translation and whether it might become a driving force for language change.

2. Evaluation

2.1. Data collection

The evaluation presented here is based on a general-purpose collection of translation and post-editing data including key logs and eye-tracking protocols recorded with Translog-II[1]. In the experiment sessions, advanced students of translation were asked to lightly post-edit, fully post-edited and translate from scratch text snippets from a dish washer manual (12 students) and a medical leaflet (9 students). The order of tasks was permutated so that each text snippet was translated 4 resp. 3 times in each mode. All texts were about 150 words long. The texts were automatically pre-translated by Google Translate for the PE tasks. Students were given instruction for the post-editing; part of the instructions for the full post-editing as opposed to light post-editing was that they were specifically asked to ensure terminological consistency.

2.2. Evaluation

For the evaluation of term variation, we marked nominal terms in the English original texts which appeared at least 3 times in the text snippet. Nominal lexemes which were too general, such as *item* or *disease*, were ruled out as terms on an intuitive basis. For each term, we checked which variants (if any) were used in translation. This was mapped onto event types, with the most frequent translation as the preferred translation type, the second most frequent as the synonym 1 translation type etc.

Using this mapping, we were able to calculate translation probabilities and on top of this perplexity values for each translation type. A statistical analysis of perplexity values for terms in the machine translated and fully post-edited texts shows patterns of (non-)variation for machine translation and full post-editing very similar of

[1] https://sites.google.com/site/centretranslationinnovation/translog-ii

each other, but very different for human translation (Čulo and Nitzke 2016). In other words, on the terminological level, there is a *shining through* (Teich 2003) of lexical patterns from machine translation to post-edited texts in our data set.

Session	term translation	term frequency	event type
P21_FPE	Geschirrspülmaschine	4	pref.t.trans
	Spülmaschine	2	syn.1.trans
P10_FPE	Spülmaschine	2	pref.t.trans
	Geschirrspülmaschine	2	syn.1.trans
	Geschirrspüler	1	syn.2.trans

Table 1: Mapping of term variant to translation event type

In a second experiment, we evaluated the use of cognates in translations corpus data (taken from the English-German Translation Corpus of TU Chemnitz[2]) with machine translated data (Google Translate) for 13 English-German cognates that occurred more than five times in the corpus data. Cognates "are those translation words that have similar orthographic-phonological forms in the two languages of a bilingual […]; non-cognates are those translations that only share their meaning in the two languages […]" (Costa, Caramazza, and Sebastian-Galles 2000, 1285). In the language pair English-German, *system* and *System* are for example cognates, while *government* and *Regierung* are non-cognates. *System* is not always the best translation for *system*, however; depending on the context, translations like *Anlage* (roughly 'installation') or *Verfahren* 'procedure' may be better options. So called false friends are different from cognates in that false friends are two words that share the same (or a very similar) form, but not the same meaning across two languages, like the English word *actual* (real, existing) vs. the German *aktuell* (current, latest).

A comparison of the machine translated texts reveals that the machine translation system used has a strong affinity towards the use of cognates in translation, while human translations show variation (see Čulo and Nitzke 2016) such as in the following example:

Source: "[…] political stability rested on the acceptance in all classes of the legitimacy […]"
Target: "[…] beruhte ihre Stabilität darauf, daß alle Klassen die Legitimität […] akzeptierten"
Lit.: […] *rested their stability on that all classes the legitimacy […] accepted*

Here, the English noun *acceptance* was translated by the German verb *akzeptieren* 'accept'. We assume that a word class shift is a cognitively different operation from translation by a cognate in the same word class and thus do not count this as cognate translation. The reason for such variation may be motivated by the fact that noun and verb of the same root may not always have exactly the same meanings: we would argue that the German word *Akzeptanz* is not used in the same way as *acceptance*, but rather as some sort of rate of acceptance, i.e. a rising or falling acceptance towards some measure or phenomenon.

3. Conclusions and further work

In the two evaluations made on existing machine translations, post-edits and translations-from-scratch we identified the lexical level, more specifically terms and cognates, as dimensions along which these three types of translations products can be contrasted. The exact nature of results will, of course, to a large extent depend on the characteristics of the machine translation systems used.

Further questions arise from the observations made: If the texts from the cognate evaluation were post-edited, would patterns of cognate use remain in the post-edited texts? If so, which cognitive processes would need to be trained or triggered in order to introduce desired (because motivated) patterns of variation? And ultimately, could patterns of cognate use "spill over" into original language production, e.g. replacing non-cognates over time? If so, and if one could show that this was due to exposure of (partially) machine-generated translations, then machine translation would be established as a driving force of language change.

4. References

Carl, Michael, Barbara Dragsted, Jakob Elming, Daniel Hardt, and Arnt Lykke Jakobsen. 2011. 'The Process of Post-Editing: A Pilot Study'. In *Proceedings of the 8th International NLPSC Workshop. Special Theme: Human-Machine Interaction in Translation*, edited by Bernadette Sharp, Michael Zock,

[2] http://ell.phil.tu-chemnitz.de/search/

Michael Carl, and Arnt Lykke Jakobsen, 131–42. Copenhagen Studies in Language 41. Frederiksberg: Samfundslitteratur.

Čulo, Oliver, and Jean Nitzke. 2016. 'Patterns of Terminological Variation in Post-Editing and of Cognate Use in Machine Translation in Contrast to Human Translation'. *Baltic Journal of Modern Conputing* 4 (2): 106–14.

Lapshinova-Koltunski, Ekaterina. 2013. 'VARTRA: A Comparable Corpus for the Analysis of Translation Variation'. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora*, 77–86. Sofia, Bulgaria.

———. 2015. 'Variation in Translation: Evidence from Corpora'. In *New Directions in Corpus-Based Translation Studies*, edited by Claudio Fantinuoli and Federico Zanettin, 93–113. Translation and Multilingual Natural Language Processing 1. Berlin: Language Science Press.

O'Brien, Sharon. 2011. 'Towards Predicting Post-Editing Productivity'. *Machine Translation* 25 (3): 197–215.

O'Brien, Sharon, Laura Winther Balling, Michael Carl, Michel Simard, and Lucia Specia, eds. 2014. *Post-Editing of Machine Translation Processes and Applications.* Newcastle upon Tyne: Cambridge Scholars Publishing. http://public.eblib.com/choice/publicfullrecord.aspx?p=1656492.

Teich, Elke. 2003. *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Vol. 5. Text, Translation, Computational Processing. Berlin/New York: Mouton de Gruyter.

Speech Is Golden – On ASR at the Service of the Danish Public Sector

Peter Juel Henrichsen

Department of International Business Communication,
Copenhagen Business School
Solbjerg Plads 3, DK-2000 Frederiksberg, Denmark
pjh.ibc@cbs.dk

In this talk I'll present the introduction of automatic speech recognition in the Danish municipalities. Ready to begin to adopt ASR, most municipalities remain nervous following a long series of bad business cases in the recent past. Complaints are voiced over costly licences and low service levels, typical effects of a de facto monopoly on the supply side. As a counteract, DanCAST (Danish Center for Applied Speech Technology) established a consortium three years ago involving more than half of the Danish municipalities, a number of SMEs, NGOs, and other organizations, together constituting a body strong enough to formulate a number of requirements to be met by suppliers and tendering material allowing fair competition. The consortium has also formulated guidelines for the collection and unrestricted sharing of a corpus of transcribed speech samples and text materials necessary (and sufficient) for developing professional ASR applications. The corpus is steadily growing, paving the way for fair competition among large and small ASR suppliers to the public sectors. In the talk I'll break down our motivations, events and actions, and comment on the benefits of national-scale cooperation in ASR development for small language areas.

The Text Encoding Initiative: 30 Years of Accumulated Wisdom and Its Potential for a Bright Future

Laurent Romary*‡

*INRIA
‡Digital Research Infrastructure for the Arts and Humanities DARIAH-EU
laurent.romary@inria.fr

Since its launch in 1987, the Text Encoding Initiative has become the reference platform for the representation of text based digital assets in the humanities. As such, it has shaped the current Digital Humanities landscape by bringing a wide multi-disciplinary community into sharing a common set of concepts, and above all, by leveraging the global technical competence of scholars in XML technologies. In this talk we would like to show how the technical, but also editorial, experience gained over the years may be the base for providing even more service to the humanities. We will present the main architecture and components of the TEI and describe new mechanisms through the example of the future stand-off component. We will also show how the TEI can also liaise closely with international standardisation environment such as ISO in the domain of language resources.

The Landscape of Digital Annotation and Its Meaning

Niels-Oliver Walkowski*

* TELOTA, Berlin-Brandenburg Academy of Sciences and Humanities
Jägerstraße 22/23, 10117 Berlin
walkowski@bbaw.de

Abstract

In this paper I argue that due to the development of computational environments the usage scenarios and the interpretation of annotations have become an increasingly complex issue. Furthermore, I evaluate the different contextual dimensions which constitute the meaning of annotation data and present a formal scheme for their systematization.

1. The Rise of Annotations

The topic of annotations has raised significant interest in the field of Digital Humanities and beyond during the last years. Many ESFRI[1] projects from the Humanities or Social Sciences have worked on annotations in corresponding working packages. With the pund.it[2] EUROPEANA[3] has also built its own annotation tool. In the W3C two working groups developed a web compliant standard model for annotations (Sanderson, Ciccarese, & Van de Sompel, 2013) and specifications for a web annotation architecture[4]. Finally, projects like annotator.js[5] and hypothes.is[6] gained tremendous community interest as well as significant funding. All these activities have significantly multiplied the usage and usage scenarios of annotations.

How do these activities affect the understanding of annotations or the ways of annotating? There are two possibilities to read the overambitious title of this study. The first version asks for the meaning of the landscape itself that is to say what is the meaning of the way this landscape is shaped. For instance, what does the structure of this landscape tell us about the state of annotations in the digital age. The second version refers to the meaning that is represented in annotations itself and which forms part of this landscape. To put it more concisely, the question here is: what do all these annotations mean and how do we find out?

The work which is presented in the current paper was financed by DARIAH-DE and benefits from a survey that was carried out by the 'DARIAH-EU Working Group Digital Annotations' (DARIAH-EU, 2016). Hence, the question is also how does this development challenge the work of infrastructure projects and why should infrastructure projects take care about these questions.

Infrastructure projects are building storage, services and tools among other things. As I mentioned before, this holds especially true for annotation data at the moment.. However, as both Atkins (Atkins, 2003) and Rockwell (Rockwell, 2010) points out infrastructure needs to be more to become successful. It needs to also take care about the research ecology which it seeks to serve. In the case of annotations this means to be informed about new digital annotation practices and to assure that annotations are used in a sound way. The dynamic which has been described before as well as the peculiar nature of annotation data makes this task especially challenging for annotations. They are mostly a granular, highly context dependent piece of information.

Another reason is given by the fact that DARIAH is a project in the (Digital) Humanities that means embedded into the Humanities research tradition. In the Humanities research about the modes of knowledge production itself is a crucial part of the research portfolio. Additionally, annotating has a long tradition in humanist research practice. Thus, research on digital annotations is also a great chance to push forward the integration of digital methods in the Humanities as Meister (2015) points out.

Having all this said, the title of this paper could be transformed into two questions:

- What needs to be known from annotation contexts so that annotation data can be reasonably used elsewhere? In technical terms, what are the metadata needs?
- What are annotations today or in the language of infrastructure projects, which best practices in annotating exist?

In the rest of this paper I will try to come closer to an answer for these questions.

2. The Meaning in Annotations

For the purpose of providing rich information to support the evaluation of these questions the DARIAH-EU Working Group Digital Annotations developed a sophisticated questionnaire with over 40 questions. The main goal of the survey is not to derive statistical claims but to generate a dense description that represents the complexity of the topic. Thus, the number of investigated use-cases is relatively low and include 17 filled out questionnaires.

Before getting into detail, a general result of the survey addresses the value of the whole effort as such. More precisely, it was asked if project members think that annotations from their work could be fully understood on its own and without further knowledge about the project. A positive answer to this question seems appropriate in a data sharing context

1 The European Strategy Forum on Research Infrastructures is a policy making body of the European Commission for the development of (digital) research infrastructures
2 http://thepund.it/
3 http://www.europeana.eu/portal/
4 https://www.w3.org/annotation/
5 http://annotatorjs.org/
6 https://hypothes.is/

Curiously enough, the response pattern resembles a common pattern of similar questionnaires regarding research data sharing in general. Only one person answered negatively while another one abstained. However, the number of purely positive answers were also only four. The majority of people wrote 'Yes, but …'. I do not only interpret this result as a clear expression in favour of further research on metadata needs for sharing annotation data but also for uncertainty about the status of annotations in terms of data sharing.

2.1. Meaningful Dimensions of Annotations Identified in the Survey

Next, I will give a short overview about important different dimensions for the meaning of annotations in the light of the survey. Some of these dimensions were explicitly mentioned by the participating projects. Others were extracted from answers where they implicitly address issues of meaning construction.

The first aspects which influences the appropriate interpretation of the meaning in annotations concerns the technological production of the annotated object. In a use-case from the field of visual anthropology[7] the participant remarks that knowledge about the process of ethnographic film making is very supportive to understand the annotations about these films. Likewise, the Monasterium[8] use-case reveals that certain annotations on documents can only be understood with knowledge about the creation of digital copies of these documents.

In the DARIAH-DE Fellowship use-case a comparable issue is mentioned but evaluated slightly different. This use-case addresses the issue of annotations in which their content might not be enough for its interpretation. However, the investigation of the annotated object region provides sufficient context information. The link between both might seem obvious. However, in digital annotating the target might not be at the same place as the annotation body. This can cause problems of different types. Dereferenceability and even more renderability of annotated objects should therefor be a crucial aspect.

Another dimension does also concern the annotated object. However, this time it is about the question what is technically referenced. The Video Annotation in Transcultural Studies[9] (VATS) as well as the Semantic Topological Notes[10] (SemToNotes) use-case emphasize that many annotation services do not offer the possibility to exactly reference a shape in an image. Instead, the annotation reference creates a box around the shape of interest. This is not precise and can lead to information retrieval and interpretation issues. We can call this the fragment dimension.

The fragment dimension is part of a bigger issue. This issue is about the concrete object layer which is addressed by the annotation. Some examples will clarify what is meant. The Relations in Space use-case in which inscriptions in Jewish gravestones are annotated distinguishes between annotations about the carrier of the inscription (the gravestone) and the inscription itself. Accordingly, annotations which reference the same area in the digital copy address completely different facets.

In the Monasterium illuminated areas in the digital copies are annotated next to layout information. Thus, the first group of annotations do not address the material nor some basic semantic concepts (title, paragraph, among others), they describe aspects of the digitization.

In e-Metaphor annotating a part of text as a metaphor does not just mean the metaphor itself but the 'focus and frame of metaphorical construction'. Within the illustrative terminology from literature studies it addresses an intratextual dimension of the metaphor and not just the metaphor.

Alluding to a common term in the research field of Systemic Functional Grammar the context dimension which discriminates the annotated objects in levels between materiality and intertextuality can be called the strata dimension.

Both the use-case Visual Anthropology as well as Ethnomusicology[11] highlight that knowledge about the way Ethnologists or Musicologists work and process content are supportive facets to understand well corresponding annotations. This dimension is the methodology or practice dimension.

Another meaningful dimension for the sound use of annotations might seem too obvious and trivial to consider. However, it is a very important dimension. The form and properties of the annotation itself needs to be clear. To put it in technical terms, the model needs to be transparent. I mentioned the standardized Open Annotation Data Model before. Nevertheless, the fact that such a model exists does not mean that it is always used or can be used everywhere - technically as well as semantically.

For instance, the e-Codicology[12] use-case has defined its annotation model in a proprietary SKOS model. The VATS use-case creates attention for fact that the annotation body can be encoded in a way that needs information about what is required to render it. This can be something like MIME-Type information for example. On a semantic level the DBPedia Spotlight use-case indicates that it needs an explanation of a specific property called the 'popularity score'. This dimension is the model dimension of annotations.

Knowledge about structure and properties of annotations is one thing, knowledge about concepts which are used in annotations are another. Certainly, this issue is well discussed in the Semantic Web domain and a lot of annotation data is produced following the Semantic Web compliant cause of conduct. Nonetheless, this issue is a lot more complicated and numerous examples in the survey give evidence about this fact.

For instance, e-Codicology uses TEI but in many cases this information is not sufficient and encoding principles are necessary. The e-Poetics use-case applies a technical terminology which complies with the rules of literature study. The issue is that no technological representation for this terminology exist. The example of e-Metaphor is even more complicated. E-Metaphor's concept of metaphors is defined very precisely for the purpose of the project. In this case it is the peculiarity of the definition linking to a

7 http://isn3.zrc-sazu.si/avl_arhiv/index.php (registration required)
8 http://www.monasterium.net
9 http://vad.uni-hd.de/
10 http://hkikoeln.github.io/SemToNotes/
11 http://etnofletno.si/
12 http://www.ecodicology.org/

specific type of theory of metaphors which complicates the appropriate use of annotations and which might create misunderstandings. Having all this said, the problem of the semantic dimension of annotations goes far beyond the question if a formal and technical representation of its concepts exist.

An interesting dimension of annotations has already been investigated and named very well by Agosti, Bonfiglio-Dosio, & Ferro (2007). In the cited work the authors remark that the meaning of annotations is often shaped by relations between annotations of the same annotating process. They call this the 'dialogic' aspect of annotations. The use-cases DHWork and Visual Anthropology also highlights this aspect in some of their answers. Accordingly, annotations should always contain information which make it possible to dereference corresponding annotations..

The correct angle to understand annotations is often set by knowing the purpose of an annotation process together with the research goals. There are no better examples for this link than the use of annotations in e-Metaphor and in the DARIAH-DE Fellowship use-case. In both cases annotations are produced in a training process of mining algorithms. As documentation for the development of this algorithm these annotations are incredibly interesting. However, as serious annotations about the annotated object part of them do not serve well.

In the CATMA[13] use-case annotations are created in a crowdsourcing environment. They are meant to be heterogeneous for the purpose to engender a dense description. Thus, their function must not be interpreted as normative classification. Finally, the DHWork use-case remarks that one of its goals is to evaluate the difference between annotation and comment. This distinction shapes what information might be published as annotation data and what is not. Thereby, it puts specific information in a specific context which depends on the definition of annotation. Thus, annotation should reference the results of the research process in which they were created.

The last dimension which significantly shapes the meaning of annotations is their intended audience. This phenomenon was intensively discussed by Chiang (2010). Accordingly, form and content of annotations differ when they are produced to support an individual researcher, a research group or meant to be public. The Visual Anthropology use-case highlights a scenario in which this issue is quite obvious. Annotations in correspondence with field diaries in ethnology often contain information which are ethically problematic. However, the issue exists also in more subtle scenarios.

2.2. Evaluation

Several approaches have already been carried out to systematize context dimensions of digital annotations into consistent models. Before I introduce my own approach I would like to quickly outline the drawbacks that these attempts still possess.

One problem of comparable approaches is that they are often carried out on the ground of settled annotation scenarios. For instance, Chiang (2010) develops a sophisticated 'Annotation Function Coding Scheme' based on 'A Multi-Dimensional Approach to the Study of Online Annotation'. However, in her research online annotations are annotations produced in the interpretative reading process of text based web documents. Thus, the work tackles an individual annotation scenario which is relatively well known.

In other cases the research which is carried out only focus on specific dimensions of annotations. Likewise Bauer & Zirker (2015) tackle the problem of different interpretation levels (called strata before) while Bélanger (2010) or Gradmann et al. (2015) concentrate on the relationship between research practice and annotations.

Approaches like oa:Motivation in the Open Annotation Data Model albeit improved over time by adding oa:hasPurpose to oa:motivatedBy are still limited, inconsistent and contingent as I have argued elsewhere (Walkowski, 2015, 2016). This might relate to the fact that originally motivation was included into the model to provide interesting ways of querying annotation data[14]. The issue of correct interpretation and usage of annotations was not the driving force.

Other approaches which consider a variety of computational annotation scenarios like Agosti et al. (2007) make transparent the complexity of the issue but do not intent to get into greater detail. Finally, many approaches only tackle the topic in prose but not in a formal manner.

In this paper I want to introduce a first attempt to systematize the different context dimensions of annotations which were addressed at least partially in the survey before. For this purpose, I would like to introduce the diagram presented in figure 1.

In the center of the figure there is the annotation which consists of a body and a target. The target addresses both the annotated part of the object as well as the entire object. The body holds the annotation content. The upper half of the figure addresses aspects of practice and semantics while the lower half references aspects of technology and structure.

Furthermore, a production flow exist from the left to the write in which objects for annotations are created, then annotated and annotations are processed. The left half represents both the production of an object to be annotated and the annotation itself. Likewise, goal and publication can identify corresponding activities which belong to annotations or output for which a peculiar annotation is created. Finally, two types of relationships exist between annotation target and annotation body. Depending on the situation, each dimension can be formally instantiated in a way that expresses its contribution to the overall meaning of peculiar annotation.

For instance, the revision of annotations in the visual anthropology use-case caused by ethical concerns is part of the publication dimension while algorithm testing in the e-Metaphor use-case provides a goal dimension. However, the dimension can also be more obvious. A semantic tag which is taken from a RDF taxonomy and references this taxonomy by namespace completely opens up the semantic dimension in annotation bodies.

13 http://www.catma.de/

14 refer to the project wiki for further details at https://www.w3.org/community/openannotation/wiki/

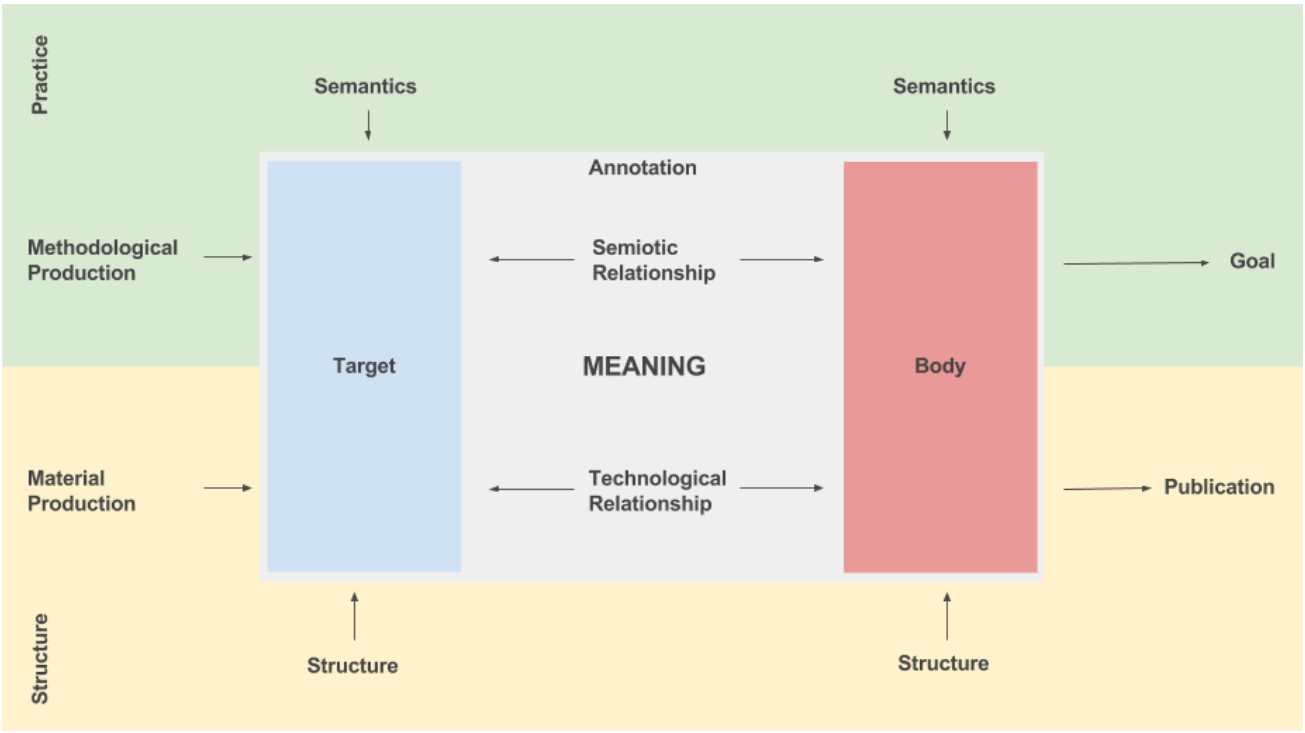


Figure 1: Meaningful context dimensions of annotations and annotated objects.

In other situations in which no technical representation of a formal vocabulary exists the formalization needs to be achieved by using other strategies. Dimensions like publication, goal, methodological or material production of annotations and annotation targets are even more complicated. Standards like Open Annotation are only partially supporting these dimensions albeit their relevance became clear in this section.

3. The Meaning of Annotations

I stated at the beginning of this paper that digital technologies have significantly multiplied the contexts in which annotations are used as a tool for research. Another way to look at this situation is to say that people speak of annotations in situations where th term would not have been applied before.

A productive way to look at this issue implies to conceive of this development as a co-dynamic. In this co-dynamic technologically defined annotation concepts and services are transfered to new research situations, modify the perception and concept of annotating on a theoretical level and are modified themselves by an updated discourse about annotations. For the purpose to illustrate these changes I want to give a few examples.

3.1. New Prospects in Contemporary Annotations

A traditional way to look at annotations implies a hierarchical relationship between the annotation and the object that is annotated. This relationship is visually well illustrated in medieval glosses which are often arranged around the text in the center of the page. The relationship exists also on the level of production. A book is produced for the main text and gives reason to add annotations 'in the margins'.

In definition of the concept of annotations within the Open Annotation Data Model this relationship vanishes. It says:

Annotating, the act of creating associations between distinct pieces of information, […] (Sanderson et al., 2013, p. 1)

The formal semantics still contain the concept of body and target but methodologically there is no necessary difference in the way Open Annotation understands the relationship between body and target.

In the Pelagios[15] project for instance sources are annotated with data about places. However, the services Pelagios provide completely blur this dependency. If the places are annotations to the texts or the other way around depends on one's point of interest.

This has to do with another principle of historical annotations that becomes more and more fragile: the existential dependency of annotations from the annotated object. Annotations exist on the paper of books and vanishes away with it. Digital annotations can be physically stored and disseminated independently. That means annotation data is a primary research output in itself and not solely anymore a documentation of the path that was taken to these results. By using Pelagios, annotations are brought to the level of first class research results.

Roorda (2013) addresses this aspect more explicitly when he calls annotations 'a new paradigm in Archiving'. In his opinion annotations are most importantly the smallest publishable information unit today. Such a use-case which completely abstracts from the linking aspect of annotations and highlights the information structure aspect

15 http://commons.pelagios.org/

is also elaborated in the Wissensspeicher[16] project of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW, 2016).

Roorda introduces another interesting prospect. In his Shebanq project (Roorda, 2016) he uses annotations for queries. In this scenario the annotated object is the volatile entity. The target of the annotation might look differently each time the annotation is rendered. The stable unit is the query as an interface, that means the representation of a certain way to look at things.

A better known issue is the shift of annotations from individual private contexts into open and collaborative spaces (Meister, 2012).

3.2. What is Annotating Today

All the examples demonstrate how fundamentally the scope of annotations and the deployment of annotating have changed by virtue of digital technologies. The level of change might be even more obvious considering that some of the main features of web annotations are just a real world implementation of how hypermedia research has envisioned the world wide web ever since (Carr, De Roure, Hall, & Hill, 1995; van Ossenbruggen, Hardman, & Rutledge, 2006). Measured by the quantity of researchers from hypermedia research which are active in the subject of web annotations this subject is hypermedia research. Thus, a technology and architecture oriented angle is driving our understanding of annotations.

The second question from the beginning of this paper asked: What are annotations today? The motivation is not to be essentialist or restrictive. The issue is that both the sound use of annotating in research and of annotation data as research resources depend on and improve with our understanding about digital annotations.. It is simply a question of methodology. The Digital Humanities community should therefor extend their otherwise often used concept of tools to discourse by definition. The benefit of such an approach is the creation of a deeper level of understanding of what is going on and consequently a more productive use of annotations. Definitions can be changed as Digital Humanists change their tools. It is the struggle which creates the benefit.

Fortunately, at least two recent initiatives aim at evaluating research practices in the Digital Humanities. The first is the Scholarly Domain Model (SDM) which appeared in the EUROPEANA project cluster (Gradmann et al., 2015) and was first defined in the DM2E satellite project[17]. The second initiative comes from the Digital Curation Unit in Athens[18] and started in the NeDiMAH project[19]. Its name is the 'NeDiMAH Method Ontology' (Digital Curation Unit, 2016; Hughes, Constantopoulos, & Dallas, 2016). In DARIAH-EU the DiMPO[20] is trying to use the NeDiMAH Method Ontology (hereinafter NeMO) to make progress on mapping Digital Humanities activities.

While NeMO's strategy adheres to a bottom-up strategy SDM at least partially partially follows a top-down approach. Nonetheless, both are well suited to record, structure and synthesize information about annotating today. Furthermore, both project clusters represent big communities which offer potentially rich content in this respect. In the case of SDM the primer even illustrates the model on the basis of an annotation example.

4. Feasibility, Strategies and Prospects

In this paper I demonstrated that the meaning which is embedded in annotations is fragile and often hard to grasp. I introduced a systematic approach to gain more control over the expression and interpretation of meaning in annotations. In the first case the systematology offers new insights for the application of metadata to annotations. I will come back to this issue below. In the second case it is a tool which can be used to look at or research on the context of annotation data before it is used. That being said, the systematology is still a device of understanding even if annotation data does not provide sufficient metadata.

I also indicated how fundamentally the concept of annotations is changing due to computational environments and argued in favour of broad evaluation of annotation activities. These two topics are only two different topics in the first place. In the long run, they belong to the same effort and contribute to each other. More precisely, the systematization of annotation activities into profiles will greatly enhance the understanding of context in annotation data. It will make it easier and more standard compliant to instantiate and describe context dimensions of annotation data. Likewise, deeper elaboration of the context dimensions systematology will make it easier to map annotation activities and identify profiles.

In the first section I indicated the complexity of annotation data and its reasons. The second section tried to make implicit things explicit and created a formal systematology. Furthermore, I criticized that current standardized annotation models like Open Annotation and its concepts of motivation and provenance are not sufficient. Thus, I am indeed arguing that annotation data needs more metadata applied to it than it is the case today.

However, it is also not feasible to describe annotation data in all its facets. The example of Semantic Web compliant tagging demonstrates that a complete description is not always necessary. The vocabulary in use is referenced implicitly in the namespace of the tag. In contrast, the e-Poetics use-case showed that these implicit means do not hold where no Semantic Web compliant vocabulary exists.

The information density of metadata which needs to be attached explicitly depends very much on aspects like the one that has just been described. It also depends on the structural and semantic model which is used to model annotation data as well as on the conditions for its instantiation in peculiar annotation scenarios. These scenarios create options to go without extra metadata or eliminate these options. Further research is needed to clarify which options for each context dimension exist in relation with which technological environments.

The argument of implicitly given informations can be pushed even further and up to the socio-cultural level. For

16 http://wissensspeicher.bbaw.de/
17 http://dm2e.eu/
18 http://www.dcu.gr/
19 http://nedimah.eu/
20 https://dariahre.hypotheses.org/working-groups/digital-methods-practices-and-ontologies

instance, the use of annotations for data publication in Bioinformatics is a straight-forward and well known practice. This means many informations about context dimensions have become part of common knowledge. In general, the level up to which this tacit knowledge exists for specific annotation scenarios influences the need for explicit metadata.

Computer science distinguishes between technological, structural and semantic interoperability. There is also something like socio-cultural interoperability which refers to questions of how public and consistent things are within a socio-cultural configuration. On the other hand this level of interoperability only exists insofar it is actively designed. In this sense the approach that has been presented in this paper tried to create better conditions for socio-cultural interoperability in annotating. Its success depends on further theoretical systematization of annotation activities in similar environments like DARIAH.

5. Disclaimer

The current study was financed by the Federal Ministry of Education and Research (BMBF) and carried out in the context of Cluster 6 in DARIAH-DE and the DARIAH-EU Working Group Digital Annotation. Special thanks go out to all projects that took the time to provide descriptions of their annotation use-cases.

6. References

Maristella Agosti, Giorge Bonfiglio-Dosio, Nicola Ferro. 2007. A historical and contemporary study on annotations to derive key features for systems design. In: *International Journal on Digital Libraries*, 8(1), pages 1–19. http://doi.org/10.1007/s00799-007-0010-0

Daniel Atkins. 2003. *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*.

Matthias Bauer, Angelika Zirker. 2015. Whipping Boys Explained: Literary Annotation and Digital Humanities. In: *Literary Studies in the Digital Age: An Evolving Anthology*. MLA Commons. https://dlsanthology.commons.mla.org/whipping-boys-explained-literary-annotation-and-digital-humanities/

Marie-Eve Bélanger. 2010. *Annotations and the Digital Humanities Research Cycle: Implications for Personal Information Management*. http://hdl.handle.net/2142/15035

Les Carr, David De Roure, Wendy Hall, Gary Hill. 1995. *The Distributed Link Service: A Tool for Publishers, Authors and Readers*. http://eprints.soton.ac.uk/250739/.

Chia-Ning Chiang. 2010. *A multi-dimensional approach to the study of online annotation* Ph.D thesis.

Stefan Gradmann, Steffen Hennicke, Gerold Tschumpel, Kristin Dill, Klaus Thoden, Alois Pichler, Christian Morbidoni. 2015. *Beyond Infrastructure! Modelling the Scholarly Domain*.

Lorna Hughes, Panos Constantopoulos, Costis Dallas. 2016. Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines. In: S. Schreibman, R. Siemens, J. Unsworth, eds., *A New Companion to Digital Humanities*, pages 150–170. John Wiley & Sons, Chichester.

Jan Christoph Meister. 2012. Crowd Sourcing 'True Meaning': A Collaborative Markup Approach to Textual Interpretation. In: M. Deegan, W. McCarty, eds., *Collaborative Research in the Digital Humanities*, pages 105–122. Ashgate, London.

Jan Chrsitoph Meister. 2015. *All dressed up and nowhere to go? The strategic role of digital humanities annotation tools*. Presentation. Hamburg. https://lecture2go.uni-hamburg.de/veranstaltungen/-/v/18469

Geoffrey Rockwell. 2010. As Transparent as Infrastructure: On the research of cyberinfrastructure in the humanities. In: M. Jerome, eds., *Online Humanities Scholarship: The Shape of Things to Come*, pages 461–487. Rice University Press , Houston.

Dirk Roorda, Charles Heuvel. 2013. Annotation as a new paradigm. In: *Proceedings of the American Society for Information Science and Technology*, 49(1), pages 1-10. Presentation. Baltimore. http://doi.org/10.1002/meet.14504901084

Robert Sanderson, Paolo Ciccarese, Herbert Van de Sompel. 2013. *Designing the W3C Open Annotation Data Model*.

Jacco van Ossenbruggen, Lynda Hardman, Lloyd Rutledge. 2006. Hypermedia and the semantic web: A research agenda. In: *Journal of Digital Information*, 3(1).

Niels-Oliver Walkowski. 2015. *Provenance and Motivation in Open Annotation*. Presentation. Cologne. http://cutuchiqueno.webfactional.com/slides/koeln052015

Niels-Oliver Walkowski. 2016. *Digitale Annotationen: "Best Practices" und Potentiale I,* Report No. 6.2.1 I. Göttingen, DARIAH-DE.

Universal Dependencies for Slavic Languages

Daniel Zeman

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University
Malostranské náměstí 25, 118 00 Praha/Prague, Czech Republic
zeman@ufal.mff.cuni.cz

Universal Dependencies (UD) is an international project that seeks to define guidelines for annotation of morphology and syntax, applicable to all natural languages. Its motivation is to find a common lingua franca for people and tools that deal with annotated linguistic data. Besides guidelines, UD also releases dependency treebanks converted to the UD annotation style; this data is becoming an important resource for multilingual NLP research and applications. In my talk, I will give a brief introduction to UD in general, and then I will focus on phenomena specific to Slavic languages. The UD release 1.2 includes six Slavic languages and others are being worked on. This gives us plenty of material for comparative studies, that in turn can (and should) further contribute to improved cross-linguistic consistency and fine-grained UD guidelines for Slavic languages. I will present observations from the UD data, as well as some results of dependency parsers trained on the data.

Asistent – A Machine Translation System for Slovene, Serbian and Croatian

Mihael Arčan* **Maja Popović†** **Paul Buitelaar***

*Insight Centre for Data Analytics, NUI Galway, Ireland
[firstname.lastname]@insight-centre.org

†Humboldt University of Berlin, Germany
maja.popovic@hu-berlin.de

Abstract

The META-NET research on language technologies in 2012 showed a weak support on tools for crossing the language barrier for many European languages, including the south Slavic languages. Therefore, we describe a statistical machine translation system, called *Asistent*, which enables automatic translations between English, Slovene, Croatian and Serbian. In addition to make this system publicly accessible, we focus on parallel data preparation as well as on using multiple pivot languages for translation quality improvement of the targeted Slavic languages. A comparison of translations generated by the *Asistent* translation system shows a significant improvement of translation quality over *Google Translate*.

1. Introduction

The statistical machine translation (SMT), in particular phrase-based SMT (Koehn et al., 2003), has become widely used to cross the language barrier in the last years. Nowadays, open source tools such as Moses (Koehn et al., 2007) have made it possible to build translation systems for many language pairs, domains or text types within days. Despite the fact that for certain language pairs, e.g. English-French, high quality SMT systems have been developed, a large number of languages and language pairs still suffer from underdeveloped resources. The largest study about European languages in the Digital Age, the META-NET Language White Paper Series[1] in year 2012 showed that only English has good machine translation support, followed by moderately supported French and Spanish. More languages are only fragmentary supported (such as German), whereby the majority of languages are weakly supported. Many of those languages are also morphologically rich, which makes the SMT task even more challenging, especially if translations are performed into the morphologically rich languages. A large part of the weakly supported languages consists of Slavic languages, where Slovene, Serbian and Croatian belong (Krek, 2012). Therefore, we describe *Asistent*,[2] an SMT system, which enables automatic translations between English, Slovene, Croatian and Serbian language. Despite the limited amount of resources and domain variations, specifically among the Slavic languages, we collected existing data and developed a system aimed at supporting human translators and enabling cross-lingual language technology tasks.

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 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 Antonio Toral and Ramírez-Sánchez (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 data sets 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.

[1] http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

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

3. Experimental Setup

The proposed system, called *Asistent*, is a freely accessible translation system, based on the widely used phrase-based SMT framework. It supports translations from English into Slovene, Croatian and Serbian and vice versa. In addition to that, translations between the Slavic languages can be obtained.

3.1. Statistical Machine Translation

Our approach is based on SMT, where we wish to find the best translation $\mathbf{e}$, of a source string $\mathbf{f}$, given by a log-linear model combining a set of features. The translation that maximizes the score of the log-linear model is obtained by searching all possible translations candidates. The decoder or search procedure, respectively, provides the most probable translation based on statistical translation model learned from the training data.

For generating the translation models, we use the statistical translation toolkit Moses (Koehn et al., 2007). Word alignments were built with GIZA++ (Och and Ney, 2003) and a 5-gram language model was built with KenLM (Heafield, 2011).

3.2. Automatic Translation Evaluation

For the automatic evaluation of translations between the targeted languages we report results based on the BLEU (Papineni et al., 2002), Meteor (Denkowski and Lavie, 2014) and the chrF3 (Popović, 2015) metric. The approximate randomization approach in MultEval (Clark et al., 2011) is used to test whether differences among system performances are statistically significant.

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

Meteor is based on the harmonic mean of precision and recall, whereby recall is weighted higher than precision. Along with standard exact word (or phrase) matching it has additional features, i.e. stemming, paraphrasing and synonymy matching.

chrF3 is a tokenisation-independent metric, which has shown very good correlations, especially for morphologically rich(er) languages, with human judgements on the WMT2015 shared metric task (Stanojević et al., 2015), both on the system level as well as on the segment level.

In addition to the described evaluation metrics, we performed an automatic error classification with the help of *Hjerson* (Popović, 2011). The publicly available tool estimates the frequencies of five error types, i.e. morphological (inflectional) errors, word order errors, omissions, additions and lexical errors (mistranslations).

4. Parallel Data Sets

In this section we describe the parallel corpora used to build the translation models as well as the data preparation approach for better translation performance.

Corpus Name	En-Sl	En-Hr	En-Sr	Sl-Hr	Sl-Sr	Hr-Sr
DGT	1.8M	196K	/	/	/	/
ECB	79K	/	/	/	/	/
EMEA	253K	/	/	/	/	/
Europarl	599K	/	/	/	/	/
Gnome	998	5K	126	4K	600K	300K
hrenWaC	/	86K	/	/	/	/
JRC-Acquis	29K	38K	/	/	/	/
KDE	58K	/	32K	85K	49k	33.2k
LangCourse	/	/	3K	/	/	
PHP	1K	/	/	/	/	/
OpenSubtitles	10.1M	16.3M	20.5M	6.1M	13.3M	22.3M
SETimes	/	198K	209K	/	/	200K
Tatoeba	/	777	633	/	/	/
TED	13K	76K	1K	/	/	/
Ubuntu	/	8K	/	557	86K	51K
Training Data	**En-Sl**	**En-Hr**	**En-Sr**	**Sl-Hr**	**Sl-Sr**	**Hr-Sr**
L1 words:	161M	165M	194M	39M	90M	137M
L2 words:	133M	133M	159M	40M	94M	139M
unique L1 w.:	631K	626K	658K	468K	775K	1.22M
unique L2 w.:	1.00M	1.26M	1.37M	579K	966K	1.24M
Par. sentences:	13.1M	16.9M	20.7M	5.5M	12.6M	19.4M

Table 1: Statistics on parallel corpora used to build the translation models accessed by the *Asistent* system (explanation: En-Sl → L1=En, L2=Sl).

4.1. Data Sets Description

The parallel data used to train the SMT system were mostly obtained from the OPUS web site (Tiedemann, 2012), which contains various corpora 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.

Table 1 illustrates the various corpora used to train the *Asistent* system. The upper part of the table shows the original amount of parallel entries in each corpus, while the lower part shows details on the concatenated and preprocessed data set (cf. Subsection 4.3.) used to train the translation models. 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).[3]

4.2. Evaluation Data Set

The in-domain data set used for evaluating *Asistent*'s performance consists of around 2.000 sentences for each language pair of various domains.[4] When translating from or into English, sentences from different corpora[5] were

[3] http://www.opensubtitles.org/

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

[5] 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

	English → Slovene		Slovene → English	
	non-preprocessed	preprocessed	non-preprocessed	preprocessed
BLEU	49.56	**49.97**	61.37	**63.52**
parallel sentences in training data	15.4M	13.1M	15.4M	13.1M
entries in translation model	201M	230M	201M	230M
unique source words in translation model	553K	604K	893K	972K

Table 2: Results on translation quality based on BLEU and statistics on training data and translation models before and after data preparation.

added to the evaluation data set (isolated from the training data set). 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. Data Preparation

The parallel corpora used for the proposed SMT systems were obtained from the OPUS web site, which contains various corpora of different sizes and domains. However, some of the corpora are rather noisy and therefore certain preprocessing steps were performed.

First, for Serbian as a bi-alphabetical language (Cyrillic and Latin), segments containing letters from both alphabets were removed (such segments were very frequent in the OpenSubtitles corpus). Cyrillic and Latin parts were separated, whereby the Cyrillic parts were converted into Latin. The original Latin parts were removed from falsely encoded special characters. The same approach was performed on the Croatian and Slovene corpora as well. After that, for all languages, technical texts were cleaned from segments containing "#", "%" and "@" symbols. In OpenSubtitles, the hyphen signs appearing at the beginning of a sentence were removed in all texts in order to obtain better consistency. Apart from the described conversions, a large portion of Slovak text was removed from the Slovene part of the Tatoeba corpus.

The next step consisted of filtering of all corpora based on the sentence length proportions. The source/target and target/source sentence length proportions were calculated on the preprocessed texts, and the confidence intervals were extracted based on average proportions and standard deviations. Then, for the texts to be cleaned, only the sentence pairs with proportions within the confidence intervals were kept. The confidence intervals based on average proportions and standard deviations were calculated on the preprocessed texts, i.e. Europarl (Koehn, 2005) for Slovene-English and SETimes (Tyers and Alperen, 2010) for Serbian-English and Croatian-English. For all other corpora, all sentence pairs with proportions within the corresponding confidence interval were kept, and the rest was removed. The last step was removing repetitions, i.e. keeping only unique sentence pairs in all corpora.

After preprocessing the data, tuning and evaluation data sets were extracted for each language pair, containing mostly clean segments from a diverse set of domains. Additionally, these data sets were extracted following the findings of (Song et al., 2014). Namely, too short and too long sentences were not included into the data set, with an optimal average sentence length of 25 words (between 10 and 40 words). Due to the heterogeneity of the used data sets, we extracted such segments from the Europarl and SETimes corpora, and shorter segments (5 to 15 words) from OpenSubtitles and technical texts in the IT domain.

Evaluation on preprocessed data set Due to the noisiness of the parallel corpora, we evaluated first the translation quality of translations generated from a translation model, which was learned from the concatenated data obtained directly from OPUS. We compared these results with translations obtained from the translation model learned from the preprocessed data set, using cleaning steps explained in Subsection 4.3. As seen in Table 2, we gain minor improvements in term of BLEU when translating from English into Slovene, but larger improvements are shown when translating from Slovene into English. Although the non-preprocessed training data set contains more parallel sentences (15.4M vs. 13.1M), the amount of entries as well as the vocabulary stored in the translation models based on the preprocessed data set increases. This illustrates that with this data set, more bilingual alignments can be learned in comparison to a non-preprocessed data set.

5. Evaluation

To evaluate the translation quality of our proposed system, we perform an in-domain and out-of-domain evaluation. The first is done on the evaluation set, which is constructed and isolated from the aforementioned preprocessed parallel corpora. The out-of-domain evaluation is performed on a domain, which is not primary used in the training step. We support the evaluation by illustrating the most frequent n-gram mismatches as well as an analysis of error classes.

5.1. In-domain Translation Evaluation

In this section we present the translation evaluation based on the data set isolated from parallel corpora described in Section 4. Table 3 shows the performance of the *Asistent* system for translating text between English, Slovene, Serbian and Croatian. We compare the translation quality in terms of the BLEU, Meteor and chrF metric to the *Google Translate* system.[6] As seen, we significantly outperform *Google Translate* in most of the examined language pairs ($p<0.01$). Only when translating from

[6] `https://translate.google.com/`, translations performed on April 3rd, 2016

	English → Slovene			Slovene → English		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	34.46	32.90	61.75	44.41	40.59	62.94
Asistent	**49.82**	**36.36**	**69.26**	**64.14**	**45.77**	**76.71**
	English → Croatian			Croatian → English		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	29.27	26.87	57.19	46.08	39.35	67.61
Asistent	**42.15**	**33.02**	**64.95**	**48.07**	**40.05**	**68.18**
	English → Serbian			Serbian → English		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	27.48	26.33	55.80	**46.05**	**39.35**	**67.53**
Asistent	**42.47**	**32.63**	**63.59**	42.35	39.11	64.96
	Slovene → Serbian			Serbian → Slovene		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	12.27	16.83	34.27	14.05	18.32	37.16
Asistent	**23.46**	**23.23**	**43.23**	**29.23**	**25.33**	**47.42**
	Croatian → Serbian			Serbian → Croatian		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	59.88	41.72	73.84	64.90	46.11	78.65
Asistent	**67.39**	**46.34**	**77.98**	**70.09**	**48.97**	**80.89**
	Slovene → Croatian			Croatian → Slovene		
	BLEU	Meteor	chrF3	BLEU	Meteor	chrF3
Google	13.47	17.81	37.36	16.07	19.60	39.54
Asistent	**34.63**	**27.02**	**50.98**	**38.64**	**29.78**	**54.98**

Table 3: Automatic translation evaluation based on BLEU, Meteor and chrF for the *Asistent* and *Google Translate* translation system.

Serbian into English, *Google Translate* performs better than the *Asistent* translation system.

Table 4 reports the frequencies of five *Hjerson* error classes, i.e. morphological-inflectional errors (infl), word order errors (order), omissions (miss), additions (add) and lexical errors (lex). The last column represents the overall sum of errors. It can be seen that there is a larger number of inflectional errors when translating from English into a Slavic language, indicating that one of the first steps towards improving the current version of the system should be dealing with morphological generation. Apart from this, a high percentage of mistranslations is present, which is typical for state-of-the-art SMT systems and can be overcome with enlarging the training data set.

In addition to the automatic translation evaluation, we performed a semi-automatic analysis of the most frequent errors based on unmatched words and word sequences. The automatic evaluation tool *rgbF* (Popović, 2012), based on word n-gram F-score, enables the annotation of unmatched n-grams in the automatically generated translations in regards to reference translation. A manual inspection of these n-grams revealed some frequent patterns, which are shown in Table 5. It can be seen that prepositions, conjunctions, relative pronouns and auxiliary verbs are problematic for

	infl	order	miss	add	lex	$\sum$
Slovene → English	1.4	6.4	5.8	5.5	13.8	33.0
English→ Slovene	7.3	6.2	5.9	5.6	16.5	41.4
Croatian → English	2.8	7.3	6.0	5.1	17.8	39.0
English → Croatian	8.7	5.4	5.3	5.3	20.0	44.7
Serbian → English	2.8	8.8	6.4	5.8	18.6	42.4
English → Serbian	8.9	7.9	6.0	6.7	23.5	53.0

Table 4: Identified translation error classes of the *Asistent* system by the *Hjerson* tool for the in-domain evaluation set.

all translation directions. For translations into English, articles and pronouns are frequently problematic since these two classes are non-existing or often omitted in the Slavic languages. For other translation directions, negation and reflexive pronouns represent frequent issues.

Pivot Language Evaluation Additionally to the direct translation (source language → target language) evaluation, we performed an experiment on pivot translation (source language → pivot language → target language). This approach can enable a bridge between languages, when existing parallel corpora are under-resourced (Babych et al., 2007). Due to the language coverage within the *Asistent* system, we could use two pivot languages for our additional translation experiment. Since we do not know in advance which pivot language can contribute most in pivot translation, we perform a mixed approach, where we select the best translation out of the set of translations coming from the different pivot languages. In our approach we translate first the source sentence into a pivot language and use the most probable translation to translate it further into the target language. In this last step we collect the best 100 translations for each source sentence. Since the language model of the target language is same regardless which pivot language is used, we identify out of the set of 200 translations, provided by two different pivot languages, the most accurate target sentence based on the language model probability.

As seen in the last column in Table 6, the pivot translation quality declines mostly for the English-Slovene and English-Croatian language pairs. Only for the English-Serbian translation direction, the mixed pivot approach provides better translation quality compared to the direct translation. Focusing on translations between Slavic languages only, the proposed approach frequently shows improvements over the direct translations for the less resourced Slavic language pairs.

5.2. Out-of-Domain Translation Evaluation

Besides the in-domain translation evaluation, we perform an evaluation on a data set, which differs from the parallel data used to build the translation models. Massive Open Online Courses (MOOCs) have been growing in impact and popularity in recent years. However, the materials are available mostly in English and the translation solutions provided so far have been fragmentary and human-based. Therefore, in addition to the in-domain evaluation campaign, the *Asistent* system has been tested on a set of

Class	English → Slovene	Slovene → English
Article	/	the, a, an$_{ref}$
Preposition	z, v, na, z, s, pri$_{ref}$, o$_{ref}$,	of, in, to, on, for, with, as
Conjunction	in, da, ki, kot, pa$_{ref}$, tudi$_{ref}$, tako$_{ref}$,	that, which
Pronoun	se (reflexive), to	it, I, you, that, this, we
Verb	je, so, bi, sem, bo, bilo	is, are, have, be
Negative particle	ne, ni,	/
Sequence	,+da ,+ki ,+da+je, to+je ,+in, da+bi, da+se,	of+the, in+the, to+the, ,+the
Class	English → Croatian	Croatian → English
Article	/	the, a, an$_{ref}$
Preposition	u, na, za, s, sa, iz,	of, in, to, on, for, with, at$_{ref}$, by$_{ref}$,
Conjunction	i, a, da, koji, koje, koja, kao, kako, te$_{ref}$,	that, and, which
Pronoun	se(reflexive), to, ti,	it, I, you, that
Verb	je, su, biti, bi, će	i, are, be, have, will, has$_{ref}$, was$_{ref}$
Negative particle	ne	/
Sequence	to+je, ,+i, ,+a$_{ref}$, da+se$_{ref}$, bi+se$_{ref}$,	of+the, in+the, to+the, ,+the, ,+and,
class	English → Serbian	Serbian → English
Article	/	the, a, an$_{ref}$
Preposition	u, na, za, s, sa, od, iz$_{ref}$, o$_{ref}$,	of, in, to, on, for, with, at$_{ref}$, by$_{ref}$, as$_{ref}$
Conjunction	i, a, da, koji, koje, kao, što$_{ref}$,	that, and
Pronoun	se(reflexive), to	it, I, you, that, its$_{ref}$
Verb	je, su, će, bi$_{ref}$	is, are, has, was, have, will, be$_{ref}$
Negative particle	ne	/
Sequence	da+se, da+je, to+je, ,+koji, koji+je, ,+a$_{ref}$, da+će$_{ref}$	of+the, in+the, to+the, ,+the$_{ref}$

Table 5: Examples of most frequent unmatched n-grams by the *Asistent* translation system.

Translation Direction	BLEU with Pivot language		Mixed
English → Slovene	21.55 (Hr)	14.60 (Sr)	20.44
Slovene → English	24.20 (Hr)	26.77 (Sr)	28.77*
English → Croatian	19.04 (Sl)	36.23 (Sr)	38.42*
Croatian → English	29.82 (Sl)	44.84 (Sr)	34.87
English → Serbian	19.19 (Sl)	**60.80** (Hr)	**59.91**
Serbian → English	21.64 (Sl)	**52.44** (Hr)	31.39
Slovene → Serbian	18.03 (En)	**25.45** (Hr)	19.68
Serbian → Slovene	**30.21** (En)	**31.97** (Hr)	**35.39***
Croatian → Serbian	24.23 (En)	30.79 (Sl)	25.89
Serbian → Croatian	41.29 (En)	34.49 (Sl)	37.96
Slovene → Croatian	**35.95** (En)	30.25 (Sr)	**40.44***
Croatian → Slovene	**39.81** (En)	**48.62** (Sr)	**52.09***

Table 6: Automatic translation evaluation based on BLEU using pivot language (in brackets; bold results = improved translation quality compared to direct translation; *-improvement over individual pivot translations).

	Croatian → English							
	BLEU	CHRF3	infl	order	miss	add	lex	∑
Asistent (d)	23.7	48.0	2.7	8.2	14.8	3.4	25.7	54.8
Asistent (p)	19.7	44.1	2.7	7.0	17.4	3.5	30.3	60.8
Google	**26.2**	**51.9**	2.7	6.7	13.7	3.1	25.1	51.3
	English → Croatian							
	BLEU	CHRF3	infl	order	miss	add	lex	∑
Asistent (d)	15.6	45.3	9.0	5.4	5.4	9.5	30.3	59.6
Asistent (p)	12.9	38.8	8.4	6.6	9.5	6.7	35.7	66.9
Google	**18.4**	**50.4**	7.9	5.5	2.6	12.2	27.9	56.1
	Serbian → English							
	BLEU	CHRF3	infl	order	miss	add	lex	∑
Asistent (d)	23.0	48.2	2.6	7.8	11.6	4.4	28.6	55.1
Asistent (p)	18.6	42.5	2.6	8.3	14.3	3.9	34.0	63.2
Google	**24.6**	**50.8**	2.7	8.2	10.7	4.2	28.0	53.7
	English → Serbian							
	BLEU	CHRF3	infl	order	miss	add	lex	∑
Asistent (d)	12.8	38.9	8.3	7.0	7.7	6.1	36.4	65.5
Asistent (p)	10.0	33.8	7.6	6.5	11.7	5.6	40.4	70.8
Google	**17.0**	**46.4**	7.9	6.6	4.7	8.9	30.8	59.0

Table 7: Identified translation error classes of *Asistent* by the *Hjerson* tool for the out-of-domain test set (d=direct translation; p=pivot translation).

out-of-domain texts, originating from educational domain, i.e. lecture subtitles from Coursera. However, it should be noted that these data were not available for the Slovene-English language pair.

The results for Serbian-English and Croatian-English are shown in Table 7 in the form of BLEU, chrF3 as well as the aforementioned five *Hjerson* error rates. Additionally we perform the same evaluation for translations generated by *Google Translate*. It can be seen that although the results for *Google* are better for these texts, they are rather close. Differently to the in-domain evaluation, the pivot translation could not improve the translations over the direct translation approach. As for detailed error rates, the main advantage of *Google* is the smaller amount of omis-

sions (miss) and lexical errors (lex), which is usually the case when larger data sets are used.

6. Translation System as a Web Service

The generic translation models built with the default Moses settings are in general very large, and cannot be use in an online scenario. Therefore, to provide a user translations as good and as fast as possible, we limit the length of the source and target translation candidates in the translation models to five-grams.[7] Additionally we filter out those translation candidates, which are below the direct phrase probability p(e|f) of 1.0E-4.[8] With these strategies we exclude more than 80 million entries for the English-Slovene language pair without to significantly decrease the translation quality. At last, we compared the performance between the OnDisk binarization of the translation model (Zens and Ney, 2007) against the Compact implementation (Junczys-Dowmunt, 2012), where the compressed translation model relies on a perfect minimum hash for look-up.

We evaluated the results of an unfiltered translation models (binarized and non-binarized) against translation models, filtered on aforementioned thresholds. Compared to the default setting we observed insignificant differences in terms of BLEU (≈49.6) using thresholds between 1.0E-5 and 1.0E-3. Only when the threshold is set to 1.0E-2 or above, the performance declines in translation quality in terms of the BLEU score. Additionally, we did not detect any significant quality difference between the OnDisk and Compact implementation.

Optimization evaluation Considering the online scenario, we compress the translation models for all language pairs based on the 1.0E-4 threshold (direct translation probability) and binarize it with OnDisk implementation.[9] Table 8 shows the performance of the *Asistent* translation system, comparing unfiltered translation models for each language pair with filtered and binarized ones. As seen, although we reduce the amount of possible translation candidates in the translation models, the BLEU score does not always decreases significantly. In fact, by using the filtered model we observe improvements for the English → Serbian (+3.59 BLEU) and Slovene → Croatian (+1.3 BLEU) translation direction. This indicates that the filtering approach can exclude an extensive amount of misaligned translation candidates in the original models that may cause translation errors. On the other hand, a decrease in performance for English → Croatian (-1.44) has been observed. Nevertheless, the decrease of translation quality for other translation directions is moderate.

Webdemo API service The translation models, which are accessed through the *Asistent* web interface, can also be accessed by third-party tools.[10] When the *Asistent* service receives a translation request in form of a JSON object (upper part of Figure 1), the service queries the translation models for the best candidate translations. A ranked list based on log probabilities of candidate translations (accessible with JSON key $possible\_translations$, seen in the lower part of Figure 1) is generated from the web service and sent back to the user that can select either the best probable translation or a translation among the proposed translations.

Translation Direction	Asistent	org. models	δ
English → Slovene	49.82	**49.97**	-0.15
Slovene → English	**64.14**	63.52	+0.62
English → Serbian	**42.47**	38.88	+3.59
Serbian → English	42.35	**43.79**	-1.44
English → Croatian	42.15	**43.38**	-1.23
Croatian → English	48.07	**48.90**	-0.83
Slovene → Serbian	**23.46**	23.34	+0.12
Serbian → Slovene	**29.23**	28.97	+0.26
Slovene → Croatian	**34.63**	33.3	+1.33
Croatian → Slovene	38.64	**38.73**	-0.09
Serbian → Croatian	70.09	**70.29**	-0.20
Croatian → Serbian	67.39	**67.54**	-0.15

Table 8: Comparison of BLEU scores between default (original) translation models and *Asistent* accessed compressed translation models.

7. Conclusion

This paper presents a publicly accessible SMT system for translating between English, Slovene, Croatian and Serbian, called *Asistent*. Through the publicly accessible web interface and API request, the SMT system can support human translators and enable information access across languages. Based on the automatically extracted evaluation data set, *Asistent* outperformed *Google Translate* for the majority of the targeted translation directions. Furthermore, experiments on pivot translation show improvements in translation quality between closely related Slavic language pairs over a direct translation approach. Our ongoing work focuses on a better combination of the pivot translation and the comparison of feature based (linguistically annotated data sets) and hierarchical (synchronous context-free grammar rules) SMT for the Slavic languages.

8. Acknowledgements

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 and the TRAMOOC project (Translation for Massive Open Online Courses), partially funded by the European Commission under H2020-ICT-2014/H2020-ICT-2014-1 under grant agreement number 644333.

[7] Moses in its default setting aligns maximum seven source/target words.

[8] We tested different thresholds between 1.0E-5 and 1.0E-1, whereby 1.0E-4 showed best performance.

[9] In our experiments we observed that translating only one sentence at a time, the OnDisk implementation performs at fastest. On the other hand, Compact implementation of the translation model performs fastest when translating an entire document. This implementation also benefits more from parallelizing the translation approach.

[10] more on: `http://server1.nlp.insight-centre.org/asistent/rest_service.html`

```
{
  "nbest":"5",
  "translation_direction":"en_sl",
  "method":"phrase_based",
  "text2translate":[
     {
        "source":"Accusations of witchcraft are also common in other African countries."
     }
  ]
}
```

```
{
  "time":"6 wallclock secs ( 0.02 usr  0.01 sys +  5.16 cusr  0.42 csys =  5.61 CPU)",
  "translation_direction":"en_sl",
  "nbest":"3",
  "method":"phrase_based",
  "text2translate":[
     {
        "source":"Accusations of witchcraft are also common in other African countries.",
        "possible_translations":{
           "obtožbe so pogosti tudi v čarovništva , druge afriške države . ":"-9.741",
           "obtožbe čarovništva so pogosti tudi v drugih afriških državah . ":"-9.644",
           "obtožbe o čarovništvu so pogosti tudi v drugih afriških državah . ":"-9.706"
        },
       "best":"obtožbe čarovništva so pogosti tudi v drugih afriških državah . "
     }
  ],
  "key":""
}
```


Figure 1: Illustration of JSON representations provided to and from the *Asistent* translation service.

9. References

Raphael Rubino Antonio Toral 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.

Bogdan Babych, Anthony Hartley, and Serge Sharoff. 2007. Translating from under-resourced languages: comparing direct transfer against pivot translation. In *Proceedings of the MT Summit XI*, pages 412–418, Copenhagen.

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 Linguistics*.

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.

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.

Marcin Junczys-Dowmunt. 2012. Phrasal rank-encoding: Exploiting phrase redundancy and translational relations for phrase table compression. *Prague Bull. Math. Linguistics*, 98:63–74.

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, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT03*, pages 127–133, Edmonton, Canada.

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.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for

Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand.

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`.

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.

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.

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.

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. In *ACL05-DDMT*, pages 41–48, Ann Arbor, MI.

Maja Popović. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–68.

Maja Popović. 2012. rgbf: An open source tool for n-gram based automatic evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 98:99–108.

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. 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.

Xingyi Song, Lucia Specia, and Trevor Cohn. 2014. Data selection for discriminative training in statistical machine translation. In *17th Annual Conference of the European Association for Machine Translation*, EAMT, pages 45–53, Dubrovnik, Croatia.

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 Translati on (WMT-15)*, Lisbon, Portugal.

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.

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.

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

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. ISCA.

Richard Zens and Hermann Ney. 2007. Efficient phrase-table representation for machine translation with applications to online mt and speech translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting*, pages 492–499, Rochester, NY.

Ohranjanje jezikovne zahtevnosti besedil pri prevajanju testov PISA

Špela Arhar Holdt,*♦ Iztok Kosem*♦

* Zavod za uporabno slovenistiko Trojina (CUJT),
Trg republike 3, 1000 Ljubljana
♦ Filozofska fakulteta Univerze v Ljubljani,
Aškerčeva 2, 1000 Ljubljana
spela.arhar@trojina.si, iztok.kosem@trojina.si

Povzetek

Eden glavnih izzivov mednarodnih testiranjih je zagotoviti ohranjanje jezikovne zahtevnosti besedil, še zlasti v primerih, ko testiranja preverjajo jezikovno kompetenco testirancev, kot velja za raziskavo bralne pismenosti PISA. V prispevku argumentirava, da je ohranjanje jezikovne zahtevnosti testov mogoče evalvirati in izboljšati z uporabo korpusnih podatkov o tipičnosti jezikovnih pojavov v realni jezikovni rabi. Z medjezikovno korpusno primerjavo pokaževa, da so slovenski testi spričo redkega besedišča in atipičnih kolokacij zahtevnejši od angleških, in predlagava postopek, s katerim bi bilo mogoče identificirano stanje v prihodnje izboljšati.

Preserving Text Difficulty in Translations of PISA Tests

One of the main challenges of international assessments is to maintain the same level of difficulty of texts across different languages, which is particularly relevant for (reading) literacy assessments such as PISA which test language abilities of participants. This paper argues that the level of text difficulty in different languages can be evaluated and maintained by using corpora, which contain information on (a)typicality of words, phrases etc. in real language use. Using an interlanguage corpus-based comparison we show that Slovene versions of PISA texts are more demanding than their English counterparts considering the amount of rare vocabulary and atypical collocations. Finally, we propose a procedure that could be used to address such problems and ensure that the level of text difficulty is maintained in different languages.

1 Uvod

PISA (*The Programme for International Student Assessment*)[1] je mednarodna raziskava, ki primerja uspešnost 15-letnih učencev na področju matematike, naravoslovja in branja. Raziskava, ki jo je leta 2000 lansirala OECD, je bila v Sloveniji prvič izvedena leta 2006, od takrat pa se ponavlja vsaka tri leta.

Ker gre za mednarodno raziskavo, se med dejavniki vpliva omenjajo – poleg denimo ekonomske razvitosti države oz. lokalnega okolja, vsebine učnih načrtov, motivacije in izvežbanosti šol ter učencev za testiranje – tudi prevodi testov v nacionalne jezike (npr. Grisay et al., 2007; Arffman, 2012; Solano-Flores et al., 2013), pri čemer je med glavnimi izzivi potreba po ohranjanju jezikovne zahtevnosti besedil. Ta potreba je v smernicah za prevajanje in prilagoditev testov izpostavljena s splošnimi vodili (OECD 2010: 11–13), vendar ob tem ni opredeljena metodologija, ki bi omogočila objektivno zagotavljanje in preverjanje zahtevanega stanja.

Namen raziskave, ki jo predstavlja pričujoči prispevek, je zato oceniti, ali se v slovenskih prevodih testov PISA pojavlja sprememba jezikovne zahtevnosti, in razviti postopek, s katerim bi bilo tovrstna težavna mesta prevoda mogoče sistematično detektirati.

2 Ohranjanje jezikovne zahtevnosti testov

Ohranjanje jezikovne zahtevnosti testov v tem prispevku povezujemo s tipičnostjo jezikovnih pojavov v realni rabi. Če se zgodi, da prevajalec pri svojem delu med več jezikovnimi variantami izbere tisto, ki je v primerjavi z izvirnikom v rabi manj tipična – ali iz dveh besed, ki sta vsaka zase sicer v rabi pogosti, tvori besedno zvezo, ki je v rabi redka – je posledično jezik prevoda zahtevnejši od jezika izvirnika. Podobno velja, če prevajalec izbere skladenjsko strukturo, ki je v jeziku prevoda manj tipična za izbrani jezikovni kontekst. Primer, h kateremu se vračamo v razdelku 4.1, je prevod zveze *mobile phone.* Slednja je v angleški jezikovni rabi relativno pogosta in mogoče je predvideti, da z dekodiranjem pomena testirani po večini ne bodo imeli težav. V slovenščini je na voljo več poimenovalnih možnosti, med katerimi je prevajalec izbral *prenosni telefon.* Ker je ta zveza v realni jezikovni rabi redka (precej redkejša od npr. *mobilni telefon*), izbira pomeni manjšo verjetnost, da testirani zvezo pozna, kar lahko povzroči počasnejše in tudi manj uspešno dekodiranje besedila. Na enak način je seveda mogoče predvideti tudi obraten vpliv, tj. da z določeno jezikovno izbiro prevajalec zahtevnost testa zniža.

Če zaradi medjezikovnih razlik ni mogoče zagotavljati povsem primerljive pogostnosti na ravni posameznih besed oz. zvez, pa je uravnoteženost vsekakor treba zagotoviti na ravni besedila kot celote. Navedena naloga v praksi ni enostavno rešljiva, saj je pogostnost jezikovnih elementov v rabi pri pripravi in redakciji prevoda težje sistematično spremljati in medjezikovno primerjati.[2] Kot možno rešitev v tem prispevku predlagava dopolnitev postopka prevajanja testov s statistično primerjavo jezikovnih pojavov, kakor se kažejo v referenčnih korpusih jezika izvirnika in prevoda. V raziskavi po naročilu Pedagoškega inštituta sva predlagano metodo preizkusila na testih bralne pismenosti PISA v letih 2009 in 2012. V nadaljevanju predstavljava metodo in rezultate raziskave, v razpravi pa predlagano metodo oceniva z vidika uporabljenih korpusnih virov.

[1] http://www.oecd.org/pisa/home/

[2] Kot tudi ni mogoče zgolj z jezikovno intuicijo sintetično spremljati drugih jezikovnih značilnosti, ki lahko vplivajo na zahtevnost prevoda, npr. pomenskih specifičnosti izbranih prevodnih ustreznic ali kompleksnosti skladenjskih struktur. Vendar so vsaj na ravni posamezne pojavitve ti izzivi opaznejši, morda tudi zato, ker so bolj pričakovani, obenem pa so tudi lažje rešljivi s pomočjo obstoječih jezikovnih priročnikov.

3 Kvantitativna analiza

3.1 Metoda primerjave korpusnih podatkov

Za namene raziskave sta bila uporabljena referenčna korpusa, ki predstavljata reprezentativna vzorca sodobne slovenščine oz. angleščine: 1,2-milijardni korpus Gigafida (Logar Berginc et al., 2012) za slovenski jezik in 2,1-milijardni Oxford English Corpus (OEC)[3] za angleški jezik. Primerjalna analiza rabe in frekvenc besedišča je bila izvedena z orodjem Sketch Engine (Kilgarriff et al., 2004).[4]

V prvem koraku so bile besede in besedne zveze, ki se pojavljajo v besedilih analiziranih testov, urejene v primerjalne tabele, skupaj z (ustrezno relativiziranimi) podatki o pogostnosti iz obeh korpusov. V analizo so bile zajete polnopomenske besede, saj so funkcijske med najpogostejšimi v obeh jezikih in kot take manj relevantne za primerjavo. Analiza besed in besednih zvez je bila nato dopolnjena s širšimi podatki o kolokacijah, ki zajemajo večji nabor skladenjskih vzorcev in omogočajo primerjavo kombinacij besed, ki se ne pojavljajo neposredno skupaj, npr. glagolov in samostalnikov, glagolov in predlogov ipd.

V tabelah so bili nato označeni primeri, kjer prihaja do največjih odstopanj na ravni pogostnosti rabe: v Tabelah 1 in 2, ki prikazujeta izsek podatkov za test z naslovom *Varnost prenosnih telefonov*, je besedišče, ki se v enem jeziku pojavlja več kot 50 % redkeje (na milijon besed) kot v drugem, natisnjeno okrepljeno.

Tabela 3 prinaša izsek podatkov kolokacijske analize, kjer so bili kot relevantni za nadaljnjo analizo označeni primeri, kjer je statistična moč kolokacije za več kot 25 % nižja od moči kolokacijskega para v drugem jeziku. Kolokacije smo opazovali v razponu od -5 do +5, upoštevana statistika pa je MI.log_f (v orodju Sketch Engine predhodno poznana kot *salience*).[5]

slovenska beseda	bes. vrsta	pogostnost Gigafida	pogostnost na milijon	angleška beseda	bes. vrsta	pogostnost OEC	pogostnost na milijon
mladi	**sam.**	**1430**	**1,2**	(the) young	sam.	78121	37,7
možgani	sam.	54190	45,7	brain	sam.	145120	70,0
nakazati	**gl.**	**28049**	**23,6**	suggest	gl.	562670	271,4
namen	sam.	241383	203,4	purpose	sam.	253924	122,5
napačen	**prid.**	**56106**	**47,3**	wrong	prid.	323375	156,0
napravica	sam.	4734	4,0	gadget	sam.	12190	5,9
navodila	sam.	54336	45,8	**instructions**	**sam.**	**49273**	**23,8**

Tabela 1: Primerjava odstopanj na ravni pogostnosti rabe v korpusih Gigafida in OEC: besede.

slovenska besedna zveza	pogostnost Gigafida	pogostnost na milijon	angleška besedna zveza	pogostnost OEC	pogostnost na milijon
prenosni telefon	**4922**	**4,1**	mobile phone	47752	23,0
protislovno poročilo	**9**	**0,0**	conflicting report	903	0,4
radijski val	2569	2,2	**radio waves**	**1687**	**0,8**
radiofrekvenčno valovanje	1	0,0	radio frequency waves	22	0,0
rak pri otrocih	**121**	**0,1**	childhood cancer	719	0,3
sodobni življenjski slog	153	0,1	modern lifestyles	240	0,1
stanje pripravljenosti	1479	1,2	on standby	2595	1,3

Tabela 2: Primerjava odstopanj na ravni pogostnosti rabe v korpusih Gigafida in OEC: zveze.

slovenska kolokacija	pogostnost Gigafida	pog. na milijon	moč kolokacij	mesto na seznamu	angleška kolokacija	pogostnost OEC	pog. na milijon	moč kolokacij	mesto na seznamu
kupiti / telefon	3513	3,0	52,645	190	**buy / phone**	**1808**	**0,9**	**36,591**	**265**
opazovati / v	11874	10,0	26,98	270	observe / under	1279	0,6	26,479	743
povezan / z	195453	164,7	75,838	5	linked / to	58607	28,3	66,741	3
sevanje / iz	**702**	**0,6**	**23,859**	**380**	radiation / from	3589	1,7	38,282	133
energija / povezava	**290**	**0,2**	**13,962**	**> 1000**	power / communicate	249	0,1	19,684	586
močen / sevanje	455	0,4	41,202	90	**high / emission**	**433**	**0,2**	**28,77**	**252**
razprava / o	55026	46,4	75,163	5	**debate / about**	**964**	**0,5**	**26,547**	**197**

Tabela 3: Primerjava kolokacijske moči v korpusih Gigafida in OEC.[6]

[3] http://www.oxforddictionaries.com/words/the-oxford-english-corpus, uporabljena verzija korpusa je iz februarja 2012.

[4] Korpus Gigafida smo analizirali v lokalni inštalaciji podjetja Amebis, d. o. o., Kamnik, korpus OEC pa v inštalaciji podjetja Lexical Computing Ltd., pri čemer smo za uporabo korpusa OEC pridobili dovoljenje založbe Oxford University Press.

[5] Enačba je navedena v: https://trac.sketchengine.co.uk/raw-attachment/wiki/SkE/DocsIndex/ske-stat.pdf

[6] Podatek o pogostnosti besed oz. mesto kolokatorja na seznamu kolokatorjev ponuja dopolnilo podatku o kolokacijski moči, kar pride prav v primerih tipa *velikopotezna zamisel – ambitious idea*, ki sta po moči primerljivi, vendar pa sta v rabi besedi *velikopotezen* in *zamisel* precej redkejši od *ambitious* in *idea*.

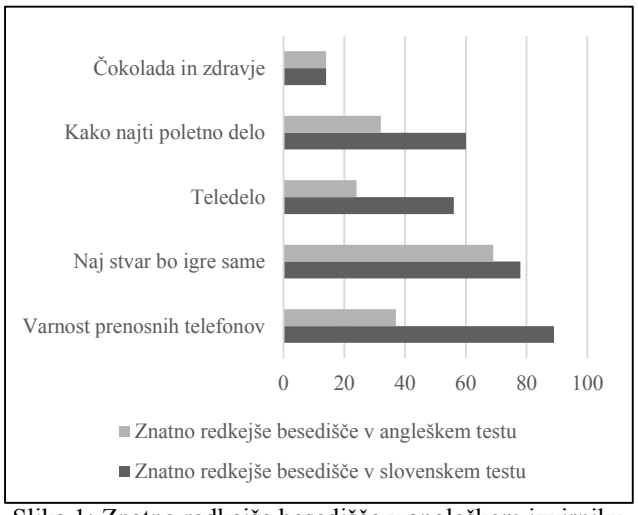
Pri interpretaciji podatkov je treba upoštevati različne vzroke, ki lahko vodijo do razlik v pogostnosti. V prvi vrsti so to razlike v pomenski členjenosti, kot npr. velja za par *napačen – wrong* iz Tabele 1. Za *napačen* referenčni slovarski viri navajajo en sam pomen, medtem ko ima *wrong* poleg slovenščini primerljivega *napačen* še pomen, ki ga v slovenščini izraža *narobe* ('Is anything wrong?' – 'Je kaj narobe?'). Upoštevati je treba tudi slovnične razlike med jezikoma, npr. kategorijo dovršnosti pri interpretaciji pogostnosti angleških in slovenskih glagolov (slovensko *nakazati* in *nakazovati,* angleško *to suggest*). Pomemben dejavnik je tudi struktura in označenost korpusov, npr. pri primeru *mladi* lahko pričakujemo napake pri pripisovanju besednovrstne oznake, in podobno. Nenazadnje je treba izpostaviti, da sta bili meji (50 % redkejša pogostnost besed in zvez ter 25 % redkejša kolokacijska moč) v tej fazi dela izbrani arbitrarno oz. intuitivno in da način primerjave podatkov med korpusoma daje natančnejše rezultate za pogoste primere, pri redkejših pa so prikazane razlike lahko do določene mere napihnjene. Zaradi naštetih razlogov lahko podatke razumemo zgolj kot izhodišče, ki lahko nakaže potencialne probleme, ki pa jih je treba v nadaljevanju natančneje kvalitativno raziskati.

Ker se podatki dotikajo izključno besediščne ravni jezika, uporabljava za predstavitev izsledkov raziskave namesto širšega pojma *jezikovna zahtevnost* poimenovanje *besediščna zahtevnost.*

3.2 Rezultati primerjave

Predstavljeni postopek je bil uporabljen na petih besedilih, ki so se uporabljala za testiranja bralne pismenosti v letih 2009 in 2012:

- *Varnost prenosnih telefonov* (2009) je polstrokovno besedilo na temo varnosti oz. nevarnosti, ki jih (morda) prinaša mobilna telefonija.
- *Naj stvar bo igre same* (2009) je dramski odlomek, v katerem se tri osebe pogovarjajo o tem, kako težko je začeti pisanje gledališke igre.
- *Teledelo* (2009) prinaša izjavi dveh oseb, ki pojasnjujeta svoji mnenji glede dela na daljavo.
- *Kako najti poletno delo* (2012) je letak, ki svetuje mladim, kako poiskati poletno zaposlitev.
- *Čokolada in zdravje* (2012) je kratko polstrokovno besedilo, ki navaja pozitivne učinke epikatehina.



Slika 1: Znatno redkejše besedišče v angleškem izvirniku in slovenskem prevodu testov PISA.

Slika 1 predstavlja rezultate kvantitativne analize, v katerih je združeno število besed in besednih zvez, ki se v enem od korpusov pojavljajo več kot 50 % redkeje kot v drugem, in kolokacij, pri katerih je statistična moč v enem od korpusov vsaj 25 % šibkejša kot v drugem.

Razmerja so od besedila do besedila različna. Besedilo *Čokolada in zdravje* se zdi glede na podatke najbolj uravnoteženo (je pa tudi daleč najkrajše), prav tako je primerljiva pogostnost besedišča v dramskem odlomku *Naj stvar bo igre same.* Pri preostalih treh besedilih je opaziti zelo visoka odstopanja, ki kljub predhodno opredeljenim metodološkim zadržkom dovolj jasno dokazujejo, da je besediščna zahtevnost slovenskih testov v splošnem višja kot pri angleških različicah. Ta ugotovitev sama na sebi seveda ne dokazuje tudi dejanskega vpliva na uspešnost pri testiranju, je pa močan indikator, da je v prihodnje vprašanju pogostnosti (in s tem jezikovne avtentičnosti) pri prevodih nujno posvetiti več pozornosti.

4 Kvalitativna analiza

Ker do razlik v pogostnosti lahko prihaja iz različnih razlogov, morajo biti primeri, ki glede na rezultate v slovenskem testu izstopajo po svoji redkosti, natančneje raziskani. V nadaljevanju predstavljava nekaj izbranih primerov tovrstne dodatne analize.

4.1 Prenosni vs. mobilni telefon

Kot je razvidno iz Tabele 2, se zveza *mobile phone* v OEC pojavi 23-krat na milijon zadetkov, medtem ko se izbrana prevodna ustreznica *prenosni telefon* v korpusu Gigafida pojavi le 4,1-krat na milijon. Pregled korpusnih podatkov pokaže, da sta alternativni možnosti *mobilni telefon* in *mobilnik* v rabi znatno pogostejši in zato morda za prevod ustreznejši: prva se pojavi 31,8-krat na milijon, druga pa 18,4-krat na milijon (Tabela 4). Dodatna potrditev, da je za slovenščino *mobilni telefon* bolj tipična izbira je dejstvo, da se slednja zveza pojavlja tudi kot del daljše zveze *varnost* oz. *nevarnost mobilnih telefonov* (čeprav z nizkim številom pojavitev). Nenazadnje korpusni podatki pokažejo, da se zveza *prenosni telefon* za razliko od *mobilni telefon* uporablja tudi v pomenu 'brezvrvični stacionarni telefon', kar je lahko za razumevanje besedila dodatno obremenjujoče.

Besedna zveza	Pogostnost v Gigafida	Pogostnost na milijon
prenosni telefon	4922	4,1
mobilni telefon	37720	31,8
mobilnik	21808	18,4
varnost prenosnih telefonov	0	0,0
varnost mobilnih telefonov	13	0,0
varnost mobilnikov	1	0,0
nevarnost prenosnih telefonov	0	0,0
nevarnost mobilnih telefonov	10	0,0
nevarnost mobilnikov	0	0,0

Tabela 4: *Prenosni telefon* in *mobilni telefon* v korpusu Gigafida.

4.2 Teledelo vs. delo na daljavo

V nadaljevanju navajamo del besedila *Teledelo,* v katerem so sivo obarvani primeri, ki se glede na opravljeno analizo pojavljajo v obravnavani ediciji testa znatno redkeje. Slika 2 prikazuje izsek v angleškem testu, Slika 3 pa slovenskega. V prikazu so podatki za redko besedišče in atipične kolokacije združeni, kar pomeni, da so primeri, ki se pojavljajo na ravni posamezne besede in obenem kot del zvez in kolokacij, označeni kot večbesedni problem.

TELECOMMUTING

The way of the future

Just imagine how wonderful it would be to "telecommute" to work on the electronic highway, with all your work done on a computer or by phone! No longer would you have to jam your body into crowded buses or trains or waste hours and hours travelling to and from work. You could work wherever you want to – just think of all the job opportunities this would open up!

Molly

Slika 2: Redkejše besedišče v odlomku *Telecommuting*.

TELEDELO

Delo prihodnosti

Predstavljajte si, kako krasno bi bilo imeti "teledelo" in potovati po elektronski avtocesti, pri čemer bi vse delo opravili na računalniku ali prek telefona! Ne bi se vam bilo treba več gnesti na natrpanih avtobusih ali vlakih ali zapravljati dolgih ur za vožnjo v službo in domov. Lahko bi delali, kjer bi hoteli - samo pomislite, koliko priložnosti za delo bi to odprlo!

Maja

Slika 3: Redkejše besedišče v odlomku *Teledelo*.

Oznake opozarjajo na mesta, ki zahtevajo dodaten prevajalski premislek. Pojav primerov *služba – job* in *pomislite – think* je mogoče pripisati razlikam v pomenski členjenosti besed (razdelek 3.1). Za preostale primere je mogoče s spletnim vmesnikom korpusa Gigafida relativno enostavno[7] poiskati v rabi pogostejše ustreznice:[8]

- *kako krasno* (90 pojavitev) – *kako čudovito* (418 pojavitev);
- *natrpan avtobus* (24 pojavitev) – *prepoln avtobus* (33 pojavitev);
- *gnesti/avtobusu* (8 pojavitev) – *drenjati/avtobusu* (12 pojavitev);
- *zapravljati dolge ure* (0 pojavitev) – *zapravljati ure in ure* (5 pojavitev) – *zapravljati čas* (1391 pojavitev);
- *hoteti* (517.056 pojavitev) – *želeti* (882.338 pojavitev);
- *priložnost za delo* (321) – *možnosti za zaposlitev* (1.326 pojavitev) – *delovno mesto* (111.160 pojavitev).

Če k temu dodamo še primerjavo med besedo *teledelo* (222 pojavitev) in zvezo *delo na daljavo* (354 pojavitev), ki omogoča nekoliko bolj tekoč prevod prve povedi besedila, je mogoče na podlagi opravljene analize pripraviti novo različico besedila (Slika 4).[9] S prenosom predstavljene statistične analize na skladenjski nivo bi bilo mogoče zagotoviti, da bi bil prevod še bližje značilnostim slovenskega jezika, vendar že izboljšave na ravni besedišča pripomorejo k berljivosti in vtisu avtentičnosti besedila.

DELO NA DALJAVO

Delo prihodnosti

Predstavljajte si, kako čudovito bi bilo delati na daljavo in se peljati v službo po elektronski avtocesti, pri čemer bi vse delo opravili na računalniku ali prek telefona! Ne bi se vam bilo treba več drenjati na prepolnih avtobusih ali vlakih ali zapravljati časa za vožnjo v službo in domov. Lahko bi delali, kjer bi želeli - samo pomislite, koliko možnosti za zaposlitev bi to odprlo!

Maja

Slika 4: Ponovni prevod odlomka *Delo na daljavo*.

5 Diskusija

Čeprav opravljena raziskava nakazuje, da podatki o razmerjih v pogostnosti jezikovnih pojavov v rabi lahko pripomorejo k ohranjanju jezikovne zahtevnosti prevoda, se ob predlaganem postopku odpira nekaj pomembnih poudarkov. Prvi je, da pogostnosti pojavov v referenčnem korpusu seveda ni mogoče razumeti kot indikator, kateri jezikovni elementi so del dejanske jezikovne rabe mladih. Na eni strani zato, ker referenčni korpusi tipično vsebujejo besedila, s katerimi se srečujejo predvsem odrasli govorci,[10] na drugi strani zato, ker v nobenem primeru na osnovi podatkov o jezikovni recepciji ni mogoče sklepati o jezikovni produkciji. Slednjo je za slovensko šolajočo se populacijo mogoče raziskovati s korpusom Šolar (Rozman et al., 2012; Kosem et al., 2012; Arhar Holdt et al., 2016), vendar le določen del, tj. pisno produkcijo, ki poteka v okviru šolskega pouka.

Za dano nalogo je uporaba referenčnega korpusa utemeljena zato, ker testi PISA prinašajo besedila, povsem

[7] www.gigafida.net, sinonime je mogoče poiskati s seznama kolokatorjev v zavihku Okolica, nato pa preveriti pogostnost in kontekst rabe v konkordančnih nizih, ki so rezultat enostavnega ali naprednega iskanja.

[8] V prispevku navajava nekaj izbranih alternativnih možnosti za vse identificirane pare in iz navedenih primerov je razvidno, da niso vse razlike enako relevantne. Za namene prikaza upoštevava v prevodu na Sliki 4 tudi manj relevantne rezultate (kjer so frekvence obeh različic bodisi zelo nizke bodisi zelo visoke), pri uporabi korpusa za izboljšavo prevoda dejanskih testov pa je treba upoštevati tako razmerja kot frekvence dobljenih rezultatov.

[9] Prevod je provizoričen in služi ponazoritvi možnosti, ki jih prinaša predlagani postopek za avtomatsko identifikacijo besediščno problematičnih mest.

[10] Ob čemer je nujen poudarek, da je slabo raziskano, katera besedila mladi dejansko berejo in katerih ne. Vsekakor je mogoče predvideti, da so med njimi dela s seznamov šolskega branja in mladinska besedila, ki jih je najti na seznamih najbolje prodajanih in izposojanih v knjižnicah. Načrti za nadgradnjo slovenskega referenčnega korpusa že predvidevajo nadgradnjo s tovrstnim gradivom (Krek et al., 2016). O samostojnem branju neumetnostnih besedil je podatkov manj.

primerljiva tistim, zajetim v referenčne korpuse (strokovna besedila, različne publicistične zvrsti, odlomke iz leposlovja, ipd.). V tem smislu je referenčni korpus ustrezen vir za primerjavo pogostnosti oz. redkosti alternativnih ubeseditvenih možnosti ali tipičnosti oz. atipičnosti kolokacij v ciljnem jeziku prevoda. Pri tem je ključna medjezikovna primerjava, ki razkrije najbolj problematična mesta, torej tista, kjer je določen jezikovni pojav v izhodiščni ediciji testa pogost, v ciljnem jeziku pa redek.

Korpusne podatke bi bilo sicer mogoče uporabiti tudi na druge načine, npr. za označevanje besed, ki so v jeziku zelo pogoste, srednje pogoste in redke (West, 1953; Xue & Nation, 1984; Coxhead, 2000). Po Nation in Waring (1997) naj bi imel jezik štiri skupine besed: splošno besedišče, akademsko besedišče, terminološke besede in zelo redke besede. Splošno besedišče naj bi obsegalo približno 2000 besed, ki jih pri komunikaciji redno uporabljamo in naj bi jih poznali vsi materni (ali spoznali vsi nematerni) govorci nekega jezika. Glede na to opredelitev bi bilo od 15-letnika, ki piše test PISA, mogoče pričakovati, da bo poznal večino 2000 najpogostejših besed, medtem ko bo zahtevnost besed izven te skupine naraščala s padanjem pogostnosti rabe. Takšen pregled testov sicer ne bi izpostavil prevodnih težav enako neposredno kot medjezikovna primerjava, zelo uporaben pa bi bil za preverjanje besediščne zahtevnosti nacionalnih preverjanj znanj ter raznovrstnih učnih gradiv oz. za spremljanje, usmerjanje in vrednotenje razvoja pisne produkcije posameznikov, vključenih v šolski proces.

V vsakem primeru se zdi vključitev podatkov o realni jezikovni rabi v prevajanje testov nujna. Raziskava razkriva težave predvsem na kolokacijski ravni, kjer najdemo več primerov, pri katerih je slovenska kolokacija statistično šibkejša kot angleška, kot primerov, pri katerih je angleška kolokacija redkejša oziroma statistično šibkejša od slovenske.[11] Rezultat so besedila, ki na prvi pogled delujejo neproblematična, vendar so zaradi vrste atipičnih kombinacij manj tekoča in »naravna«, kot to velja za angleško edicijo testa. Zveze, ki jih v korpusu Gigafida ni moč najti, se pojavljajo tudi v delu testa z vprašanji oziroma nalogami za učence, kar ima lahko še bolj neposreden vpliv na uspešnost reševanja.

Rezultati raziskave vodijo k ugibanju, ali na besediščno zahtevnost ne vpliva tudi želja prevajalca izbrati najbolj nesporno standardno oz. knjižno ustreznico (kot nakazuje že večkrat omenjeni primer izbire *prenosni telefon* vs. *mobilni telefon*). Na tej točki je razpravo potrebno povezati s problemom pomanjkljivega slovarskega opisa za slovenščino (Gorjanc et. al., 2015), predvsem na ravni predstavitve kolokacij in sinonimije. Kot kažejo rezultati, bi bila v obeh primerih zelo zaželena vključitev podatkov o pogostnosti jezikovnih pojavov v rabi. Čeprav je postopek, ki ga predlagava v prispevku, lahko v vsakem primeru za prevajalca dobrodošla povratna informacija, bi seveda v temelju kazalo poskrbeti za to, da bo imel slednji že v prvem koraku na voljo podatke, ki jih potrebuje. Seveda pa je pri razpravi o težavah in možnih rešitvah potrebna dobršna stopnja previdnosti, saj jih je – tudi zaradi kompleksnega prevodnega postopka, v katerega je vključenih več oseb in vsebuje številna usklajevanja – težko enoznačno določiti.

[11] Nekaj primerov iz testa *Varnost prenosnih telefonov: izražanje genov*, *laboratorijske razmere, neodvisno testirati, protislovno poročilo, poškodba zaradi toplote, zaščitne napravice* in *zmanjšana zbranost.*

6 Sklep

Za korektno pridobivanje in primerjavo rezultatov mednarodnih testiranj morajo biti prevodi testov v smislu jezikovne zahtevnosti primerljivi z izvirnikom. Predlagani postopek primerjalne rabe referenčnih korpusov za identifikacijo spremembe besediščne in kolokacijske zahtevnosti (in avtentičnosti) pri prevajanju testov PISA se je izkazal kot koristen korak pri doseganju tega cilja: s postopkom je mogoče identificirati potencialno problematična mesta pri izbiri prevodnih ustreznic in omogočiti prevajalcu širši, sintetični pogled na zahtevnost izvirnika in besedila. Rezultati raziskave so pokazali, da so slovenski testi v primerjavi z angleškimi na besediščni ravni zahtevnejši in posledično težji za razumevanje. Vključitev predlaganega koraka v postopek priprave testov se torej kaže kot smiselna in potrebna. Čeprav se postopek osredotoča zgolj na enega od številnih dejavnikov vpliva, namreč omogoča izboljšanje stanja z relativno nizkim izhodiščnim finančno-časovnim vložkom, ki pa bi ga bilo z nadaljnjimi prilagoditvami metodologije mogoče še dodatno optimizirati. Pomemben korak za prihodnje delo pa je ugotoviti, v kolikšni meri izboljšave prevoda dejansko vplivajo na rezultate testiranj bralne pismenosti PISA.

7 Zahvala

Analiza testov PISA je bila delno financirana s strani Pedagoškega inštituta, delno pa s strani ARRS v okviru infrastrukturnega programa *Center za uporabno jezikoslovje* pri zavodu Trojina (šifra I0-0051). Zahvala gre tudi strokovnjakom na seminarju Šolskega polja Centra za študij edukacijskih politik za dragocene povratne informacije glede interpretacije pridobljenih rezultatov ter anonimnima recenzentoma prispevka za koristna dopolnila.

8 Literatura

Inga Arffman. 2012. Unwanted Literal Translation: An Underdiscussed Problem in International Achievement Studies. *Education Research International*, 2016 (ID 503824): 1–13.

Špela Arhar Holdt, Iztok Kosem in Polona Gantar. 2016. Corpus-Based Resources for L1 Teaching: The Case of Slovene. V: A. Marcus-Quinn in T. Hourigan, ur. *Handbook on Digital Learning for K-12 Schools*. Springer, v tisku.

Averil Coxhead. 2000. A New Academic Word List. *TESOL Quarterly* 34(2)**:** 213–238.

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek, ur. 2015. *Slovar sodobne slovenščine: problemi in rešitve*. Znanstvena založba Filozofske fakultete UL.

Aletta Grisay, John H.A.L. de Jong, Eveline Gebhardt, Alla Berezner in Beatrice Halleux-Monseur. 2007. Translation equivalence across PISA countries. *Journal of Applied Measurement*, 8(3): 249–266.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz in David Tugwell. 2004: The Sketch Engine. V.: G. Williams in S. Vessier, ur. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004*, str. 105–116, Lorient. Universite de Bretagne - sud.

Iztok Kosem, Mojca Stritar Kučuk, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012.

Analiza jezikovnih težav učencev: korpusni pristop. Trojina, zavod za uporabno slovenistiko.

Simon Krek, Polona Gantar, Špela Arhar Holdt in Vojko Gorjanc. 2016. Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres. *Konferenca Jezikovne tehnologije in digitalna humanistika 2016*, v pripravi.

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.

Paul Nation, Robert Waring. 1997. Vocabulary size, text coverage and word lists. V N. Schmitt, M. McCarthy (ur.) *Vocabulary: Description, Acquisition and Pedagogy, str.* 6–19. Cambridge: Cambridge University Press.

OECD. 2010. *Translation and adaption guidelines for PISA 2012*. Dostop 5. 3. 2016: http://www.oecd.org/pisa/pisaproducts/49273486.pdf

Tadeja Rozman, Irena Krapš Vodopivec, Mojca Stritar Kučuk in Iztok Kosem. 2012. *Empirični pogled na pouk slovenskega jezika*. Trojina, zavod za uporabno slovenistiko.

Guillermo Solano-Flores, Luis Ángel Contreras-Niño in Eduardo Backhoff. 2013. The measurement of translation error in PISA-2006 items: An application of the theory of test translation error. V: M. Prenzel, M. Kobarg, K. Schöps in S. Rönnebeck , ur. *Research on PISA*, str. 71–85. Springer Netherlands.

Michael West. 1953. *A General Service List of English Words*. London: Longman, Green and Co.

Gun-yi Xue, Paul Nation. 1984. A University Word List. *Language Learning and Communication* 3**:** 215–229.

Predstavitveni portal spletnih jezikovnih virov za slovenščino

Špela Arhar Holdt,*♦ Kaja Dobrovoljc,* Iztok Kosem*♦

* Zavod za uporabno slovenistiko Trojina (CUJT),
Trg republike 3, 1000 Ljubljana
spela.arhar@trojina.si, kaja.dobrovoljc@trojina.si, iztok.kosem@trojina.si
♦ Filozofska fakulteta Univerze v Ljubljani,
Aškerčeva 2, 1000 Ljubljana

Povzetek

Prispevek predstavlja Portal jezikovnih virov, rezultat dveh manjših projektov, ki ju je Ministrstvo za kulturo RS sofinanciralo v okviru *Javnega razpisa za sofinanciranje projektov, namenjenih predstavljanju, uveljavljanju in razvoju slovenskega jezika* v letih 2014 in 2015. V prispevku opišemo zasnovo, izvedbo in rezultate projektov. Portal jezikovnih virov je zasnovan kot strukturirana knjižnica videoposnetkov, ki na poljuden način predstavljajo vsebino in strukturo izbranih jezikovnih virov za slovenščino. Za izbrane vire so bila gradiva na portalu v sklopu aktivnosti *Vir meseca* dopolnjena z raznovrstnimi zanimivostmi, ki so v strnjeni obliki na voljo za nadaljnje diseminacijske in izobraževalne namene.

Portal for the Presentation of Language Resources for Slovenian Language

The paper presents the Portal of Language Resources, a result of two small projects funded in 2014 and 2015 by the Ministry of Culture of the Republic of Slovenia under the *Public call for funding projects focused on the presentation, promotion and development of the Slovene language*. The conception and rationale of both projects are discussed, and the results presented. The Portal is conceived as a structured library of videos which in a straightforward and clear manner present the content and structure of different language resources for Slovene. Various interesting aspects of the resources in the Portal were also promoted through the *Resource of the month* activity. All the information coming out of this activity is available for further dissemination and use for educational purposes.

1 Namen portala jezikovnih virov

S prehodom družbe v digitalno dobo se tudi za slovenščino viša število jezikovnih virov, priročnikov in orodij, ki so za rabo prosto dostopni na spletu. Avtorji teh izdelkov vedno uspešneje izrabljajo možnosti novega medija in veliko truda namenjajo zagotavljanju dostopnosti in prijaznosti svojih izdelkov, vendar ti v javnosti pogosto ostajajo premalo opaženi. Za diseminacijo, ki je v projektnih časovnicah tipična zadnja naloga, v praksi namreč pogosto zmanjka časa, izvedba izobraževanj uporabnikov pa običajno presega domet projektov, v katerih se viri pripravljajo. Na drugi strani naraščajoč izziv za uporabnike predstavlja raznolikost razpoložljivih možnosti, tako v smislu namena oz. vsebine kot konkretnih vmesniških rešitev. Tudi v primeru, da uporabniki določen vir na predstavitvah ali izobraževanjih dovolj natančno spoznajo, ob neredni uporabi hitro izgubijo veščine, potrebne za uspešno pridobivanje in interpretacijo jezikovnih podatkov.[1]

Da bi avtorjem jezikovnih virov, priročnikov in orodij poenostavili diseminacijo projektnih rezultatov, uporabnikom pa na enem mestu omogočili pregledno seznanitev z možnostmi, ki so trenutno na voljo, smo s pomočjo *Javnega razpisa za sofinanciranje projektov, namenjenih predstavljanju, uveljavljanju in razvoju slovenskega jezika* Ministrstva za kulturo RS v letih 2014 in 2015 pripravili predstavitveni portal spletnih jezikovnih virov za slovenščino. Portal, ki trenutno predstavlja 15 virov, je dostopen na spletni strani http://viri.trojina.si. Namen prispevka je predstaviti zasnovo, izvedbo in rezultate obeh projektov: *Izdelava spletne strani z opisi jezikovnih virov in orodij za slovenščino ter osnovnimi (video)navodili za njihovo uporabo* ter *Nadgradnja in popularizacija predstavitvenega portala spletnih jezikovnih virov za slovenščino*.

2 Poglavja s predstavitvenimi posnetki

Glavni doprinos Portala jezikovnih virov so izobraževalni posnetki, ki na zgoščen, poljuden način predstavljajo vsebino in strukturo posameznega obravnavanega jezikovnega vira, in posnetki, ki na konkretnih primerih uporabe kažejo, kako lahko v določenem viru najdemo odgovor na specifično jezikovno vprašanje. Posnetki so urejeni v obliki spletnega portfelja, v katerem vsako poglavje prinaša tudi povezavo na obravnavani jezikovni vir in projektno stran ter informacije o projektu in avtorjih.

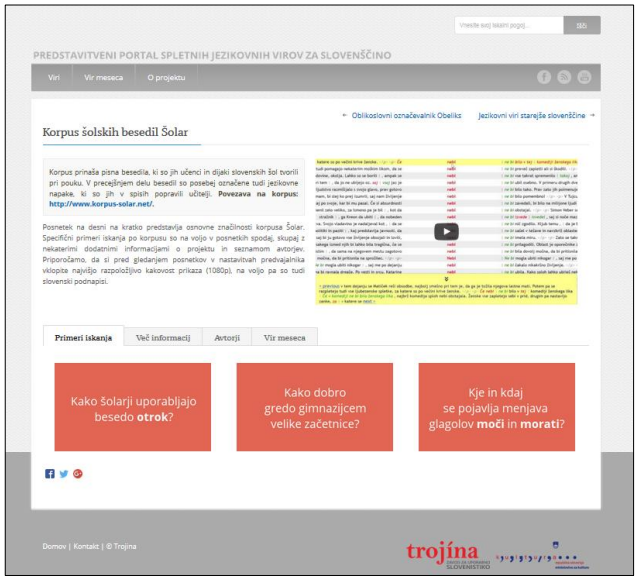
Prioriteta pri pripravi posnetkov je bila zagotoviti optimalno uporabniško izkušnjo na različnih vrstah naprav, od računalnikov z velikimi zasloni do mobilnih telefonov z majhnimi. Proces editiranja posnetkov je bil v večji meri pogojen s prikazom na telefonih, za katerega je bilo treba zagotoviti ustrezno približevanje obravnavanim delom ekrana, vendar na način, da uporabnik pri gledanju posnetka ne izgubi občutka za vmesnik kot celoto. Na drugi strani smo širšo dostopnost vsebin za različne skupine uporabnikov in v različnih situacijah rabe skušali omogočiti tudi s pripravo slovenskih in angleških podnapisov.

Posnetke smo pripravljali po naslednjem postopku: (I) pregled obravnavanega vira, preizkus iskalnih možnosti, rezultatov itd., (II) izbira reprezentativnih primerov in pisanje prve različice scenarija, (III) testno snemanje in popravljanje scenarija, (IV) snemanje posnetka, (V) editiranje posnetka, npr. krajšanje, približevanje, dodajanje

[1] V tem smislu indikativne so povratne informacije udeležencev jezikovnotehnološkega izpopolnjevanja učiteljev, ki je potekalo po slovenskih šolah v letih 2013 in 2014 (http://ucitelji.sdjt.si/). Učitelji, ki so nove možnosti po večini ocenjevali zelo pozitivno, so kasneje poročali, da je novih vsebin preveč in so preveč raznolike, da bi imeli nad njimi celovit pregled.

napisov,[2] (VI) izvoz in objava na kanalu *YouTube*, (VII) dodajanje podnapisov in (VIII) vgradnja novega posnetka v ustrezno poglavje na strani portala. Pred objavo posnetka na portalu smo zbrali in upoštevali tudi povratne informacije predstavnikov avtorjev obravnavanih virov.

Slika 1 predstavlja poglavje, posvečeno korpusu šolskih besedil Šolar. Desno zgoraj je na voljo posnetek, ki pregledno predstavi glavne značilnosti korpusnega vmesnika in možnosti za izvedbo različnih vrst iskanj. Spodaj so na voljo povezave do treh krajših posnetkov, ki kažejo, kako poiskati in interpretirati korpusne podatke, da dobimo odgovor na specifična jezikovna vprašanja (npr. »Kje in kdaj se pojavlja menjava glagolov *moči* in *morati*?«). S klikom na zavihke med gornjim in spodnjim delom ekrana dostopamo do osnovnih informacij o jezikovnem viru in projektu, podatkov o avtorjih projekta, pri nekaterih poglavjih pa so na voljo tudi dodatne zanimivosti, ki so bile pripravljene v okviru akcije *Vir meseca* (več o tem v razdelku 4).



Slika 1: Poglavje na Portalu jezikovnih virov.

3 Predstavljeni jezikovni viri

Vire za predstavitev na portalu smo izbrali ob upoštevanju izkušenj z izvedbo jezikovnotehnološkega izpopolnjevanja učiteljev (Stritar in Dobrovoljc 2013), rezultatov spletnega vprašalnika (http://viri.trojina.si/drugi-viri) ter zanimanja avtorjev jezikovnih virov za sodelovanje pri projektu. V nadaljevanju naštevamo trenutno predstavljene vire, skupaj z referenčno literaturo in povezavo na spletno mesto:

- referenčni pisni korpus Gigafida (Logar et al., 2012; www.gigafida.net),
- korpus govorjene slovenščine GOS (Verdonik in Zwitter Vitez, 2011; www.korpus-gos.net),
- korpus šolskih besedil Šolar (Kosem et al., 2012; www.korpus-solar.net),
- slovarski portal Fran (Ahačič et al., 2015; www.fran.si),
- slovarski portal Termania (Romih in Krek, 2012; www.termania.net),
- rezultati projekta Viri starejše slovenščine IMP (Erjavec, 2015; http://nl. ijs.si/imp/),
- rezultati projekta Signor (Vintar et al., 2012; http://lojze.lugos.si/signor/),
- kolaborativni slovar Razvezani jezik (Dolar, 2014; http://razvezanijezik.org/),
- Jezikovna svetovalnica ISJFR ZRC SAZU (Dobrovoljc in Bizjak Končar, 2015; http://isjfr.zrc-sazu.si/svetovalnica#v),
- Terminologišče ISJFR ZRC SAZU (Žagar Karer, 2015; http://isjfr.zrc-sazu.si/terminologisce#v),
- jezikovnodidaktični Pedagoški slovnični portal (Arhar Holdt et al., 2016; www.slovnica.slovenscina.eu),
- leksikon besednih oblik Sloleks (Dobrovoljc et al., 2015; http://www.slovenscina.eu/sloleks),
- oblikoslovni označevalnik Obeliks (Grčar et al., 2012; www.slovenscina.eu/tehnologije/oznacevalnik),
- digitalizirani starejši pravopisi (http://www.trojina.org/pravopisi),
- in digitalizirane starejše slovnice (http://www.amebis.si/slovnice).

Čeprav gornji seznam seveda ni popoln in zaključen, v trenutnem obsegu predstavlja solidno izhodišče za rabo portala v (samo)izobraževalne oz. diseminacijske namene. V nadaljevanju si želimo portal dopolnjevati z novimi poglavji, izziv pa predstavlja tudi posodabljanje vsebin pri virih, ki se razvijajo, k čemur se vračamo v razdelku 5.

4 Diseminacijska aktivnost Vir meseca

V letu 2015 je potekala v časovnem smislu najbolj obsežna projektna aktivnost, ki smo jo imenovali *Vir meseca*. Namen akcije je bil kontinuirano (v trajanju enega meseca) opozarjati širšo javnost na izbrani jezikovni vir in s pomočjo zanimivosti izobraževati uporabnike ter jih motivirati, da vire v praksi preizkusijo.

Za promocijo so bili izbrani: korpus Šolar (april), Viri starejše slovenščine IMP (maj), korpus Gigafida (junij), korpus GOS (september), portal Termania (oktober) in slovarski iskalnik Fran z Jezikovno svetovalnico in Terminologiščem ISJFR ZRC SAZU (november). Promocija je potekala po treh glavnih kanalih: prek predstavitvene strani na omrežju Facebook (https://www.facebook.com/jezikovniviri), v obliki mesečnih e-novičk in na poštnem seznamu SlovLit (https://mailman.ijs.si/mailman/listinfo/slovlit).

Tekom akcije smo vsakega od zgoraj navedenih virov promovirali z rednimi (tipično tremi na teden) objavami na omrežju Facebook. Objave so vsebovale zanimivosti glede priprave vira, informacije o gradivu (npr. različne statistke o jeziku, besedne oblake in druge vrste slikovnega gradiva), predstavitvene posnetke, ideje za uporabo vira v didaktične namene, kratke posnetke pogovora z avtorji in podobno.[3]

[2] Pri prvem od projektov smo za snemanje uporabili program *Community Clips,* pri drugem *Debut Video Capture Software*, ki je za razliko od prejšnjega plačljiv, vendar omogoča globlje zajemanje barv na zaslonu in nekatere dodatne funkcije, kot npr. barvno označevanje premika miške ali glasovno okrepljeno klikanje. Za editiranje posnetkov smo uporabili *iMovie*.

[3] Akcija je potekala v sodelovanju s predstavniki avtorjev virov, ki so posredovali statistične podatke in zanimivosti v zvezi z gradnjo, svoje projektne izkušnje in vizijo za prihodnji razvoj. Sodelovali so: dr. Tadeja Rozman (korpus Šolar), dr. Tomaž Erjavec in Katja Zupan (viri starejše slovenščine), dr. Nataša Logar (Gigafida), dr. Ana Zwitter Vitez (GOS), Miro Romih (Termania), dr. Helena Dobrovoljc (Jezikovna svetovalnica) in dr. Mojca Žagar Karer (Terminologišče). Ostali avtorji (po večini gre za projekte z visokim številom sodelujočih) so našteti v ustrezajočih poglavjih na spletni strani portala.

V nadaljevanju predstavljamo nekaj primerov gradiva, ki je bilo pripravljeno za promocijo korpusa Gigafida. Sliki 2 in 3 predstavljata besedna oblaka, ki sta bila uporabljena za ponazoritev, kako lahko pri pouku slovenščine podatke o sopojavljanju besed uporabimo za obravnavo sinonimije v jeziku. Pomen pridevnikov *ekonomski* in *gospodarski* je podoben, vendar v rabi nastopata z različnimi samostalniki (ali z istimi samostalniki različno pogosto).

Slika 2: Besedni oblak: *ekonomski* (korpus Gigafida).

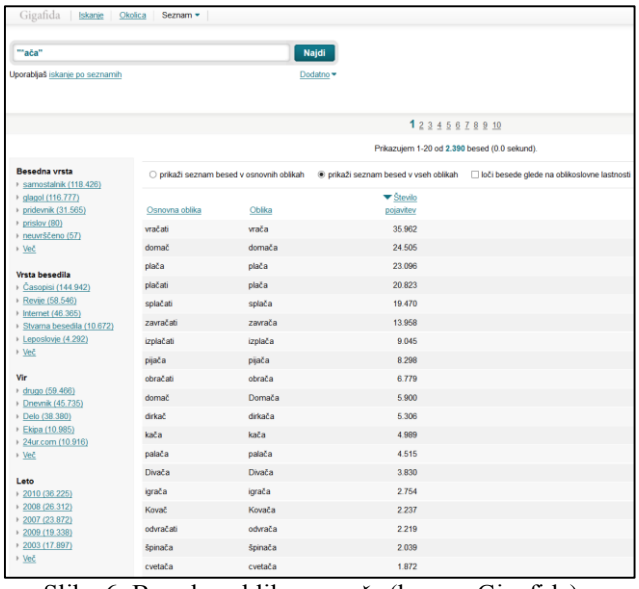
Slika 3: Besedni oblak: *gospodarski* (korpus Gigafida).

Kot ideja za rabo korpusnih podatkov v jezikovnodidaktične namene so bili predstavljeni tudi besedni oblaki najpogostejših pojavitev pri posamezni besedni vrsti. Slika 4 predstavlja podatke za pridevnik in Slika 5 za prislov.

Slika 4: Besedni oblak: pridevniki (korpus Gigafida).

Slika 5: Besedni oblak: prislovi (korpus Gigafida).

Slika 6 predstavlja še zadnji primer, in sicer ponazoritev, kako lahko vmesnik korpusa Gigafida uporabimo za iskanje rim. V zavihku *Seznam* lahko poiščemo besede, ki vsebujejo enak niz črk. Če v iskalno okence npr. vnesemo **ača*, dobimo besede, ki se končajo na -ača (*plača, pijača, palača, igrača*). Če pogoj postavimo še v narekovaje, dobimo seznam vseh ustrezajočih besednih oblik (npr. *vrača, domača, plača, dirkača*).



Slika 6: Besedne oblike na *-ača* (korpus Gigafida).

Zaenkrat nerealizirana je ostala ideja, da bi v promocijo vključili tudi ankete, vprašalnike in kvize za uporabnike, vendar ocenjujemo, da je bila v smislu raznolikosti in privlačnosti gradiva akcija kljub temu uspešna: skupno smo pripravili vsebine za več kot 70 objav. Od začetka akcije (april 2015) do časa pisanja prispevka je stran na Facebooku všečkalo 911 uporabnikov, kar je v skladu s pričakovanji, spodbudno pa je predvsem, da je bilo v času promocije število všečnikov, kot tudi njihov odziv (klikanje na vsebino) v stabilnem porastu. Sočasno s pripravo strani na Facebooku smo na Portal jezikovnih virov dodali možnost, da se uporabniki naročijo na mesečne e-novičke, s strnjenim povzetkom objavljanjih vsebin. Adrema, ki sestoji iz učiteljev, sodelujočih na jezikovnotehnoloških izpopolnjevanjih, ter novih naročnikov, je ob koncu akcije obsegala 910 uporabnikov (v določenem fragmentu so prekrivni s sledilci na Facebooku, vendar večinoma ne).

Za izdelavo mesečnih novičk smo izbrali program *MailChimp*, ki ob kreiranju poštnega sporočila ustvari spletno mesto, na katerem ostane vsebina sporočila trajno dostopna. Z uporabo te možnosti smo promocijske vsebine vključili na spletno stran portala, kjer so prosto na voljo za nadaljnje promocijsko-diseminacijske in izobraževalne aktivnosti.

Slika 7: Arhiv *Vir meseca* na Portalu jezikovnih virov.

Za konec navajamo Tabelo 1, v kateri so podatki o številu ogledov pripravljenih posnetkov do 17. maja 2016. Podatki so pridobljeni s kanala Youtube in združujejo oglede osnovnih predstavitvenih posnetkov, kratkih posnetkov na temo izbranih jezikovnih vprašanj in intervjujev, ki so bili pripravljeni v akciji *Vir meseca*. Podatki posredno razkrivajo, za spoznavanje katerih virov je med uporabniki največ interesa, in nakazujejo mesta, ki bi se jim bilo v prihodnje smiselno dodatno posvetiti.

Število ogledov	**Osnovna predstavitev**	**Kratki posnetki**	**Intervju z avtorjem**	**Skupaj - vir**
Termania	79		69	148
Signor	44			44
Jezikovna svetovalnica	188		94	282
Pedagoški slovnični portal	125			125
Razvezani jezik	107			107
Terminologišče	94		74	168
Gos	189	129		318
Gigafida	208	171	173	552
Šolar	251	200	211	662
Sloleks	248	134		382
IMP	105	149		254
Fran	237			237
Obeliks	128			128
Starejše slovnice	61			61
Starejši pravopisi	78			78
Skupaj - portal				**3546**

Tabela 1: Število ogledov posnetkov na kanalu Youtube.

5 Zaključek in nadaljnje delo

Po zaključku projektov je Portal jezikovih virov dosegel velikost, ko ga je mogoče uporabljati kot samostojen vir za (samo)izobraževalne namene. Ker so vsebine pripravljene za širšo javnost in optimalno uporabniško izkušnjo na različnih vrstah naprav, je vrednost predstavitvenih posnetkov velika. Potencial za nadaljnjo rabo imajo tudi zanimivosti, zbrane za akcijo Vir meseca, ki bi jih bilo mogoče preoblikovati, da bi bile neposredno uporabne za izobraževalne namene.

Glavna naloga za naprej je dopolniti portal z novimi poglavji in z obstojem vsebin seznaniti čim širši nabor potencialnih uporabnikov. Projektno financiranje razvoja portala se je sicer zaključilo, zato bi bilo v nadaljevanju k pripravi vsebin smiselno aktivneje vključiti zainteresirane avtorje virov, ki bi prek obstoječe platforme lahko ponudili lastne diseminacijske vsebine. Kontinuirano sodelovanje bi bilo dobro vzpostaviti tudi z avtorji virov, ki se nadgrajujejo ali pogosteje spreminjajo oz. dopolnjujejo. Za slednje bi bilo mogoče v predstavitvena poglavja vnesti posebno rubriko, kjer bi bile sproti predstavljene nadgradnje oz. posodobitve. Tako bi zagotovili ažurnosti in uporabnost portala tudi v prihodnje.

6 Zahvala

Projekta *Izdelava spletne strani z opisi jezikovnih virov in orodij za slovenščino ter osnovnimi (video)navodili za njihovo uporabo* in *Nadgradnja in popularizacija predstavitvenega portala spletnih jezikovnih virov za slovenščino* je sofinanciralo Ministrstvo za kulturo Republike Slovenije v sklopu *Javnega razpisa za sofinanciranje projektov, namenjenih predstavljanju, uveljavljanju in razvoju slovenskega jezika* v letih 2014 in 2015 (JPR-UPRS-2014 in JPR-UPRS-2015). Posebej se zahvaljujemo avtorjem jezikovnih virov, ki so sodelovali pri pripravi promocijskega gradiva: Helena Dobrovoljc, Tomaž Erjavec, Nataša Logar, Miro Romih, Tadeja Rozman, Katja Zupan, Ana Zwitter Vitez in Mojca Žagar Karer, ter anonimnima recenzentoma prispevka za koristna dopolnila.

7 Literatura

Kozma Ahačič, Nina Ledinek in Andrej Perdih. 2015. Portal Fran – nastanek in trenutno stanje. V: M. Smolej, ur., *Slovnica in slovar – aktualni jezikovni opis*, *Obdobja 34,* str. 57–66. Znanstvena založba Filozofske fakultete, Ljubljana.

Špela Arhar Holdt, Iztok Kosem in Polona Gantar. V pripravi. Corpus-based resources for L1 teaching: The case of Slovene. V: A. Marcus-Quinn, ur.: *Handbook on Digital Learning for K-12 Schools.* Springer.

Helena Dobrovoljc in Aleksandra Bizjak Končar. 2015. Pravopisno slovaropisje na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU. *Slavia Centralis*, 8(1): 34–50.

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

Kaja Dolar. 2014. Kolaborativni slovar Razvezani jezik. *Slavistična revija* 62(2): 235–252.

Tomaž Erjavec. 2015. The IMP historical Slovene language resources. *Language resources and evaluation* 49/3, str. 753–775.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik V: T. Erjavec, J. Žganec

Gros, ur., *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012: zbornik 15. mednarodne multikonference. Informacijska družba – IS 2012, zvezek C,* str. 89–94. Institut Jožef Stefan, Ljubljana.

Iztok Kosem, Mojca Stritar Kučuk, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012. *Analiza jezikovnih težav učencev: korpusni pristop*. Zavod za uporabno slovenistiko Trojina, Ljubljana.

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*. Zavod za uporabno slovenistiko Trojina; Fakulteta za družbene vede, Ljubljana.

Miro Romih in Simon Krek. 2012. Termania – prosto dostopni spletni slovarski portal V: T. Erjavec, J. Žganec Gros, ur., *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012: zbornik 15. mednarodne multikonference Informacijska družba - IS 2012, zvezek C,* str. 163–166. Institut Jožef Stefan, Ljubljana.

Mojca Stritar in Kaja Dobrovoljc. 2013. Korpusi na poti v šole: jezikovnotehnološko izpopolnjevanje učiteljev. *Slovenščina 2.0,* 1(1): 181–194.

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

Špela Vintar, Boštjan Jerko in Marjetka Kulovec. 2012. Compiling the Slovene sign language corpus. V: *8th International Conference on Language Resources and Evaluation, 21–27 May 2012, Istanbul, Turkey. LREC 2012: proceedings*, str. 159–162. ELRA, Istanbul.

Mojca Žagar Karer. 2015. Terminologišče – kraj, kjer terminolog išče. *Slavia Centralis*, 8(1): 22–33.

Integrating Natural Language and Formal Analysis for Legal Documents

Shaun Azzopardi,[*] Albert Gatt,[†] Gordon J. Pace[*]

[*]Department of Computer Science
Faculty of ICT
University of Malta
shaun.azzopardi@um.edu.mt
gordon.pace@um.edu.mt

[†]Institute of Linguistics
University of Malta
albert.gatt@um.edu.mt

Abstract

Although much research has gone into natural language legal document analysis, practical solutions to support legal document drafting and reasoning are still limited in functionality. However given the textual basis of law there is much potential for NLP techniques to aid in the context of drafting legal documents, especially contracts. Furthermore, there is a body of work focusing on the formal semantics of norms and legal notions which has direct applications in analysis of such documents. In this paper we present our attempt to use several off-the-shelf NLP techniques to provide a more intelligent contract editing tool to lawyers. We exploit these techniques to extract information from contract clauses to allow intelligent browsing of the contract. We use this surface analysis to bridge the gap between the English text of a contract and its formal representation, which is then amenable to automated deduction, specifically it allows us to identify conflicts in the contract.

1. Introduction

Many fields of inquiry that have traditionally fallen under humanistic studies have benefited from the development of language technologies. For example, computational linguists have investigated several aspects of literary text and, more generally, "creativity", yielding important insights into the mechanisms underlying the creation of such texts (Gervás, 2013). One area of the humanities that has been the focus of increasing interest in recent years is legal studies.

The legal profession is wide and varied, however a good amount of legal work involves the drafting of documents, specifically contracts. Contracts are themselves linguistic artefacts and amenable to NLP techniques such as keyword and named entity extraction. Tools which leverage such NLP techniques to bridge between the linguistic "surface" structure of a contract and the underlying technical and logical content, have the potential to do for the drafting of contracts what intelligent design solutions have done for architects and engineers. Some tools already exist that provide some form of help to the contract drafter, e.g. (Gabbard et al., 2015); however what is lacking is an actual analysis of the semantics of the text, that is, a bridge between the language and the underlying semantics.

Contracts have drawn the attention of researchers interested in the formalisation of some of their features, such as norms, rights and obligations, often using some form of deontic logic (Fenech et al., 2009; Gao and Singh, 2014). Such formalisations constitute an abstraction of core aspects of the content of a contract, supporting reasoning and detection of errors and/or conflicts. However there remains a gap between the linguistic "surface" of a contract document, and its underlying structure and semantics. As a result, software tools that genuinely support contract editing, by providing on-demand analysis and reasoning of the linguistic content of a contract, remain scarce.

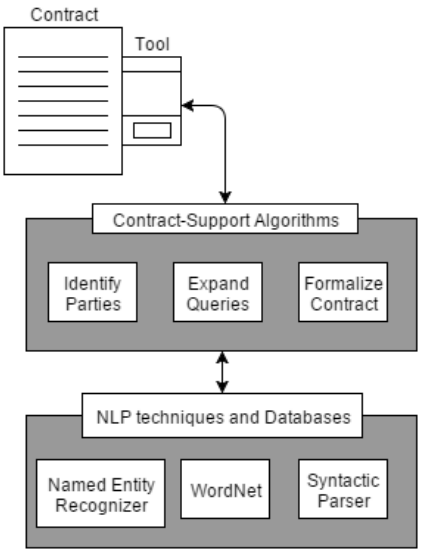


Figure 1: Three-layered (UI, Contract-Support, and NLP and Databases) tool architecture.

The present paper seeks to address this gap. In particular, we describe work on an intelligent contract editor which (a) exploits well-understood NLP techniques to extract information from the text as it is being drafted, using this (b) to enable intelligent browsing of contract clauses, and (c) to automatically construct a partial formal representation that supports automated reasoning about the core elements of the contract.

2. Architecture

In designing the tool we were motivated by the need to have an extensible architecture in order to support the integration of further external tools in order to allow the inclusion of new features in the tool and to improve the quality of the analysis we are already performing. As shown in Figure 1, the architecture is split into three modules that encapsulate the UI, the contract-support algorithms, and the off-the-shelf NLP tools and databases. In each module, a component-based approach is also taken, keeping each algorithm separate, and using dependency injection to enable exchange of these components without the need to change and re-compile the system.

Off-the-shelf NLP tools, such as dependency parsers and part-of-speech taggers (with multiple tools included in both cases), are also included in the system. To integrate them into our system (written in C#), the architecture exploits loosely coupled modules and C# wrappers for the off-the-shelf tools to enable interfacing with the tool.

3. Information Extraction

Our contract-editing tool is implemented as an add-in to Microsoft Word, allowing analysis of the contract side-by-side with contract editing. For this we exploit several information extraction algorithms, whose output we then associate with each contract clause as a set of features.

We developed an algorithm that uses a mixture of regular expressions and named entity extraction to identify the parties to the contract. Since the structure of contracts depends on the drafter, rather than on some universal template, this is not always effective; thus we allow the user to specify these themselves to further refine the outcomes.

We use keyword and named entity extraction in this manner, such that each clause is labelled with the keywords specific to it and the named entities mentioned by it, along with the parties mentioned.

These sets are used to enable the user to browse the contract in question quickly, by highlighting the specific keyword, party, and/or entity in question. This can be useful, for example, to identify clauses which talk about a certain party or a specific concept (e.g. clauses involving a payment, or involving a specific party).

Drafting contracts can also require or benefit from cross-referencing against a body of legal text, including a country's laws or other relevant legal documents. Thus, our tool also includes the capacity to search through laws and related documents. (For the time being, this functionality uses a database consisting of the laws of Malta.) A database containing information about companies is also included, to allow cross-referencing with official company details which are important to get right in a contract.

To improve on both the results of these searches and the clause browsing we employ query expansion. Here, a user's search term is expanded using synonyms, as well as hypernyms and hyponyms which are within a fixed distance from the query in the Wordnet lexical database (Miller, 1995).

4. Formal Analysis

Extracting sufficient information from a contract to support such reasoning remains an understudied problem. Indeed, most approaches to legal texts that apply NLP techniques tend to view the task as a form of information retrieval whose results are insufficient to support automated reasoning (Gao and Singh, 2014; Dragoni et al., 2015; Wyner and Peters, 2011).

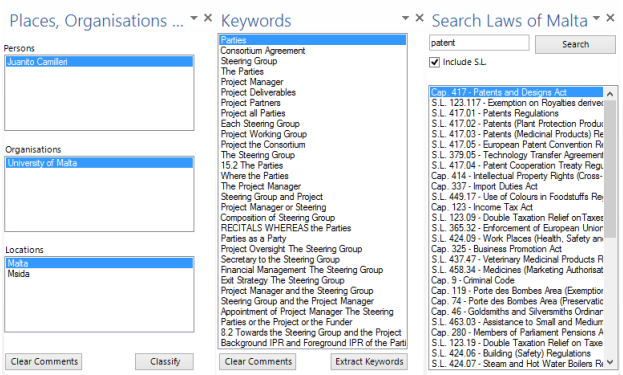


Figure 2: Task pane views of law search, keyword and named entity extraction.

Reasoning about contracts can be done by modelling these using deontic logic (Von Wright, 1999), which views contracts as agreements between two or more parties, with norms (i.e. obligations, permissions, and prohibitions) and structures over these (e.g. temporal sequential composition). To bridge the gap between a natural language contract and such a model we have constructed a deontic logic that we can use to reason about a contract which is only partially known. This is important primarily in case of inaccuracies in the outcomes of the algorithm used to translate English contracts to their formal representation, but is also intended to help during contract-drafting, when the contract is not yet complete.

Definition 1 illustrates a simple subset of the deontic logic we use, without partiality, for simplicity of illustration. Note how the deontic norms are modelled as predicates over an action α, labelled with the acting party p. Simple contract clauses can then either be an obligation (O), a permission (P), or a prohibition (F). These clauses can then also either be sequentially composed ($C \triangleright C'$), reparated[1] ($C \blacktriangleright C'$), concurrently composed ($C \& C'$), or conditioned on actions occurring ($[e]C$).

Definition 1. *A contract C, where α is an action label and p is a party label, is defined as follows:*

$$\begin{aligned} C &:= O_p(\alpha) \mid P_p(\alpha) \mid F_p(\alpha) \mid [e]C \mid \\ &\quad\; C \triangleright C \mid C \blacktriangleright C \mid C \& C \\ e &:= \alpha \mid 0 \mid 1 \mid e.e \mid e + e \mid e \& e \end{aligned}$$

Our approach to translate English contracts into this formal representation uses syntactic parsing. Consider, for example, the sentence "The passenger should check in" (an obligation clause). Structurally, this has the form **S → NP (VP → MD VP')**, as in Figure 3. Note how this structures the sentence such that it separates the party **(NP → the passenger)**, the norm **(MD → should)**, and the action **(VP' →**

[1] A reparation clause C' for a contract C comes into effect if and after C is violated.

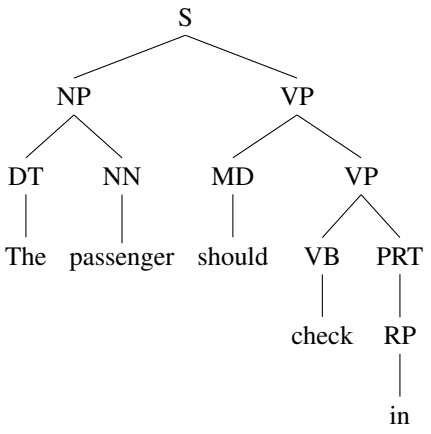


Figure 3: Parse tree of a normative sentence.

check in) into different sub-trees. Our approach is to 'read off' these aspects of the sentence's argument structure from the parse tree, mapping them to elements of our formalism.

To extract these from a sentence we define a number of pattern-matching expressions using Tregex (Levy and Andrew, 2006), which allows us to separately grab exactly the relevant features of a normative sentence. Thus with an appropriate expression we can get the formal counterpart of the clause, i.e. $O_{passenger}(checkIn)$, indicating an obligation on the passenger to check in. We have defined several such expressions that correspond to a certain parse tree structure (along with the presence of a norm specifier like *should*, or *permitted to*). Although it is not clear whether such constructions always correspond to a normative sentence (e.g. "The receptionist should have been here" does not specify an obligation, although it can be seen to imply a perceived one), in the limited context of contracts this is quite likely, given that the phrasing of such constructs typically follows well-worn templates.

This approach is however limited by the number of expressions defined (and their quality). Another issue is that some sentences may not have correlates in our logic e.g. a distinction is made between state-based and action-based clauses, between which there is no one-to-one correlation (Hage, 2001).

With this formal representation we can detect conflicts automatically, through an appropriate trace semantics (e.g. $O_p(a)$ is satisfied if a is done, while $C \rhd C'$ is satisfied if C is satisfied, after which C' applies and is satisfied). We generate an automaton with states labelled with the applicable norms there, and transitions by actions, according to the trace semantics, and using the minimal method delineated in (Fenech et al., 2009). We axiomatise conflicts as in Definition 2, from (Pace and Schapachnik, 2012). By comparing the contracts at the states of the generated automaton we can detect any conflicts and report back to the user.

Definition 2. *Two contracts are said to be in conflict if there is no trace that satisfies both at the same time. The conflict relation is denoted by* ✠*, so that that C and C' are conflicting is denoted by C✠C'. Note also that we denote two mutually exclusive actions as $a \bowtie a'$.*

Contract	TP	TN	FP	FN	Precision	Recall	F_1	$F_{0.5}$
C14	14	33	5	2	0.739	0.875		
C21	9	170	56	0	0.139	1		
C41	16	61	9	0	0.64	1		
C69	12	37	4	0	0.75	1		
C199	5	37	18	10	0.217	0.333		
Results					0.497	0.842	0.625	0.541

Table 1: Formalizing norms evaluation.

Axioms:

$$\vdash P_p(a) ✠ F_p(a) \quad (1)$$

$$\vdash O_p(a) ✠ F_p(a) \quad (2)$$

$$a \bowtie a' \vdash O_p(a) ✠ O_p(a') \quad (3)$$

$$a \bowtie a' \vdash O_p(a) ✠ P_p(a') \quad (4)$$

$$C ✠ C' \vdash C' ✠ C \quad (5)$$

$$C ✠ C' \wedge C' \equiv C'' \vdash C ✠ C'' \quad (6)$$

5. Evaluation

We evaluated our research in two ways: (i) by testing our English to deontic logic translation on a random selection of contracts from the Australian Contract Corpus (Curtotti and McCreath, 2011); and (ii) by obtaining feedback from notaries who used the Microsoft Word add-on over a few days.

As our gold-standard we selected five contracts from the corpus (of varying length), and hand-tagged each clause with a suitable representation in our logic, where possible. Clauses were also tagged as *normative* or not, and as *formalizable* in our logic or not.

The results of the automated translation are shown in Figure 1. In our evaluation, true positives correspond to those clauses which can be formalized and have been formalized correctly; while false positives correspond to the clauses which cannot be formalized but have still been (incorrectly) formalized. True negatives are clauses that cannot be formalized and were not attempted, while false negatives are clauses that can be formalized but were not.

As can be seen the amount of false positives is not negligible, especially with contracts **C21** and **C199**. Through an analysis of their text we observed that these false positives mostly occur in the definitions section of these contracts. These are only tagged as normative (since norm specifiers were present), but with the translation failing. Methods however exist to extract definitions automatically (e.g. (Curtotti et al., 2013)) which can be employed to preprocess the contract and dealing with definitions separately, thus improving the algorithm.

The tool as a whole was given to a number of notaries who were informally asked to give feedback after using it for a few days. Feedback received on the search functionality, especially in legal documents and companies, was mostly positive, or neutral among users who said they rarely consult such already available online databases. The other features were not perceived as useful, although it was

pointed out that they may be more applicable in the context of large contracts.

6. Discussion

Our tool thus effectively combines existing NLP tools and formal contract analysis algorithms, providing for a degree of automated analysis. However, there are other features that we did not consider that would make the tool more attractive to notaries, such as a templating system, easing the analysis of definitions (e.g. (Curtotti et al., 2013)), a versioning system (a work-in-progress), or a higher-level analysis of the components of a contract (e.g. (Gabbard et al., 2015)).

On the formal side our approach also has some limitations. A major one is the fact that we check for equality between actions simply by checking for string equality. A better measure of equality can be added to our algorithm by semantic similarity measures that use a lexical database to analyse the senses of a word, as done in (Aires et al., 2015).

The logic used needs to be augmented with state-based norms as first-class entities, since these too appear in contracts, though seemingly at a lesser incidence then action-based ones. An example of such a norm is "The passenger should be in possession of their passport during the whole trip", which we detect automatically by noting the use of "be"[2].

7. Conclusions

Professionals involved in contract-drafting have the potential to benefit from tools that employ NLP techniques that can automatically analyse the contract while it is being written. This is an area of the humanities where NLP tools have yet to make a noticeable impact.

In this paper, we have presented a tool, packaged as an add-in to Microsoft Word, that presents several legal-drafting support features as task panes. These employ keyword and named entity extraction so as to facilitate the extraction of certain key words associated with each clause, to enable easier browsing of a contract depending on these keys.

We also employ a deontic logic, and syntactic parsing to automatically (partially) translate an English contract into a deontic logic model from which automated deductions can be made. Specifically conflicts between clauses can be detected.

The tool was tested by lawyers and notaries, getting overall positive feedback with suggestions for further work (e.g. including contract templates), with the law and company search being seen as the most useful, and automated deduction as promising.

8. References

João Paul Aires, Vera Lúcia Strube de Lima, and Felipe Meneguzzi. 2015. Identifying potential conflicts between norms in contracts. In *18th International Workshop on Coordination, Organisations, Institutions and Norms (COIN 2015) @IJCAI*, July.

Shaun Azzopardi. 2015. Intelligent contract editing. Master's thesis, Department of Computer Science, University of Malta.

Michael Curtotti and Eric C. McCreath. 2011. A corpus of australian contract language: Description, profiling and analysis. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, ICAIL '11, pages 199–208, New York, NY, USA. ACM.

Michael Curtotti, Eric McCreath, and Srinivas Sridharan. 2013. Software tools for the visualization of definition networks in legal contracts. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ICAIL '13, pages 192–196, New York, NY, USA. ACM.

Mauro Dragoni, Guido Governatori, and Serena Villata. 2015. Automated rules generation from natural language legal texts. In *Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts*, pages 1–6, San Diego, USA, June.

Stephen Fenech, Gordon J. Pace, and Gerardo Schneider. 2009. Automatic conflict detection on contracts. In *Proceedings of the 6th International Colloquium on Theoretical Aspects of Computing*, ICTAC 2009, August.

Jason Gabbard, Jana Z. Sukkarieh, and Federico Silva. 2015. Writing and reviewing contracts: Don't you wish to save time, effort, and money? In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, pages 229–230, New York, NY, USA. ACM.

Xibin Gao and Munindar P. Singh. 2014. Extracting normative relationships from business contracts. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS 2014, May.

Pablo Gervás. 2013. Story generator algorithms. In P. Hühn, editor, *The Living Handbook of Narratology*. Hamburg: Hamburg University.

Jaap Hage. 2001. *Contrary to Duty Obligations - A Study in Legal Ontology*. IOS Press.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*, LREC 2006.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, November.

Gordon J. Pace and Fernando Schapachnik. 2012. Contracts for interacting two-party systems. In Anders P. Ravn Gordon J. Pace, editor, *Proceedings of Sixth Workshop on Formal Languages and Analysis of Contract-Oriented Software*, volume 94 of *EPTCS*, pages 21–30.

Georg Henrik Von Wright. 1999. Deontic logic: A personal view. *Ratio Juris*, 12(1):26–38, March.

Adam Wyner and Wim Peters. 2011. On rule extraction from regulations. In Katie Atkinson, editor, *JURIX*, volume 235 of *Frontiers in Artificial Intelligence and Applications*, pages 113–122. IOS Press.

[2]There does not exist a single action a such that for this example we can construct a norm $O_p(a)$, i.e. a norm which is satisfied by the performance of a single action.

Organiziranje projekta vizualizacije podatkov

Narvika Bovcon,[*] Jure Demšar,[*] Aleš Vaupotič[†]

* Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, SI-1000 Ljubljana
narvika.bovcon@fri.uni-lj.si
* Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, SI-1000 Ljubljana
jure.demsar@fri.uni-lj.si
†Raziskovalni center za humanistiko, Univerza v Novi Gorici
ales.vaupotic@ung.si

Povzetek

Članek bo predstavil vizualizacije zbirke literarnovednih podatkov, projekt, ki je potekal v okviru študijskega procesa na Fakulteti za računalništvo in informatiko Univerze v Ljubljani, v povezavi sodelavcev iz treh disciplin: humanistike, računalništva in informatike ter grafičnega oblikovanja.

1 Uvod

Področje digitalne humanistike je izoblikovalo specifične žanre in metode, s katerimi se loteva urejanja in raziskovanja digitaliziranih kulturnih zbirk podatkov. Osrednji poudarki in raziskovalni pristopi so opisani v petnajstih poglavjih zbornika Digital_Humanities (Burdick et al., 2012), kot teoretski okvir področja jih poskusimo strniti v naslednji seznam: (1) kuriranje večpredstavnostnih zbirk, (2) označevanje besedil, (3) oddaljeno branje in strojno učenje, (4) različni pogledi na iste podatke, (5) kulturna analitika in podatkovno rudarjenje, (6) vizualizacija podatkov in oblikovanje informacij, (7) plastenje podatkov na geolokacijah, (8) digitalne skupnosti, (9) porazdeljeno proizvajanje znanja in množičenje, (10) resne igre, (11) refleksija programja, (12) narativizacija podatkovnih zbirk, (13) kultura remixa in ponovne uporabe, (14) vseprisotna infrastruktura, (15) javna dostopnost vsebin. Gre za dejavnosti in problemske sklope, ki se med seboj povezujejo in nadgrajujejo, zato jih lahko predstavimo tudi v bolj skrčenem naboru temeljnih pojmov. Teorija novomedijskega objekta Leva Manovicha (2001) poveže podatkovno zbirko in množico vmesnikov, prek katerih dostopamo do podatkov v zbirki. Ta koncept združuje točke 1, 2, 3, 4, 5 in 12. Naslednja pomembna enota se nanaša na samo zasnovo in jo zato lahko prepoznamo kot osnovni element digitalne humanistike, to so digitalni modeli kulturnih artefaktov, "oblike argumenta, izražene v informacijskih strukturah – vmesnikih, podatkovnih zbirkah, orodjih, platformah" (Burdick et al., 2012), ta koncept poveže novomedijski objekt s točkama 11 in 13, z refleksijo lastnega obstoja in učinkovanja ter načini produkcije in uporabe. Kot poseben poudarek raziskovanja digitalne humanistike ostaja točka 6, vizualizacije, ki prenašajo podatke v vidno razumljivo obliko prek likovnega jezika. Točke 7, 14 in 15 se povežejo v enoto, ki omogoča obstoj in delovanje v mešani resničnosti, kjer je poudarek lahko tudi na telesnosti. Točki 8 in 10 se nanašata na pravila, ki so ključna tako za resne igre kot za delovanje digitalnih skupnosti; s tem se ukvarja teorija kiberteksta (Aarseth, 1997). Točka 9 izpostavi družbo in družbeno raznolikost kot enega ključnih akterjev, udeleženih v pristopih digitalne humanistike.

2 Cilji

V projektu vizualizacije literarnovedne podatkovne zbirke smo uporabljali predvsem metode, povezane z novomedijskim objektom, torej v relaciji med podatkovno zbirko in njenimi vmesniki, ter metode vizualizacije podatkov in oblikovanja informacij (Manovich, 2011). Sama izgradnja podatkovne zbirke ni bila v domeni našega projekta, saj smo izdelovali zgolj vmesnike za obstoječo podatkovno zbirko Women Writers (http://neww.huygens.knaw.nl), ki je nastala v okviru evropskega projekta Women Writers In History (COST Action, 2009–2013). Z našim pristopom, ki sodi med generativne humanistične metode (Burdick et al., 2012), in množico konkretnih partikularnih rešitev smo dejavno prispevali k oblikovanju digitalnega modela kulturnega artefakta podatkovne zbirke Women Writers: sama podatkovna zbirka namreč še ni tematizirala možnosti vizualizacij kot svojega integralnega dela, ki skupaj z zbirko tvori argument, izražen v informacijski strukturi. Pravzaprav vsaka posamezna prototipna vizualizacija podatkovne zbirke predstavlja specifično obliko tega digitalnega kulturnega artefakta. Uredniki podatkovne zbirke so izdelali drugačne oblike argumentov, s tem ko so zbirko predstavili s spletno stranjo, ki ob iskanju izpiše seznam zadetkov, ter jo povezali s posebej zanjo načrtovanim in izdelanim virtualnim raziskovalnim okoljem, ki bo v okviru projekta Travelling Texts 1790–1914 (HERA CRP, 2013–2016) dokončano letos (http://resources.huygens.knaw.nl/womenwriters).

Drugačen okvir za pregledovanje in sistematizacijo področja digitalne humanistike ponuja npr. pregled vsakoletnih nagrajenih projektov na spletišču Digital Humanities Awards (http://dhawards.org). Primerjava s podobnimi projekti omogoča natančnejše obravnavanje posebnosti, povezanih z vizualizacijami podatkovne zbirke Women Writers.

Poleg teoretskega ozadja in primerjave z referenčnimi sorodnimi projekti bo članek predstavil tudi samo organizacijo tovrstnega interdisciplinarnega ter v pedagoški proces integriranega projekta ter s tem povezane omejitve in priložnosti. V prvem podpoglavju predstavimo metodo, ki je v svojem temelju povezana z organizacijo projekta.

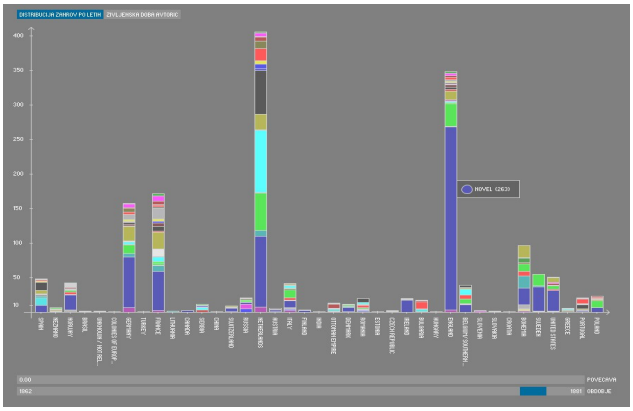
2.1 Metode in organizacija projekta

Projekt je potekal v treh organizacijskih oblikah, ki so vsaka po svoje strukturirale odnose med sodelavci z različnih raziskovalnih področij. Prva organizacijska oblika je študente računalništva in informatike soočila z literarnovedno podatkovno zbirko in mentorico, grafično oblikovalko. Druga organizacijska oblika je vpeljala literarnovedne raziskovalke, ki so soustvarjale zbirko Women Writers, da so zastavile znanstvena vprašanja, na katera so bili odgovori pridobljeni skozi vizualizacije. Tretja oblika je namesto profesionalnih raziskovalcev vključila študente humanistike in s tem razširila pedagoški proces še na drugo disciplino. V vseh organizacijskih oblikah so vizualizacije izdelali študenti Fakultete za računalništvo in informatiko Univerze v Ljubljani pod mentorstvom izr. prof. dr. Narvike Bovcon in asistentov Tadeja Zupančiča in Jureta Demšarja. Koordinacija projekta s sodelavkami zbirke Women Writers je potekala prek doc. dr. Aleša Vaupotiča, ki je bil tudi mentor študentom humanistike na Univerzi v Novi Gorici.

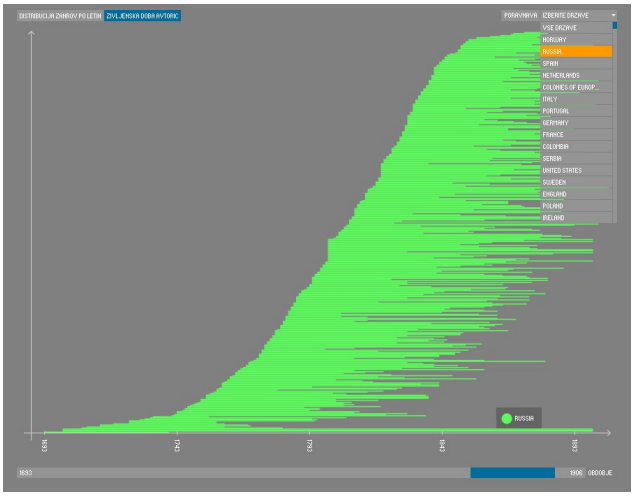
2.1.1 Množica vmesnikov do podatkovne zbirke

Metoda generativnih humanističnih raziskav izdeluje argumente s pomočjo prototipov, ki jih postopoma izboljšuje. Gre za pristop, ki je značilen za oblikovalske projekte, kjer je poudarek na izdelavi konkretnih rešitev, množice prototipov, ter njihovo ovrednotenje in nadgrajevanje (Kräutli, 2016). Prednost umestitve izdelovanja prototipov v študijski proces je ta, da razred stotih študentov, ki izdela v enem mesecu nalogo, izdela dejansko sto povsem različnih prototipov. Takemu obsegu produkcije je v okvirih industrije oz. komercialno zasnovanega projekta zelo težko slediti, saj se finančno in terminsko ne izide, z bistveno večjo težavo tudi zagotovi toliko človeških virov. Pri tem želimo poudariti, da se je v našem pristopu individualnost študentov jasno odražala v veliki raznolikosti njihovih izdelkov, medtem ko izdelki istega avtorja bistveno težje ali sploh ne morejo doseči tolikšne raznolikosti izjav. Izdelava naloge v več zaporednih študijskih semestrih (naš projekt je potekal od leta 2012 do leta 2016) prinese na stotine rešitev, med katerimi je mogoče izbrati najboljše: tiste, ki so kvalitetne na ravni vsebine, torej smiselne in zanimive glede samega izbora prikazanih odnosov med informacijami iz podatkovne zbirke, in na ravni vizualne predstavitve teh odnosov in informacij, torej grafičnega oblikovanja vmesnika. Za študente je tak projekt zanimiv (kljub težavam in predsodkom, pogosto povezanim z interdiciplinarnim sodelovanjem), ker se vključujejo v raziskovalno prakso.

Raznolikost prototipnih rešitev izvira v veliki meri tudi iz našega načrtno fragmentarnega pristopa k vizualizaciji. Namen ni bil, vizualizirati celotno zbirko oz. izbrati nekaj ustreznih grafov za prikaz vseh povezav med deli zbirke. Temu smo se odrekli zato, ker je zbirka preveč obsežna in bi že zaradi tega tovrstna naloga ne bila več izvedljiva v okvirih študentskega seminarskega projekta. Drugi, pravzaprav poglavitni razlog je bil počasno delovanje prikaza, ko se vizualizirajo vsi podatki iz zbirke, kar naredi vizualizacijo praktično nefunkcionalno za razbiranje prikazanih podatkov in odnosov med njimi (tovrsten prikaz smo poskusili izdelati v obliki različnih grafov in z različnimi tehnologijami, vendar je vsakokrat deloval prepočasi in bil zato neuporaben). Tretji razlog za izbiro parcialnih pogledov na zbirko je bila večja fleksibilnost in inovativnost pri formuliranju posameznih smiselnih znanstvenih vprašanj v zvezi s podatki iz zbirke (osnova za to so bile kategorije podatkov), saj so bili študenti motivirani, da so s pomočjo vizualizacije poiskali odgovor na konkretno vprašanje, ki se jim je porodilo ob pregledovanju in razbiranju ustroja zbirke, brez vizualizacije pa odgovora nanj ni ali pa ni očiten (Slika 1). Na ta način so bili tematizirani različni vstopi v zbirko, ki so temeljili na konkretnih vprašanjih v zvezi s podatki, torej so zbirko skozi vizualizacijo narativizirali.



Slika 1: Dva pogleda na zbirko, izbira prek gumbov v zgornji vrsti – popularnost žanrov in življenjska doba avtoric. (Študenta: Aleksandar Kojić, Dejan Grbec.)



Slika 2: Vizualizacija pokaže napako pri vnosu življenjske dobe avtorice.

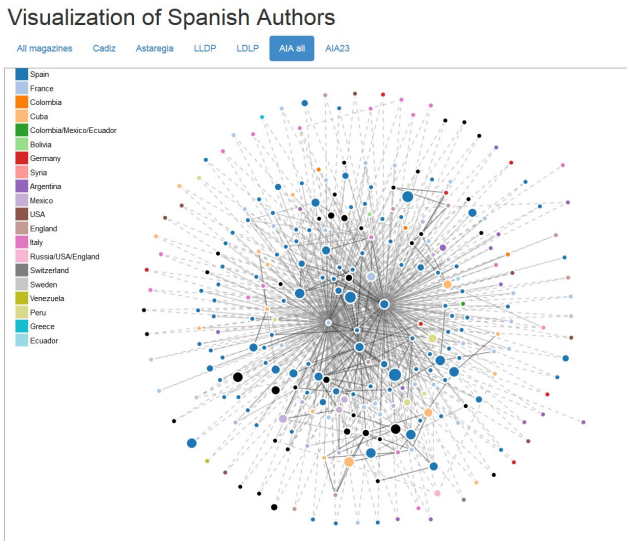
Tovrstni fragmentarni, skicozni, partikularni pristop je najustreznejši način za izdelavo eksperimentalnih vizualizacij, ki služijo mdr. namenu promocije zbirke, javnemu predstavljanju projekta za različna občinstva ter tudi odkrivanju napak in iregularnosti v zvezi z vnosi v zbirko (Slika 2). Zbirka Women Writers nujno potrebuje uspešno predstavitev in promocijo, saj gre za raziskave in predstavljanje podatkov o do sedaj zapostavljenem in skorajda nevidnem ustvarjalnem delu žensk vse do vključno t. i. dolgega 19. stoletja, ki se danes mora družbeno afirmirati in uvrstiti v šolske učne programe.

Posamezni prototipi vmesnikov so bili izdelani skozi stopnje korektur v zvezi z grafičnim oblikovanjem in strukturiranim posredovanjem informacij, ki so potekale pod mentorstvom oblikovalke, zato so nekateri med njimi – skozi postopno odstranjevanje napak in v kombinaciji s softverskimi predlogami za različne diagramske prikaze iz prosto dostopnih knjižnic, ki zagotavljajo dokaj visoko stopnjo grafične urejenosti – dosegli raven ustrezne berljivosti, smiselne razporeditve informacij po ekranski ploskvi ter za uporabnika intuitivne interakcije. Študenti računalništva in informatike so na ta način, tj. skozi interdisciplinarno sodelovanje, oblikovali čitljive in likovno urejene vmesnike za vizualizacije odnosov med podatki. Če razmislimo, ali bi bilo bolje, da bi vmesnike oblikovali študenti oblikovanja vizualnih komunikacij ali pa celo profesionalni oblikovalci, hitro ugotovimo, da obstaja ovira, ki je tehnološke narave: oblikovalci bi se morali naučiti nekaj programiranja, kar v Sloveniji danes še ni prav pogosta praksa, čeprav je bil npr. Processing (http://processing.org) razvit za oblikovalce vizualnih komunikacij v novih medijih in v tujini obstajajo številni interdisciplinarni novomedijski študiji, ki temeljijo na oblikovanju. V Sloveniji je organizacija te vrste interdisciplinarnega projekta danes lažje izvedljiva v okviru študija računalništva in informatike.

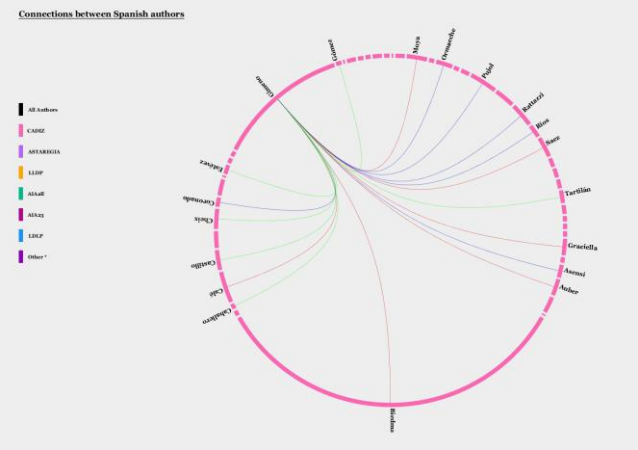
Pregled najboljših vizualizacij, ki so bile izdelane v letih 2012–2013 – te so bile dveh vrst: interaktivne, izdelane s pomočjo platforme Processing, za katere je bilo značilno približevanje in oddaljevanje pogleda na podatke prek animacij, interakcije in postopnih prikazov podrobnosti, ter na drugi strani videi in gibljive informacijske grafike, izdelane s pomočjo programa Adobe After Effects, ki so s svojo linearno obliko izrazito narativizirale podatkovno zbirko – so podrobneje opisane in dokumentirane z ekranskimi slikami v članku Narvike Bovcon (2014). Vizualizacije so jasno pokazale, da je vsakršna statistična, tj. količinska primerjava med podatki v bazi, povezana predvsem z vnosi v bazo in ne odslikuje v celoti dejavnosti avtoric, npr. količine izdanih del ali prejetih recepcij po Evropi, saj različne države in jezikovni teritoriji bolj ali manj celovito vnašajo gradivo v bazo: Holandija ima daleč največ vnosov, saj je baza nizozemski projekt. Torej pri poskusu oddaljenega branja te podatkovne zbirke ne gre za dilemo o redukciji kompleksnosti humanističnih podatkov na nekaj ali eno njihovo lastnost, prek katere se jih primerja (Drucker, 2011), ampak za strukturno razliko nezaključene baze, kakršna je Women Writers, ki je projekt v nastajanju (vizualizacije se povežejo z bazo in odražajo trenutno stanje zbirke), v primerjavi z zaključeno bazo, kakršna je neko dramsko besedilo, ki je napisano in vsebuje določeno število junakov, vizualizacija pa lahko riše razmerja med njimi, soprisotnost junakov v dramskih prizorih ali pa količino izgovorjenega teksta (Grandjean, 2015). Zaradi velikih teritorialnih in časovnih razdalj ter neenakomerne izpolnjenosti zbirke izrisovanje mrež povezav v zbirki Women Writers zahteva omejitve pri prikazih (npr. omejitev na eno državo), če želimo dobiti dokaj primerljive nabore podatkov, kot npr. v projektu, ki izrisuje mreže povezav med člani Belfastske skupine (http://belfastgroup.digitalscholarship.emory.edu/network), kjer gre za bistveno manjši in bolj povezan nabor možnih podatkov, čeprav še niso vsi zavedeni in upoštevani. V tej smeri so šle nadaljnje raziskave, v katerih so bila znanstvena vprašanja izrazito fokusirana ter podatki posebej urejeni s strani strokovnjakinj.

2.1.2 Nove raziskave – vizualizacije strokovno oblikovanih znanstvenih vprašanj v zvezi z izbranimi podatki

Druga stopnja organizacije projekta je testirala vključitev profesionalnih humanistov, torej profesorjev in raziskovalcev podatkov, vključenih v gradnjo zbirke, pri zastavitvi znanstvenega vprašanja (to je bilo v prvi stopnji projekta prepuščeno samim študentom računalništva in informatike in je bilo v več primerih zastavljeno premalo kompleksno ali pa sploh narobe – povezovalo je podatke, ki v resničnosti niso bili soodvisni). Na ta način smo zagotovili najvišjo stopnjo zahtevnosti in doslednosti v zvezi z znanstvenim vprašanjem, naloga študentov pa je bila, to vprašanje čim bolje vizualizirati: fokus naloge je bil prestavljen z načrtovanja vsebine na načrtovanje oblike vizualizacije. Poseben, skrbno pripravljen sklop podatkov, povezanih z njunima raziskovalnima znanstvenima vprašanjema, sta posredovali dr. Suzan van Dijk s Huygens ING, o povezavah med nizozemskimi avtoricami, in Judith Rideout, doktorska študentka z Univerze v Glasgowu, o moči povezav med španskimi avtoricami, ki so objavljale pri različnih literarnih revijah; avtorice so se bodisi poznale enostransko, vzajemno, ali pa so delale na skupnih projektih, kar pomeni najmočnejšo povezavo. Izdelana sta bila dva, povsem različna prikaza moči povezav med španskimi avtoricami.



Slika 3: Tri debeline črt predstavljajo tri stopnje moči povezav med španskimi avtoricami (avtorice so predstavljene s točkami na mreži), ki so pisale za konkretno revijo (izbrana v zgornjem meniju). Barve označujejo države. Ob kliku na posamezno točko na grafu se izpišejo podrobnejši podatki o izbrani avtorici. Mreža je interaktivna in jo je mogoče premikati, razplesti, povezave delujejo kot elastike – interakcija omogoča postopno razbiranje povezav, ki zaradi številnosti niso razvidne prek pogleda na celoten graf. (Študentki: Mojca Komavec, Viki Petrovič.)



Slika 4: Moč povezave med avtoricami, ki so razporejene po krogu, je označena s tremi različnimi barvami črt. Barva krožnice predstavja izbrano revijo (izbranih je lahko več revij hkrati) in se povezuje z legendo na levi. (Študentka: Marija Djurdjević.)

Posebej velja omeniti primer t. i. resne igre, ki recepcije med nizozemskimi avtoricami postavi v interaktivno virtualno resničnost, narejeno s pomočjo pogona Unity, v kateri se uporabnik sprehaja po tipični nizozemski pokrajini med hišami posameznih avtoric, vzdolž poti pa bere citat iz recepcije, ki jo je napisala avtorica, katere hišo zapušča, o avtorici, proti hiši katere hodi. V tem primeru je pomembno izumljanje vmesnika, ki omogoča uprostorjeno branje, s tem ko poveže računalniško generirani 3D prostor z zbirko citatov in mrežo povezav (Grosar in Demšar in Bovcon, 2015).

V tretjem modelu sodelovanja je bila naloga študentov humanistike, postaviti znanstveno vprašanje v zvezi s podatki iz zbirke in ga formulirati v obliki jasnega navodila, na podlagi katerega študenti računalništva in informatike izdelajo vizualizacijo. Prototipe rešitev so študenti humanistike v zaključnem eseju tudi ovrednotili, razčlenili, kako in ali so podatki prikazani na razumljiv način, kako bi bili posamezni deli lahko prikazani bolj nazorno in razumljivo, kakšno vlogo imajo sredstva grafičnega oblikovanja pri podajanju informacij. S tem so aktivno vstopili v študij digitalne humanistike.

Slika 5: Navodilo na dnu ekrana usmerja uporabnika k hiši naslednje avtorice, omenjene v citatu, ki ga bere na zgornjem delu ekrana. (Študent: Jernej Grosar.)

3 Sklep

Podatkovna zbirka Women Writers se je izkazala kot ustrezna, dovolj kompleksna zbirka podatkov, ob kateri se študenti naučijo kritično razmišljati tako o različnih smiselnih povezavah med kategorijami podatkov kot tudi o načinih grafičnega prikaza teh odnosov na interaktivnih vmesnikih. V prihajajočem mesecu bomo izdelali predstavitveno spletno stran, na kateri bodo dostopne izbrane eksperimentalne vizualizacije te zbirke.

Literatura

Espren Aarseth. 1997. *Cybertext: Perspectives on Ergodic Literature*. Baltimore: The Johns Hopkins Univ. Press.

Narvika Bovcon. 2014. Jezik gibljivih slik v računalniških vizualizacijah literarnozgodovinske podatkovne zbirke. *Primerjalna književnost,* 37(2):119–133, 235–242.

Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner in Jeffrey Schnapp. 2012. *Digital_Humanities*. Cambridge, Mass.: MIT Press.

DigitalHumanitiesAwards. 2012–2015. http://dhawards.org.

Johanna Drucker. 2011. Humanities Approaches to Graphical Display. *Digital Humanities Quarterly,* 5(1). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

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

Belfast Group Poetry: Networks. Emory University's Manuscript, Archives, and Rare Book Library. http://belfastgroup.digitalscholarship.emory.edu/network.

Martin Grandjean. 2015. *Network visualization: mapping Shakespeare's tragedies*. http://www.martingrandjean.ch/network-visualization-shakespeare.

Jernej Grosar, Jure Demšar, Narvika Bovcon. 2015. 3D walk through the refrences of Dutch women writers. V: *StuCoSReC: proceedings of the 2015 2nd Student Computer Science Research Conference*. Koper: University of Primorska: 39-42, ilustr.

Florian Kräutli. 2016. *Visualising Cultural Data: Exploring Digital Collections Through Timeline Visualizations*. Ph.D. thesis, Royal College of Art.

Lev Manovich. 2001. *The Language of New Media*. Cambridge, Mass.: MIT Press.

Lev Manovich. 2011. What is visualization? *Visual Studies,* 26(1):36–49.

Processing. http://processing.org.

Aleš Vaupotič. 2007. Literarno-estetski doživljaj in novi mediji - prihodnost literature? *Primerjalna književnost* 30(1):203–216.

WomenWriters. 2012–2016. Ur. Suzan van Dijk. http://neww.huygens.knaw.nl in http://resources.huygens.knaw.nl/womenwriters.

Razvoj učne množice za izboljšano označevanje spletnih besedil

Jaka Čibej,* Špela Arhar Holdt,* Tomaž Erjavec,† Darja Fišer*†

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
jaka.cibej@ff.uni-lj.si, spela.arharholdt@ff.uni-lj.si, darja.fiser@ff.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

Jezik spletne komunikacije se v marsikaterem vidiku razlikuje od standardnega jezika. Obstoječa orodja za označevanje besedil se z njim težje spopadajo, saj je učenje označevalnih postopkov do zdaj potekalo predvsem na standardnih besedilih. Za čimbolj natančno računalniško obdelavo računalniško posredovane komunikacije je torej treba ustrezno nadgraditi označevalne metodologije in orodja. V prispevku zato predstavljamo izdelavo učnega korpusa slovenske spletne komunikacije, ki bo uporabljen kot učna množica za izboljšano jezikoslovno označevanje slovenskih spletnih besedil. Korpus je bil vzorčen iz korpusa spletne slovenščine JANES in sprva avtomatsko označen, nato pa ročno popravljen na petih ravneh: tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladnja.

The Development of a Training Set for an Improved Annotation of Internet Texts

In many ways, the language of internet communication differs from standard language. Internet texts are more difficult to process for existing tools, which have predominantly been trained on standard language. To improve their accuracy when processing internet texts, the tools and annotation methodologies need to be upgraded. In this paper, we present the compilation of a training corpus of Slovene internet communication to be used as a training set to improve the automatic annotation of Slovene internet texts. The training corpus was sampled from the JANES corpus of Internet Slovene, then automatically annotated and manually corrected on five annotation levels: tokenisation, sentence segmentation, normalisation, lemmatisation, and morphosyntax.

1 Uvod

Jezik spletnih žanrov, kot so tviti, forumi in komentarji, se v marsikaterem vidiku razlikuje od standardnega jezika (Baldwin et al., 2013). Med pogosteje omenjanimi razlikami so npr. pogovorni zapisi besed, pogostejša raba regionalizmov in tujejezičnih prvin ter raba okrajšav in jezikovnih oz. grafičnih elementov, specifičnih za spletno komunikacijo (Crystal, 2001). Z naštetimi specifikami se obstoječa orodja za označevanje besedil težje spopadajo, saj je učenje označevalnih postopkov do sedaj potekalo predvsem na standardnih besedilih (Ljubešić et al., 2014). Če želimo kvalitetno obdelavo jezika zagotoviti tudi za računalniško posredovano komunikacijo, je torej potrebna ustrezna prilagoditev oz. nadgradnja označevalne metodologije in orodij.

V prispevku kot enega od korakov v tej smeri predstavimo izdelavo učnega (in testnega) korpusa spletne komunikacije, ki je bil vzorčen iz korpusa JANES (Fišer et al., 2015). Učni korpus je bil nato avtomatsko označen na petih ravneh (tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladnja), avtomatsko pripisane oznake pa so bile ročno popravljene. V prvem delu prispevka predstavimo cilje in delotok označevalnega procesa, v drugem pa se osredotočimo na tisti del apliciranih rešitev, ki se razlikuje od dosedanjih praks jezikoslovnega označevanja slovenščine.

2 Sorodne raziskave

Predstavljena raziskava sodi v obširno in dejavno področje normalizacije nestandardnih prvin v besedilih. Čeprav nekatere raziskovalce zanimajo specializirani problemi, npr. normalizacija zapisa velikih začetnic (Nebhi et al., 2015), rediakritizacija besed (Ljubešić et al., 2016) ali normalizacija ključnikov (Declerck in Lendvai, 2015), se večina skupnosti ukvarja s splošno normalizacijo nestandardnega besedišča v spletnih besedilih, ki izboljšuje nadaljnjo obdelavo besedil, npr. oblikoskladenjsko označevanje in lematizacija. Ker gre za razmeroma novo raziskovalno področje, za večino jezikov normalizirane učne in testne množice še ne obstajajo. Prav tako ni vzpostavljenih enotnih evalvacijskih meril, zato je orodja težko razvijati, evalvirati in primerjati. Prva prizadevanja v to smer so že obrodila sadove za angleščino v okviru delavnice in skupne naloge WNUT (Baldwin et al., 2015), za nemščino v okviru skupne naloge EmpiriST (Beißwenger et al., 2015a) in za španščino v okviru delavnice in skupne naloge Tweet-Norm (Alegria et al., 2014). Poleg prosto dostopnih učnih množic so bile za te jezike izdelane tudi smernice za ročno označevanje (npr. Beißwenger et al., 2015b). Za podobno si prizadevamo tudi s označevalsko kampanjo za slovenščino, ki jo predstavljamo v pričujočem prispevku.

Cilj kampanje je vzpostaviti načela in razviti učne množice za učenje avtomatske normalizacije nestandardnega besedišča v šumnih spletnih besedilih kot predhodni postopek jezikoslovnega označevanja teh besedil z orodji, sicer razvitimi za standardni jezik (Sproat et al., 2001). S tem bomo nadgradili in izboljšali osnovni pristop k normalizaciji slovenskih tvitov (Ljubešić et al. 2014), ki temelji na statističnem strojnem prevajanju na nivoju znakov, naučenem na ročno preverjenem leksikonu izvornih in normaliziranih parov 1000 neznanih najbolj ključnih besed v korpusu tvitov glede na referenčni korpus Gigafida. Čeprav so bili rezultati, doseženi z osnovnim modelom, spodbudni (69% točnost), je njegova pomanjkljivost ta, da pri iskanju najverjetnejše normalizirane oblike ne upošteva sobesedila, kar bomo

omogočili z ročno normaliziranim korpusom, ki ga predstavljamo v prispevku.

3 Priprava podatkov in označevalna platforma

Besedila za označevanje smo vzorčili iz korpusa JANES ter izdelali dva vzorca: Kons1, ki vsebuje tvite, in Kons2, ki vsebuje forumska sporočila ter komentarje na blogovske zapise in spletne novice. V nadaljevanju predstavimo kriterije za vzorčenje teh podkorpusov, označevalno platformo WebAnno in postopek pretvorbe, uvoza in izvoza podatkov.

3.1 Vzorčenje

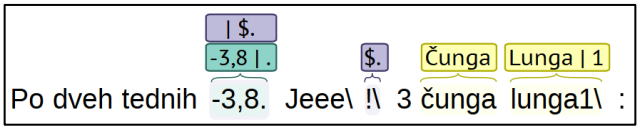
Vzorec Kons1 zajema 4.000 tvitov, ki so bili vzorčeni naključno, a z upoštevanjem nekaterih dodatnih omejitev. Izločili smo tvite, daljše od 120 znakov[1], in tvite z uradnih računov organizacij (npr. agencij in podjetij). Pri vzorčenju smo upoštevali tudi stopnjo tehnične (T1-T3) in jezikovne (L1-L3) standardnosti tvita, ki smo jo merili s posebej za to razvito avtomatsko metodo (Ljubešić et al., 2015). Ker se nismo želeli osredotočiti le na nestandardne tvite (3), temveč so nas zanimale tudi splošne specifike žanra, smo v Kons1 vključili po 1.000 tvitov iz vsake od kategorij T1L1, T3L1, T1L3 in T3L3.[2]

Na podoben način smo izdelali vzorec Kons2, ki vsebuje 4.000 besedil, razdeljenih po stopnjah (ne)standardnosti. Pri žanrih, vključenih v Kons2, ni znakovne omejitve, kakršno ima Twitter, zato smo zaradi primerljivosti z vzorcem Kons1 v Kons2 vključili samo besedila z dolžino med 20 in 280 znaki.

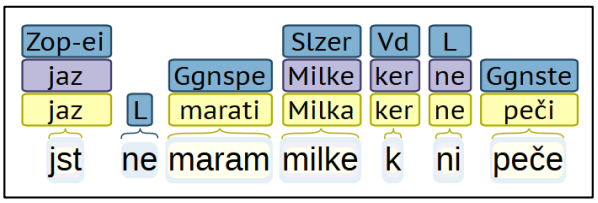
Pred ročnim označevanjem sta bila vzorca z obstoječimi orodji (Erjavec, 2011; Ljubešić et al., 2014) avtomatsko označena na vseh petih obravnavanih ravneh jezikoslovnih oznak.

3.2 Označevalna platforma

Vzorčena besedila smo razdelili na datoteke, ki so vsebovale po 10 besedil, in naložili na WebAnno (Eckart de Castilho et al., 2014), prosto dostopno spletno platformo, ki omogoča večravninsko označevanje besedila. Primeri oznak za normalizacijo, tokenizacijo in stavčno segmentacijo so prikazani na Sliki 1, primeri oblikoskladenjskih oznak in lem pa na Sliki 2.



Slika 1: Oznake za normalizacijo (rumena), tokenizacijo (zelena) in stavčno segmentacijo (vijolična) v WebAnnu.



Slika 2: Oznake za oblikoskladnjo (modra) in lematizacijo (rumena) v WebAnnu. Prikazane so tudi normalizirane oblike (vijolična).

WebAnno smo prilagodili tako, da je omogočal označevanje besedil na vseh petih ravneh, relevantnih za našo učno množico. V platformo je vgrajena tudi funkcija razsojanja, ki se jo lahko uporabi, če iste podatke označi več označevalcev. Pri tem razsodnik primerja večkratne oznake iste datoteke in izbere dokončno različico.

3.3 Pretvorba podatkov

Posebno pozornost smo namenili formatu podatkov, da bi vse ročno preverjene oznake združili v enovit zapis. Pri zapisu korpusa JANES uporabljamo priporočila za kodiranje besedil TEI (Text Encoding Initiative), ki so v uporabi pri večini slovenskih korpusov. Ker WebAnno formata TEI ne podpira, smo med razpoložljivimi formati izbrali TSV, tabelarični format, v katerem je vsaka pojavnica zapisana v svoji vrstici, ki ji je pripisan njen identifikator ter vse oznake.

Izdelali smo program, ki izvorni TEI izvozi v format TSV. Tega je nato mogoče uvoziti v WebAnno, po označevanju pa lahko popravljeni TSV znova izvozimo in združimo z izvornim TEI, tako da rezultat vsebuje vse oznake izvornega TEI, a dopolnjene s popravki ročnega označevanja. Postopek je razmeroma zapleten, saj smo pretvorbo zasnovali tako, da bo uporabna tudi za morebitne nadaljnje označevalske kampanje z bistveno drugačnimi oznakami. Težava je tudi v tem, da naša označevalska metodologija predvideva prvine, za katere WebAnno ni predviden, predvsem popravljanje pojavnic in mej med njimi. Poleg tega je z vidika pretvorbe podatkov problematično, da lahko eni pojavnici ustreza več normaliziranih oblik ali obratno.

4 Smernice za označevanje

Na podlagi preliminarnega ročnega pregleda uravnotežnega vzorca 200 tvitov sta bili izdelani dve zbirki smernic za označevanje[3]. Tehnične smernice so označevalce seznanile z označevalsko shemo v WebAnnu in s splošnimi napotki za delo s platformo (npr. (ra)združevanje pojavnic, brisanje nerelevantnih ali avtomatsko generiranih besedil, delo z večplastnimi oznakami), jezikoslovne smernice pa so obravnavale kriterije za sprejemanje jezikoslovnih odločitev pri označevanju. Za zagotavljanje kompatibilnosti podatkovnih množic so smernice v največji možni meri upoštevale navodila za označevanje slovenskih korpusov (JOS,[4] ssj500k,[5] GOS,[6] IMP[7]) in referenčnih virov za slovenščino (Fran,[8] Sloleks[9]).

[1] Daljši tviti se zaradi splošne omejitve do 140 znakov pogosto končajo z odrezanimi besedami, ki bi predstavljale šum.

[2] Kategorij s stopnjama T2 in L2, pri katerih so značilnosti nestandardnih tvitov manj izražene, nismo vključili.

[3] Smernice so prosto dostopne na naslovu http://nl.ijs.si/janes/viri/

[4] http://nl.ijs.si/jos/msd/html-sl/

[5] http://www.slovenscina.eu/tehnologije/ucni-korpus

V nadaljevanju predstavimo poglavitne specifike označevanja prvin, ki so značilne za spletno komunikacijo, ter pri vsaki ravni označevanja razpravljamo o teţa vnih primerih in rešitvah zanje.

4.1 Stavčna segmentacija

V standardnem jeziku meje med povedmi najpogosteje zaznamujejo končna ločila, v spletnih besedilih, vključenih v našo učno mnoţi co, pa smo na ravni stavčne segmentacije kot signal za konec povedi poleg klasičnih končnih ločil (pika, klicaj, vprašaj) upoštevali tudi druga ločila, ki lahko delujejo kot končna (npr. večpičje, vezaj in narekovaj), oz. druge prvine, ki se lahko pojavljajo na tem mestu:

a) emotikoni in emojiji (=D ☺),
b) ključniki (#sampovem),
c) URL- ali e-naslovi (http://youtube.com, avtor@domena.com),
d) sklici na uporabniška imena (@avtor).

Te prvine zaključujejo stavke zlasti v besedilih brez končnih ločil. Če se stavek konča z nizom prvin, se za konec stavka[10] šteje zadnja prvina v nizu:

> Liverpool zasluţ eno owna Twitter, ampak na vrhu je pa fucking Iago Aspas hahaha :) #nogomet #LFC #SOULIV http://t.co/LCyEvyoVD7 ¶

V primerih, ko se tovrstne prvine sopojavljajo s končnimi ločili (bodisi samostojno ali kot niz), jih obravnavamo kot nov stavek, četudi se nanašajo na prejšnjega:

> Ţivljenje Je Cirkus. js sm pa čefur. Luka Stigl js sm se poscal v hlače k sm se vidu. bolano. ¶ :) ... ¶ http://t.co/QyzKRZqZnS ¶

4.2 Tokenizacija

Specifične označevalne izzive pri spletnih besedilih predstavlja tudi raven tokenizacije. Določene vrste pojavnic, ki vsebujejo ločila, je tokenizator najpogosteje napačno ločil, zato jih je bilo treba zdruţiti ročno. Pogost primer so okrajšave (npr. slov.), pri katerih je tokenizator piko interpretiral kot končno ločilo in jo obravnaval kot ločeno pojavnico (ter jo na ravni stavčne segmentacije označil tudi kot konec stavka), kar je moral označevalec popraviti ročno.

Podobno je bilo z emotikoni, ki so se pogosto pojavljali v strnjenih nizih (npr. :):):**), tokenizator pa jih je (napačno) razdelil na posamezne pojavnice. V takih primerih smo celoten niz zdruţili v eno samo pojavnico:

> :) ¶ :) ¶ : ¶ * ¶ * → :):):**

Poleg ţe znanih zadreg zdruţe vanja in ločevanja elementov pisnega jezika se v naši učni mnoţici pojavlja tudi višji deleţ primerov, v katerih avtor pri zapisu besed narazen oz. skupaj ne sledi standardu (*nebi*), ne uporablja presledkov ob ločilih (*fruktoza-glukoza*) ali vpenja ločila v besede na manj predvidljiv način (*iTunes-ih*, *žen(sk)am*, *politike/o*). Tovrstne pojavnice smo zdruţe vali:[11]

> TV ¶ - ¶ ja → TV-ja
> sms ¶ - ¶ i → sms-i
> ţ en ¶ (¶ sk ¶) ¶ am → ţ en(sk)am
> politik ¶ (¶ e ¶ / ¶ o ¶) → politik(e/o)

4.3 Normalizacija

Pri normalizaciji smo upoštevali načelo minimalne intervencije in besedam nismo pripisovali standardnih sopomenk (npr. *pofarbat → pofarbati* in ne **pobarvati*). Normalizirane so bile besede v nestandardnem zapisu (*priemerjavi → primerjavi, sovascana → sovaščana, mamo → imamo*) ali z nestandardno morfologijo (*na Ptuji → na Ptuju*), v izvirni obliki pa so ostale tviterske prvine (*#krneki*, *@RTV_Slovenija*, *www.youtube.com*), samocenzurirane besede (*p*****, *pof**am*) in jezikovne napake na ravni skladenjskih razmerij (*pri Harry Potterju*, *ne rabim knjigo*), četudi so zelo verjetno naključne (*morajo delajo*). Prav tako pri normalizaciji nismo popravljali izbire besedišča (menjave glagolov *moči-morati*) ali napak na ravni sloga ali registra (*rabiti-potrebovati*).

Pri normalizaciji sta se za najbolj problematični izkazali dve kategoriji besed: nestandardne besede brez neposredne standardne ustreznice in z več različicami zapisa (*orng*, *ornk*, *oreng*, *orenk* ali *fovš*, *favš*, *fouš*, *fauš*, *fowš*) ter tujejezične prvine z različnimi stopnjami prevzetosti na ravneh zapisa in oblikoslovja (*updateati*, *updajtati*, *updejtati*, *apdejtati*), ki jim zgolj s pomočjo referenčnih virov ni bilo mogoče določiti normalizirane ustreznice.

Pri nestandardnih besedah z več različicami zapisa smo normalizirano obliko določili tako, da smo v korpusu JANES s pomočjo regularnih izrazov poiskali vse različice zapisa in izbrali najpogostejšo (v zgornjih primerih sta to *ornk* in *fouš*).

V primeru tujejezičnih prvin bi bila normalizacija v izvirno obliko problematična, saj bi s tem v korpus vnesli umetne oblike, ki jih v realni jezikovni rabi ne najdemo (npr. *poapdejtati → po-update-ati*). Tujejezične prvine smo zato obravnavali po naslednjih kriterijih:

a) če je bila beseda zapisana povsem fonetizirano (npr. *dankešn* „danke schön“, *aprišiejt* „appreciate“), smo jo obravnavali kot slovensko nestandardno besedo z več različicami zapisa (glej *fouš* in *ornk* zgoraj);

b) če je beseda še vedno izkazovala značilnosti tujejezičnega zapisa, npr. neslovenske črke (*wau*) oz. ostanke izvirnega zapisa (*meil*), smo normalizirano obliko določili tako, da smo iz korpusa JANES izbrali najpogostejšo različico med tistimi, ki so še vsebovale značilnosti tujejezičnega zapisa (npr. *updateati*, *updajtati*, *updejtati → updejtati*).

[6] http://www.korpus-gos.net/Support/About
[7] http://nl.ijs.si/imp/
[8] http://fran.si/
[9] http://www.slovenscina.eu/sloleks
[10] Konec stavka oz. mejo med pojavnicami v tem prispevku označujemo s simbolom ¶.
[11] Pri tem je treba omeniti, da napačno zapisanih nizov z manjkajočimi ali odvečnimi presledki (*hodildomov*, *porka duš*) ne popravljamo na nivoju tokenizacije, temveč pri normalizaciji.

4.4 Lematizacija

Pripisovanje lem je v največji moţni meri sledilo smernicam za označevanje korpusa ssj500k (Holozan et al., 2008), ki je v vmesniku SketchEngine sluţil kot referenčni vir za označevalce. Razlike ali dopolnitve označevalnega sistema zadevajo odločitve, vezane na specifike označevanih besedil. Pri tem gre izpostaviti tujejezične prvine in raznovrstne kratice, ki se v spletni slovenščini pojavljajo mnogo pogosteje in oblikovno bolj raznorodno kot v standardnem jeziku.

Med večjimi izzivi označevanja je določanje meje med tujejezičnim in slovenskim besediščem. V tvitih je tujejezičnih prvin veliko, pojavljajo pa se kot posamezne besede različnih besednih vrst in variant zapisa (*share/shareati/share-ati/šerati*), kot besedne zveze ali daljši segmenti. Zadnje smo označevali kot niz pojavitev v tujem jeziku, pri čemer so leme enake oblikam, oblikoskladenjska oznaka po sistemu JOS pa je *Nj*. Podobno velja za občnoimenske besedne zveze (*bonus score, sugar rush*) in posamezne besede, ki so v besedilu zapisane citatno, brez jasno razvidnih prilagoditev slovenskemu zapisu oz. pregibanju (*jailbreak, hrvatskog*). Pri besedah, ki prilagoditev izraţajo, smo lemo določili v skladu s slovenskimi oblikoslovnimi načeli (*benchmarki → benchmark, chatala → chatati*). Pri odločanju, ali besedo obravnavati kot tujejezično ali prevzeto, so bili uporabljeni tudi referenčni leksikalni viri, predvsem SSKJ in SNB ter leksikon besednih oblik Sloleks.

Vprašanja uvrščanja kratičnih poimenovanj med kratice in okrajšave na eni strani ter občna (*lol, drž.*) in lastna imena (*Sds, Slo.*) na drugi so bila rešena ţe na ravni normalizacije. V teh primerih so označevalci pri pripisovanju lem (in oblikoskladenjskih oznak) sledili normaliziranim oblikam.

Projektnospecifična je še odločitev, da se URL-naslovi lematizirajo v domeno (*http://t.co/ZaVQdnaN5p* → t.co), s čimer omogočimo preglednejše prikazovanje korpusnih podatkov v vmesniku. Pri ostalih tviterskih prvinah (uporabniška imena, ključniki, emotikoni) je lema enaka obliki.

4.5 Oblikoskladenjsko označevanje

Tudi na oblikoskladenjski ravni so bile osnovno izhodišče za označevanje smernice korpusa ssj500k. Med razlikami gre v prvi vrsti omeniti prilagoditev oz. širitev sistema za oblikoskladenjsko označevanje JOS, ki so mu bile dodane naslednje nove oznake: *Nh* za ključnike; *Nw* za URL- in e-naslove; *Na* za sklice na uporabniška imena; in *Ne* za emotikone in emojije. Z naštetimi oznakami in načelom lematizacije, pri katerem lema sledi izvorni obliki, smo na enostaven in sledljiv način rešili vprašanje označevanja tvitersko specifičnih prvin.

Pri označevanju niso bile uporabljene oznake *Nt* (zatipkana beseda) ter *Np* (tokenizacijska napaka), saj so bile tovrstne teţave ročno odpravljene ţe na ravni normalizacije.

Zaradi specifik tviterske komunikacije se je pri označevanju pojavljalo večje število pomensko nejasnih oz. dvoumnih primerov (npr. *dobr* kot pridevnik ali prislov). Kot je to veljalo za označevanje ssj500k, so označevalci take primere interpretirali in označili po principu najverjetnejše moţnosti. Podobno načelo je veljalo za označevanje samocenzuriranih besed (*v p***i* → Sozem). V primeru odstopov od norme na skladenjski ravni so bile oznake pripisane skladno z dejansko (in ne pričakovano) pojavitvijo. Tipični tovrstni primeri so na ravni rabe sklonov (*nisem oblikovala intergalaktično brisačo* → Sozet, ne Sozer), števila (*Z Martino smo se tekmovale* → Ggnd-mz, ne Ggnd-dz) in rabe kategorije ţivosti (*jaz vem za kvalitetnega centra z nba izkušnjami* → Sometd, ne Sometn).

Nazadnje je treba omeniti še označevanje zaprtih besednih vrst, ki je skladno z načeli ssj500k v izhodišču potekalo leksikonsko pogojeno, a z moţnostjo dodajanja nestandardnega besedišča. Kategorija, ki je na ta način dobila največ novih elementov, je členek (npr. *eto, evo, ajde, naka, kao, glih/lih* in *ta* v primerih tipa *ta star*). Pri drugih kategorijah se potreba po dopolnitvi pojavlja redkeje, npr. z veznikom *samo* (*Nism še vidu, sam so rekl da je dobr*).

5 Označevalska kampanja

V tem razdelku predstavljamo pregled in opis različnih stopenj označevalske kampanje, ki je zajemala tri stopnje:

a) NTS-Kons1 – normalizacijo, tokenizacijo in stavčno segmentacijo vzorca Kons1 (od decembra 2015 do marca 2016);
b) LO-Kons1 – lematizacijo in oblikoskladenjsko označevanje vzorca Kons1 (od marca 2016 s predvidenim zaključkom avgusta 2016); in
c) NTS-Kons2 – normalizacijo, tokenizacijo in stavčno segmentacijo vzorca Kons2 (od marca 2016 do maja 2016).

5.1 Usposabljanje označevalcev

Ob začetku prvega dela označevalske kampanje (NTS-Kons1) smo priredili dvodnevno delavnico, na kateri so se označevalci seznanili z delom v WebAnnu in s smernicami za označevanje. Na delavnici je sodelovalo 11 študentov jezikoslovnih smeri na magistrski stopnji. Teoretičnemu uvodu v WebAnno s praktičnim delom in predstavitvi smernic je sledila uvajalna označevalska faza, med katero so udeleţenci označili manjše število tvitov. Cilji označevanja so bili naslednji:

a) vsak tvit mora biti pravilno razdeljen na stavke;
b) vsak tvit mora biti pravilno razdeljen na pojavnice; in
c) vse pojavnice morajo imeti pripisano normalizirano obliko; dvoumne pojavnice ohranijo izvirno, nenormalizirano obliko.

Uvajalni označevalski fazi je sledila diskusija, na kateri smo z označevalci razpravljali o njihovih odločitvah in razhajanjih med njihovimi oznakami, podali pa smo tudi pravilne rešitve in razloge zanje, da bi čim bolj uskladili odločitve označevalcev in izboljšali njihovo ujemanje. V drugem delu kampanje (LO-Kons1) smo na enodnevni delavnici označevalce seznanili s konceptom oblikoskladenjskih oznak in lem ter jim predstavili smernice. Tudi tej delavnici je sledila uvajalna faza, cilj pa je bil tokrat vsaki pojavnici v tvitu (z izjemo ločil) pripisati ustrezno lemo in oblikoskladenjsko oznako JOS. Odločitve smo skupaj prediskutirali in utemeljili z načeli iz smernic.

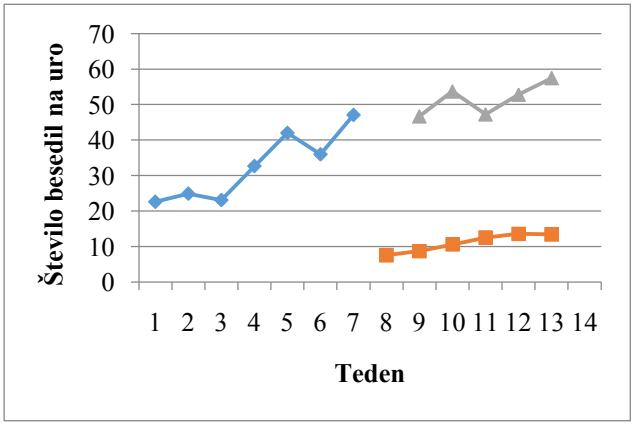
5.2 Preizkušanje označevalcev

Obema delavnicama je sledila preizkusna faza. V delu NTS-Kons1 smo označevalce razdelili v dve skupini po 5 oz. 6 označevalcev, vsaki skupini pa smo dodelili 100 tvitov iz preizkusne množice, pri katerih so morali popraviti avtomatsko pripisane oznake in dodati nove, kjer je bilo to potrebno. Pri oblikoskladenjskih oznakah in lematizaciji smo označevalce razdelili v pare, vsak par pa je označil po 50 tvitov iz preizkusne množice. Oznake sta nato ročno preverila razsodnika, ki sta ocenila tudi natančnost označevalcev. Na podlagi rezultatov sta bila v delu NTS-Kons1 iz kampanje izključena dva nezanesljiva označevalca, v obeh delih pa sta razsodnika po začetni evalvaciji dopolnila smernice za označevanje še s primeri, ki so se v preizkusni seji izkazali za problematične.

5.3 Delotok označevanja

Delotok označevanja je vključeval skupino označevalcev in dva razsodnika z dobrim poznavanjem smernic za označevanje. Razsodnika, ki sta bila zadolžena tudi za vodenje označevalske kampanje, sta v tedenskih fazah posamezni skupini označevalcev[12] dodelila določeno število datotek, po koncu vsake faze pa sta oznake ročno preverila in, če je bilo potrebno, označevalcem podala konstruktivno povratno informacijo ter na ta način odstranila najpogostejše oz. najresnejše napake. Če so označevalci med delom naleteli na posebno problematično dilemo, so bile z njo dopolnjene tudi smernice za označevanje. Ustvarjen je bil tudi e-poštni seznam, na katerem so lahko označevalci razsodnikoma zastavljali vprašanja in razreševali problematične ali dvoumne primere, ki niso bili vključeni v smernice.

Med delom smo spremljali učinkovitost označevalcev, tako da smo v vsaki fazi merili razmerje med časom označevanja in številom označenih besedil (glej Sliko 3).



Slika 3: Učinkovitost označevalcev pri označevanju vzorcev Kons1 in Kons2. Modri del predstavlja NTS-Kons1, sivi NTS-Kons2 in oranžni LO-Kons1.

Iz grafa je razvidno, da normalizacija, tokenizacija in stavčna segmentacija potekajo mnogo hitreje od lematizacije in oblikoskladenjskega označevanja. Dobro usposobljeni označevalci lahko v eni uri normalizirajo med 45 in 55 besedil, lematizirajo in oblikoskladenjsko označijo pa le nekaj nad 10 besedil.

6 Rezultati in diskusija

V tem razdelku podamo strnjeno kvantitativno analizo označenih vzorcev in razpravljamo o najpogostejših razhajanjih med označevalci na vseh nivojih označevanja.

6.1 Kvantitativna analiza rezultatov

Tabela 1 prikazuje velikosti do zdaj označenih podatkovnih množic. V prvi vrstici so navedeni podatki za oblikoskladenjsko označeni del Kons1, v drugi in tretji za celotna Kons1 in Kons2, v zadnji pa za njuno vsoto.

	Besedila	Stavki	Besede	Pojavnice	Norm.	Norm. %
Kons1-MSD	880	2.365	20.537	23.958	4.888	20,4
Kons1	3.940	9.976	86.593	102.719	11.881	11,6
Kons2	1.927	4.473	34.583	41.056	4.728	11,5
Kons	5.867	14.449	121.176	143.775	16.609	11,6

Tabela 1: Velikost označenih vzorcev.

Stolpci od leve proti desni podajajo število označenih besedil, število ročno preverjenih stavkov, število preverjenih besednih pojavnic v izvornem besedilu, število vseh pojavnic ter število (in nazadnje delež) pojavnic, ki jim je bila pripisana normalizirana oblika. Vseh ročno preverjenih besed je več kot 120.000. Več kot desetina je bila potrebna normalizacije.

Oblikoskladenjski del je bistveno manjši, saj je trenutno ročno označenih le približno 20.000 besed. Predpostavljamo pa, da je ta količina zadostna za preverjanje točnosti označevanja z razvitimi orodji in za dopolnjevanje učne množice, da lahko oblikoskladenjski označevalniki bolje označujejo uporabniške spletne vsebine. Besedila, ki so bila oblikoskladenjsko označena do zdaj, so v povprečju tudi bolj nestandardna kot preostanek vzorca Kons1 - normalizacije je bilo namreč potrebnih dobrih 20 % pojavnic.

Omeniti je treba tudi leksikon oblikoskladenjsko označenega vzorca Kons1, ki vsebuje vsega skupaj 8.033 različnih izvornih pojavnic (vključno z ločili). Od tega je normaliziranih 7.305 pojavnic (skoraj 91 %). Normalizirane pojavnice imajo 5.548 različnih lem, označene pa so bile s 592 različnimi oblikoskladenjskimi oznakami.

6.2 Najpogostejša razhajanja pri označevanju

Na ravni stavčne segmentacije je do razhajanja prišlo predvsem pri stavkih, ki niso vsebovali nobenih klasičnih ločil in so bili npr. razdeljeni z večpičji, ki jih je bilo mogoče interpretirati bodisi kot zamolk bodisi kot konec stavka.

Pri normalizaciji so bile glavni vir razhajanja besede, ki jih je bilo mogoče normalizirati v več različnih oblik (npr. *k* → *ker/ko/ki/kjer/kot* itd. ali *sm* → *sem*/*samo*), v

[12] Na začetku so bili označevalci razdeljeni v skupine po 3, pozneje pa v pare. Zelo natančni označevalci so v nekaterih fazah označevali tudi posamezno.

omejenem kontekstu pa je bila interpretacija stvar posameznega označevalca.

Na ravni lematizacije in oblikoskladenjskega označevanja je mogoče razhajanja med označevalci in njihove napake pripisati več razlogom. V prvo skupino sodijo objektivno odpravljive probleme, ki jih označevalci razumejo, a v praksi pogosto spregledajo, npr. enakopisne oblike (npr. *si* kot oblika glagola *biti* ali povratni zaimek v dajalniku; *da* kot oblika glagola *dati* ali podredni veznik). V drugi skupini so razhajanja, ki so posledica dveh različnih, a znotraj sistema legitimnih interpretacij (npr. *Džizs, to bi b'lo fajn,* kjer je prvo besedo mogoče uvrstiti med občnoimenske ali lastnoimenske samostalnike ali pa med medmete).

Zadnja skupina razkriva teţave s smernicami za označevanje, bodisi ker so slednje nejasno napisane ali pa ker predvidevajo rešitve, ki so manj intuitivne ali odstopajo od siceršnjih načel sistema. Med pogostimi napakami tega tipa so denimo pozabljeni popravki stopnjevanih prislovov tipa *večji/največji* (izjema, pri kateri je lema enaka stopnjevani in ne osnovni obliki, oblikoskladenjska oznaka pa izraţa kategorijo stopnjevanosti) ali napake pri besednovrstnem uvrščanju povedkovega določila tipa *je bilo lepo*, ki se po smernicah ssj500k označuje kot pridevnik srednjega spola, označevalci pa ga razumejo kot prislov. Pri tem je nujno omeniti, da smernice temeljijo na referenčnih virih za slovenščino in v določeni meri preslikavajo kategorizacijske teţave pri primerih, ki lahko nastopajo v različnih vlogah (npr. *nič, ves, kaj, prav* ipd.). Na ravni označevanja kratic in lastnih imen (predvsem tujih) je opremljenost še toliko bolj pomanjkljiva, saj tovrstno besedišče praviloma v vire ni zajeto. Teţavam, identificiranim v tej skupini, bi se bilo pri nadaljnjem razvoju označevanja slovenščine smiselno natančneje posvetiti.

7 Zaključek

V prispevku smo povzeli ključne vidike smernic za označevanje učnega korpusa na različnih ravneh in na kratko predstavili rešitve za označevanje prvin, specifičnih za spletno komunikacijo. Cilj priprave korpusa, ki bo prosto dostopen na repozitoriju CLARIN.SI, je dvojni: uporaben bo kot učna mnoţica za izboljšanje označevanja spletnih besedil, smernice za označevanje učnega korpusa pa bodo ponudile prvi celoviti vpogled v problematiko jezikoslovnega označevanja slovenske računalniško posredovane komunikacije ter rešitve za najbolj problematične primere. Izdelava učnega korpusa ponuja tudi moţnosti za dopolnitev obstoječih leksikonov besednih oblik z nestandardnim besediščem (npr. z besedami *evo*, *eto*, *ajde*, *naka*, *kao*, *glih* ipd. v kategoriji členkov).

8 Zahvala

Avtorji se najlepše zahvaljujejo Kaji Dobrovoljc, Simonu Kreku in Katji Zupan za konstruktivne pripombe pri izdelavi smernic za označevanje, ter vsem označevalcem, ki so sodelovali v označevalski kampanji: Teji Goli, Melaniji Koţar, Vesni Koţelj, Poloni Logar, Klari Lubej, Dafne Marko, Barbari Omahen, Eneji Osrajnik, Predragu Petroviću, Poloni Polc, Aleksandri Rajković in Izi Škrjanec.

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, ter s podporo Slovenske raziskovalne infrastrukture za jezikovne vire in tehnologije (CLARIN.SI).

9 Literatura

Iñaki Alegria, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo in Arkaitz Zubiaga. 2014. TweetNorm_es Corpus: an Annotated Corpus for Spanish Microtext Normalization. V: *Zbornik konference Ninth International Conference on Language Resources and Evaluation (LREC2014)*. ELRA, Reykjavik-Paris.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay in Li Wang. 2013. How Noisy Social Media Text, How Diffrnt Social Media Sources. V: *Sixth International Joint Conference on NLP*, str. 356–364.

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter in Wei Xu: Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. V: *Workshop on Noisy User-generated Text at ACL 2015, July 31, 2015*. Peking, Kitajska.

Michael Beißwenger, Thomas Bartz, Angelika Storrer in Swantje Westpfahl. 2015a. Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. V: *Guideline document from the Empirikom shared task on automatic linguistic annotation of internet-based communication (EmpiriST 2015)*.

Michael Beißwenger, Sabine Bartsch, Stefan Evert in Kay-Michael Würzner. 2015b. Richtlinie für die manuelle Tokenisierung von Sprachdaten aus Genres internetbasierter Kommunikation. V: *Guideline document from the Empirikom shared task on automatic linguistic annotation of internet-based communication (EmpiriST 2015)*.

David Crystal. 2001. *Language and the Internet*. Cambridge: Cambridge University Press.

Thierry Declerck in Piroska Lendvai. 2015. Processing and Normalizing Hashtags. V: *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference, 7–9 September 2015*, str. 104–109, Hissar, Bolgarija.

Richard Eckart de Castilho, Chris Biemann, Iryna Gurevych in Sied Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. V: *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands.

Tomaţ Erjavec. 2011. Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. V: *5th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*. Portland: Association for Computational Linguistics, 2011, str. 33–38. http://aclweb.org/anthology-new/W/W11/W11-1505.pdf.

Darja Fišer, Nikola Ljubešić in Tomaţ Erjavec. 2015. The JANES corpus of Slovene user generated content: construction and annotation. V: *International Research Days: Social Media and CMC Corpora for the*

eHumanities: Book of Abstracts, 23–24 October 2015, str. 11, Rennes, Francija.

Peter Holozan, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman in Aleš Velušček. 2008. *Specifikacije za učni korpus (kazalnik 2): projekt Sporazumevanje v slovenskem jeziku*. Kamnik. Dostopno na: http://projekt.slovenscina.eu/Vsebine/Sl/Kazalniki/K2.aspx.

Nikola Ljubešić, Tomaţ Erjavec in Darja Fišer. 2014. Standardizing tweets with character-level machine translation. V: *Computational linguistics and intelligent text processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6–12, 2014: proceedings: part II, (Lecture notes in computer science, 8404)*, str. 164–175, Heidelberg: Springer, 8404.

Nikola Ljubešić, Darja Fišer, Tomaţ Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. V: *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference, 7–9 September 2015*, str. 371–378, Hissar, Bolgarija.

Nikola Ljubešić, Tomaţ Erjavec in Darja Fišer. 2016. Corpus-Based Diacritic Restoration for South Slavic Languages. V: *Zbornik konference Tenth International Conference on Language Resources and Evaluation (LREC2016)*. ELRA. Portoroţ, Slovenija, str. 3613–3616.

Kamel Nebhi, Kalina Bontcheva in Genevieve Gorrell. 2015. ResToRinG CaPitaLiZaTion in #TweeTs. V: *Zbornik konference 24th International Conference on World Wide Web (WWW '15 Companion)*. ACM, New York, NY, USA, str. 1111–1115.

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf in Christopher Richards. 2001. Normalization of non-standard words. V: *Computer Speech and Language, 15 (3)*, str. 287–333.

Preizkus Googlovega govornega programskega vmesnika pri samodejnem razpoznavanju govorjene slovenščine

Simon Dobrišek, David Čefarin, Vitomir Štruc, France Mihelič

Laboratorij za umetno zaznavanje, sisteme in kibernetiko,
Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1000 Ljubljana
{simon.dobrisek,vitomir.struc,france.mihelic}@fe.uni-lj.si, david.cefarin@gmail.com

Povzetek

Samodejni razpoznavalniki govora počasi dozorevajo v tehnologije, ki omogočajo človeku bolj naravne oblike komuniciranja z različnimi pametnimi napravami in informacijsko-komunikacijskimi sistemi. Velika svetovna podjetja, kot so Google, Microsoft, Apple, IBM in Baidu, tekmujejo pri razvoj čim bolj zanesljivih razpoznavalnikov govora, ki podpirajo čim več pomembnih svetovnih jezikov. Zaradi svoje majhnosti pa podpora govorjeni slovenščini pri govornih tehnologijah zaostaja in med velikimi svetovnimi podjetji je naš govorjeni jezik kot prvi podprl samo Google. V članku predstavljamo rezultate našega neodvisnega preizkusa Googlovega govornega programskega vmesnika pri samodejnem razpoznavanju govorjene slovenščine. Za preizkus so bile uporabljene govorne zbirke, ki jih tudi sicer uporabljamo za razvoj in preizkušanje razpoznavalnikov govorjene slovenščine.

Assessment of the Google Speech Application Programming Interface for Automatic Slovenian Speech Recognition

Automatic speech recognizers are slowly maturing into technologies that enable humans to communicate more naturally and effectively with a variety of smart devices and information-communication systems. Large global companies such as Google, Microsoft, Apple, IBM and Baidu compete in developing the most reliable speech recognizers, supporting as many of the main world languages as possible. Due to the relatively small number of speakers, the support for the Slovenian spoken language is lagging behind, and among the major global companies only Google has recently supported our spoken language. The paper presents the results of our independent assessment of the Google speech-application programming interface for automatic Slovenian speech recognition. For the experiments, we used speech databases that are otherwise used for the development and assessment of Slovenian speech recognizers.

1 Uvod

Prostoročno govorno ukazovanje avtomatiziranim in robotiziranim sistemom, govorna komunikacija z virtualnimi osebnimi asistenti na pametnih telefonih in tablicah, narekovanje diagnoz, pravnih in drugih besedil, elektronskih pisem in kratkih sporočil, samodejno tvorjenje zapisnikov sestankov, sej in telefonskih konferenc, samodejno podnaslavljanje in prepisovanje oddaj in drugih več-medijskih vsebin, preverjanje pravilnosti izgovarjave pri učenju jezika in tako dalje - vse to so primeri uporabe govornih tehnologij, ki vključujejo samodejno razpoznavanje govora.

Razvoj samodejnih razpoznavalnikov govora je že desetletja zahteven izziv, ki ga spremljajo velika pričakovanja, žal pa tudi razočaranja. Govor je človeku najbolj naraven način komuniciranja, vendar ima prav zaradi svoje naravnost določene značilnosti, ki otežujejo razvoj po vsem zanesljivih samodejnih razpoznavalnikov. Zaradi okoljskih, fizioloških, sociolоških, razpoloženjskih in drugih vplivov govor posameznika izkazuje precejšnjo spremenljivost, s katero se tudi najsodobnejši razpoznavalniki govora le s težavo spopadajo. Še posebej to velja za govorjeno slovenščino, ki ima v primerjavi z nekaterimi večjimi jeziki, kot je angleščina, večje število pregibnih oblik besed in bolj prost besedni red.

Ne glede na navedeno pa je bil v zadnjih nekaj letih dosežen preboj in znaten napredek na tem področju. Razpoznavalniki govora postajajo vse bolj zanesljivi in vse več ljudi jih uporablja pri svojem vsakdanjem delu. Še posebej to velja za uporabnike, ki pri svojem delu govorno komunicirajo v enem od večjih svetovnih jezikov, kot so angleščina, španščina, portugalščina, francoščina, nemščina, italijanščina in kitajščina. Velika svetovna podjetja, kot so Google, Microsoft, Apple, IBM in Baidu, so razvila vrsto komercialnih računalniških programskih rešitev, ki vključujejo že solidno zanesljive razpoznavalnike govora, kot so Microsoft Cortana, Skype Translator, Xbox, Google Now, Apple Siri, IMB Watson Speech Recognition, Baidu Voice Search in programske rešitve podjetja Nuance.

Po črnogledih pričakovanjih pa pri navedenih programskih rešitvah podpora govorjeni slovenščini še izostaja ali vsaj zaostaja. Pozitivno presenečenje in izjema je zaenkrat le Googlov oblačni razpoznavalnik govora, ki je pred približno dvema letoma podprl tudi govorjeno slovenščino. Na demonstracijski strani Googlovega govornega programskega vmesnika[1] je že nekaj časa v spletnem brskalniku Google Chrome možno izbrati in preizkušati delovanje samodejnega razpoznavanja govorjene slovenščine. Od nedavnega je naš govorjeni jezik podprt tudi v brezplačnem Googlovem spletnem prevajalniku[2], kjer se lahko vhodno besedilo narekuje tudi v slovenščini, in tudi drugih Googlovih storitvah.

Rezultati osnovnega uporabniškega preizkusa točnosti in zanesljivosti Googlovega samodejnega razpoznavalnika govorjene slovenščine so presegli naša pričakovanja. Poleg tega pa smo dobili subjektivni vtis, da se točnost razpoznavanja sčasoma še izboljšuje. Zato smo se odločili, da Googlov govorni programski vmesnik sistematično preizkusimo z našimi obstoječimi govornimi zbirkami (Mihelič et al., 2003) in neodvisno ovrednotimo

[1] Google Web Speech API Demonstration - www.google.com/intl/en/chrome/demos/speech.html

[2] Google Translate - http://translate.google.si

zanesljivost njegovega delovanja in točnost razpoznavanja govorjene slovenščine.

V tem prispevku poročamo o naših izkušnjah in ugotovitvah, ki smo jih pridobili s preizkusom. Podajamo tudi kratek pregled napredka tehnologij in izpostavljamo tiste značilnosti sodobnih razpoznavalnikov govora, ki so največ pripomogli k napredku na tem področju.

2 Razvoj razpoznavalnikov govora

Prikriti Markovovi modeli (angl. Hidden Markov Models - HMM) in modelimešanic Gaussovih porazdelitev (angl. Gaussian Mixture Models - GMM) so leta prevladovali kot najbolj uspešen zgled akustičnega in jezikovnega modeliranja, nakaterem so te meljili samodejni razpoznavalniki tekočega govora z velikimi besednjaki (angl. Large-Vocabulary Continuous Speech Recognition - LVCSR). Za posebna področja uporabe z omejenim obsegom ožjega strokovnega jezika so tovrstni komercialni razpoznavalniki govora že dajali zadovoljive rezultate, denimo pri samodejnem prepisovanju narekovanih medicinskih diagnoz ali pa pravnih mnenj, strokovnih poročil in podobno.

V zadnjih nekaj letih pa je bil dosežen velik napredek z zamenjavo omenjenega nekdaj prevladujočega zgleda modeliranja z novim zgledom, ki temelji na uporabi t. i. globokih nevronskih omrežjih (angl. Deep Neural Networks - DNN) (Hinton et al., 2012; Deng et al., 2013; Siniscalchi et al., 2013; Yu in Deng, 2015). Nadaljnji znatni napredek je bil dosežen z uporabo t. i. konvolucijskih nevronskih omrežij (angl. Convolutional Neural Networks - CNN) (Sainath et al., 2013; Abdel-Hamid et al., 2014) in povratnih nevronskih omrežij (angl. Recurrent Neural Networks - RNN), ki modelirajo dolgi kratkoročni spomin (angl. Long Short-Term Memory - LSTM) (Sainath et al., 2015; Ravuri in Stolcke, 2015).

Nevronska omrežja niso novost, saj so dobro poznana že več kot pol stoletja. Za akustično modeliranje so se poskušala uporabljati že od začetka razvoja razpoznavalnikov govora (Juang in Rabiner, 2005). Vendar pa je vse do okoli leta 2010 njihov potencial na področju razpoznavanja govora ostal dokaj neizkoriščen. Zaradi omejene računske moči računalnikov in razmeroma zahtevnih učnih postopkov so bili doseženi rezultati razpoznavalnikov nezadovoljivi. Šele uspešno sodelovanje raziskovalcev z Univerze v Torontu z globalnimi podjetji, kot sta Microsoft in Google, je odigralo ključno vlogo pri uveljavljanju nevronskih omrežji v komercialnih razpoznavalnikih govora (Deng 2016).

Raziskovalci so razvili učinkovita orodja za učenje globokih nevronskih omrežij, poleg tega pa je bila omogočena tudi lažja in bolj učinkovita uporaba grafičnih procesorjev za splošne računske naloge, predvsem z izdajo programske knjižnica Cuda (NVIDIA). Z novimi odkritji je bilo možno pospešiti učenje obsežnih nevronskih omrežij na velikih količinah podatkov. Obsežne govorne zbirke zvočnih posnetkov so bile na voljo, saj so raziskovalne institucije zbirale, prepisovale in označevale govorne posnetke, primerne za učenje razpoznavalnikov govora, že v sajtrih deset letih. V siti dejavniki so omogočili uspešno sodelovanje raziskovalcev iz podjetij in univerz, kar je odprlo novo poglavje pri razvoju samodejnega razpoznavanja govora. Po letu 2010 je bilo na znanstvenih konferencah IEEE ICASSP in Interspeech sprejetih vedno več znanstvenih člankov o uporabi metod nevronskih omrežij pri razpoznavanju govora. Med komercialnimi aplikacijami pa se je uporaba nevronskih omrežij prav tako zelo hitro širila in danes jih uporablja že večina sistemov za avtomatsko razpoznavanje govora.

Pri Googlovi aplikaciji za samodejno razpoznavanje govora Google Voice so sčasoma zamenjali predkrmiljeno (angl. feedforward) omrežje s povratnim omrežjem z dolgim kratkoročnim spominom (LSTM). Ta omrežja imajo v primerjavi s prejšnjimi dodatne povratne povezave in spominske celice, ki jim omogočajo, da si »zapomnijo« pretekle podatke. Ta lastnost izboljša razpoznavanje, saj omrežje s tem modelira kontekst glasov oziroma besed, kar izboljša modeliranje govora.

2.1 Podpora za govorjeno slovenščino

Med danes najbolj razvitimi komercialnimi govornimi vmesniki je zaenkrat uspelo le Googlu podpreti tudi govorjeno slovenščino. Osnovni preizkus govornega vmesnika pri razpoznavanju govorjene slovenščine je možen na omenjeni Googlovi demonstracijski spletni strani, od nedavnega pa je v spletnem brskalniku Chrome razpoznavanje našega govorjenega jezika podprta tudi na popularni spletni strani Googlovega prevajalnika besedil. Pojavlja se tudi vse več aplikacij za pametne telefone in tablice, ki podpirajo govorjeno slovenščino, vendar po večini vse temeljijo na uporabi Googlovega vmesnika. Applov vmesnik Siri in Microsoftova Cortana pa zaenkrat še ne podpirata govorjene slovenščine in tudi ni novic, da se bo to v kratkem zgodilo.

3 Googlov govorni programski vmesnik

Googlov govorni programski vmesnik je zmogljiv, vendar razmeroma slabo dokumentiran in še vedno omejeno dosegljiv za širšo skupnost razvijalcev novih informacijsko-komunikacijskih aplikacij. Zaradi pomanjkanja dokumentacije točna sestava sistema javno še ni povsem znana, ljubitelju pa so nekateri razvijalci pridobili in objavili nekaj informacij o tem, kako se vmesnik lahko uporablja.

Do Googlove govornega programskega vmesnika tako lahko dostopamo z različnimi orodji ali programskimi jeziki, ki podpirajo običajno internetno komunikacijo. Vmesnik se najbolj enostavno in najpogosteje uporablja z uporabo Javascript programskega jezika in programskega vmesnika Google Javascript Speech API. V tem primeru se aplikacije, ki uporabljajo Googlov govorni vmesnik, razvijejo kot običajne spletne aplikacije, ki se naložijo v spletni brskalnik. Omejitev je le ta, da se mora aplikacijo naložiti in uporabljati v spletnem brskalniku Google Chrome, ki zaenkrat edini podpira vse komponente govornega programskega vmesnika. Delovanje takšnih aplikacij ponazarja omenjena Googlova demonstracijska spletna stran Google Web Speech API Demonstration. Za vključevanje Googlovega govornega programskega vmesnika z uporabo Javascript API obstaja tudi dokumentacija, podprta s strani Googla (Payton, 2014).

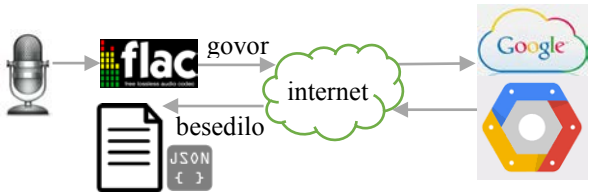
Za neposredno uporabo Googlovega govornega programskega vmesnika iz samostojne aplikacije, t. j. aplikacije, ki ni odvisna od uporabe spletnega brskalnika, pa potrebujemo ključ, ki je registriran v Googlovem oblačnem sistemu programskih vmesnikov. Ključ, ki

istoveti našo aplikacijo in ji omogoča dostop do Googlovih programskih vmesnikov, lahko pridobimo s prijavo v skupino razvijalcev aplikacij Chromium Development Group na Googlovi spletni strani https://console.developers.google.com.

Po pridobitvi ključa, ki istoveti našo aplikacijo, pa lahko Googlov govorni programski vmesnik brezplačno uporabimo le do petdesetkrat na dan in še to z omejitvijo pri dolžini obdelanih zvočnih posnetkov, ki so lahko dolgi do 15 sekund. Za intenzivnejšo uporabo govornega programskega vmesnika je potrebno plačilo po posebnem ceniku. Višina plačila je odvisna od števila uporab programskega vmesnika in dolžine vseh posnetkov, ki jih je govorni programski vmesnik obdelal.

Za raziskovalne namene in namene preizkušanja govornega programskega vmesnika pa je možna uporaba ključa, ki sicer istoveti spletni brskalnik Google Chrome. V tem primeru se lahko neka preizkusna aplikacija Googlovemu oblačnemu sistemu programskih vmesnikov dejansko predstavi na enak način kot spletni brskalnik, samo aplikacijo pa lahko spišemo v poljubnem programskem jeziku (Java, Python, C#, C++, ipd), ki podpira hkratno dvosmerno komunikacijo s strežniki z uporabo običajnih spletnih protokolov (HTTPS ipd).

Google govorni programski vmesnik lahko uporabljamo na dva načina. V prvem načinu se povežemo s Googlovim oblačnim strežnikom na internetnem naslovu https://www.google.com/speechapi/v2/recognize. Po zaključenem pošiljanju zvočne datoteke nam strežnik po istem komunikacijskem kanalu vrne rezultate razpoznavanja. Ta programski vmesnik omogoča razpoznavanje posameznih posredovanih govornih posnetkov v dolžini do okoli 15 sekund.



Slika 1: Simbolni prikaz povezave med aplikacijo in Googlovim govornim programskim vmesnikom.

Druga možnost je povezava s strežnikom na naslovih https://www.google.com/speech-api/full-duplex/v1/up in https://www.google.com/speech-api/full-duplex/v1/down. V tem primeru aplikacija s strežnikom vzpostavi dvosmerno hkratno komunikacijo v t. i. načinu »full-duplex«. To pomeni, da naša aplikacija s strežnikom hkratno komunicira v obe smeri, pri čemer strežniku pošilja zvočni signal, hkrati pa s prejema razpoznano besedilo. V tem primeru so lahko poslani govorni signali tudi daljši od 15 sekund, saj Googlov programski vmesnik sam po potrebi razdeli na primerne odseke in vrača rezultate razpoznavanja govora v več delih. Pri tem govorni programski vmesnik sprejema zvočne signale, kodirane v formatu FLAC (angl. Free Lossless Audio Codec) in Speex. Slednji ni standardiziran, zato je priporočena uporaba formata FLAC. Rezultate razpoznavanja strežnik vrača v datotečnem formatu JSON (angl. JavaScript Object Notation). Simbolni prikaz povezave med aplikacijo in Googlovim govornim programskim vmesnikom je prikazana na sliki 1.

Parametri internetne povezave z govornim programskim vmesnikom omogočajo različne nastavitve načina delovanja razpoznavalnika govora. Med drugim tako lahko nastavljamo stopnjo zakrivanja/filtriranja neprimernih besed (psovke ipd) v rezultatu razpoznavanja (parameter `pFilter`), izbiramo lahko število alternativnih oz. manj verjetnih rezultatov razpoznavanja, ki jih želimo pridobiti (parameter `maxAlternatives`), vključimo lahko stikalo za razpoznavanje neprekinjenega tekočega govora (parameter `continuous` - razpoznavalnik potem pošilja rezultate obsamodejno zaznanih premorih v govoru) ter stikalo, ki omogoči pošiljanje vmesnih rezultatov razpoznavanja še pred koncem izrečenih povedi (parameter `interim`) in drugi.

4 Preizkus govornega vmesnika

Preizkus Googlovega govornega vmesnika smo izvedli z uporabo lastnih izbranih preizkusnih govornih posnetkov, ki smo jih pridobili med raziskovalnim in razvojnim delom na področju govornih tehnologij. Največji del preizkusnih posnetkov smo uporabili iz naših govornih zbirk GOPOLIS (Mihelič et al., 2003) in VNTV (Žibert in Mihelič, 2000; Mihelič et al., 2003). Poleg omenjenih govornih zbirk smo za preizkus uporabili tudi nekaj dodatnih, posebej pridobljenih zvočnih posnetkov prebiranja elektronskih pisem v slovenščini (dolžine okoli 100 besed). Elektronska pisma so se prebirala na več načinov in sicer najprej počasi in razločno, nato normalno hitro in nekaj manj razločno ter nato še povsem spontano, manj razločno, z medmeti in s prekinitvami.

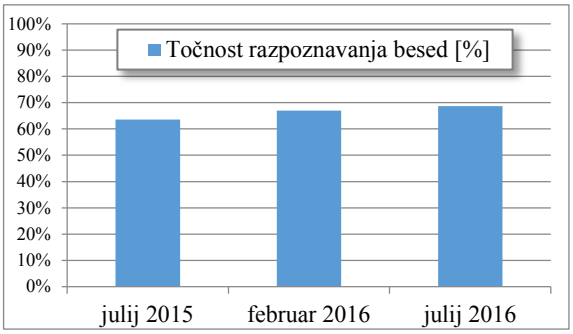
4.1 Rezultati na govorni zbirki GOPOLIS

Zbirka GOPOLIS vsebuje govorne posnetke posameznih branih povedi govornih pozivedovanj po letalskih informacijah. Po analizi dejanskih dialogov med uporabniki in uslužbenko klicnega centra za podajanje letalskih informacij je bil izdelan generator najbolj pogostih povedi. Med njimi se je nato naključno izbralo in vsakemu govorcu pripisalo od 20 do 172 povedi, ki jih je ta v bolj ali manj nadzorovanem okolju kolikor je mogoče sproščeno prebral na način, kot bi te povedi izgovoril pri komunikaciji s klicnim centrom.

Za preizkus Googlovega razpoznavalnika govora smo izbrali 10 testnih govorcev iz izvirne govorne zbirke GOPOLIS ter še 12 dodatnih govorcev, naključno izbranih med študenti, ki so imeli v okviru enega od laboratorijskih vaj pri predmetu Razpoznavanje vzorcev za nalogo zbrati nekaj število govornih posnetkov oz. govornih vzorcev. Pri študentskih posnetkih se je prav tako prebiralo naključno izbrane povedi, tvorjene z omenjenim generatorjem. Pri zbiranju posnetkov so bili študenti povsem prosti pri izbiri snemalnega programa, mikrofona in avdio opreme.

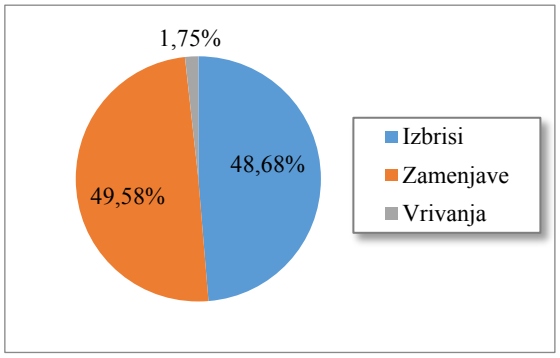
Skupaj 1925 testnih posnetkov povedi v skupnem trajanju dobre 1 ure in 37 minut se je predeko posebej izdelanega programskega vmesnika, kije te meljil na uporabi programskega orodja `curl`, poslalo Googlovemu govornemu programskemu vmesniku in pridobilo vse rezultate razpoznavanja. Pridobljene prepise govora se je nato neposredno primerjalo z referenčnimi prepisi povedi in na običajen način ugotavljalo število napak v smislu števila zamenjanih, vrinjenih in izbrisanih besed. Preizkus smo v zadnjem letu dni izvedli trikrat, ker nas je zanimalo, če se bo rezultat zaradi morebitnega prilagajanja

Googlovega r azpoznavalnika g ovora že o bdelanim govornim posnetkom v vmesnem času kaj izboljševal. Pridobljeni rezultati so potrdili, da so bila ta pričakovanja upravičena.



Slika 2: Točnost razpoznavanja besed, ugotovljena v različnih časovnih obdobjih.

Skupna točnost razpoznavanja izgovorjenih povedi se je pri preizkusih v času od julija 2015 do junija 2016 gibala od 63% do skoraj 70% pravilno razpoznanih besed. Med govorci so bile znatne razlike, saj je bila ugotovljena točnost razpoznavanja pri različnih govorcih od le 15% pa vse do 77% pravilno r azpoznanih b esed. P ri n ajslabših rezultatih je večina napak posledica dejstva, da razpoznavalnik zaradi različnih razlogov (odsotnost tišine na ko ncu p osnetka a li p risotnost ka kšnih i zrazitih motenj v p osnetku, kot j e hr up p remikanja stola ipd) n i vrnil dokončnega rezultata razpoznavanja i n s o s e vse b esede štele za i zbrisane. Zaradi večjega števila lastnih imen krajev, letališč in letalskih prevoznikov, ki so omenjena v preizkusnih p ovedih j e d el napak r azpoznavalnika t udi posledica n apak pri n jihovem or tografskem z apisu. T ako je bi lo, de nimo, l astno ime *Sheremetyevo* pogosto razpoznano ko *Šeremetjevo*, a li p a *Zuerich* kot *Cirih*, ka r se j e štelo kot napake za menjave. Povzetek r ezultatov je podan na sliki 2.



Slika 3: Pogostost različnih vrst napak pri razpoznavanju besed v preizkusnih povedih.

Analizirali smo tudi, kakšna je pogostost različnih vrst napak, torej napak izbrisa, zamenjave in izbrisa besed. Po pričakovanjih so prevladovale napake zamenjave in izbrisa, medtem ko j e b il d elež n apak vrivanja precej manjši (slika 3) . T o pr ipisujem de jstvu, da j e G ooglov razpoznavalnik večkrat vrnil prazno besedilo in, kot je že bilo omenjeno, se j e to navadno dogajalo, kadar posnetek povedi ni bila zaključene z dovolj premora oz. tišine.

Po pričakovanjih Googlov govorni vmesnik ni dosegel rezultatov razpoznavalnika, ki smo ga razvili sami in je bil razvit in posebej prilagojen danemu področju uporabe (poizvedovanje p o l etalskih informacijah). Naš r azpoznavalnik je s icer že nekoliko zas tarel i n temelji š e n a uporabi pr ikritih M arkovovih modelov (Dobrišek e t a l., 2006), v endar u porablja posebne kanonične akustične modele za sprotno pr ilagajanje govornega m odela na govorne značilnosti posameznih govorcev.

Z našim razpoznavalnikom govora smo na še večjem številu podobno i zbranih t estnih pos netkov dosegli rezultat preko 81% pravilno razpoznavanja besed. Ključen razlog za znatno boljši r ezultat pa je pr edvsem v tem, da naš r azpoznavalnik u porablja verjetnostni jezikovni model, ki i ma bistveno nižjo p erpleksnost in je b istveno bolj prilagojen ožjemu področju uporabe, ko t p a ga uporablja G ooglov r azpoznavalnik g ovora, s aj t a t emelji na splošnem generičnem jezikovnem modelu s precej višjo perpleksnostjo.

4.2 Rezultati na govorni zbirki VNTV

Poleg g ovorne zb irke GOPOLIS smo za p reizkus Googlovega r azpoznavalnika g ovora u porabili še 1548 televizijskih p osnetkov vremenskih napovedi p etih govorcev i z govorne z birke VNTV v s kupnem t rajanju 1 ure in 48 minut. Rezultati in ugotovitve so podobne kot pri preizkusu s p osnetki govorne z birke G OPOLIS. Skupna točnost razpoznavanja besed je bila 62,3%, pri čemer je bil zn aten d eleže napak ponovno pripisan iz brisom b esed pri govornih posnetkih, za ka tere vmesnik s ploh ni vrnil prepisa, k er t i niso bili zaključeni z dovolj dolgim premorom oz. tišino. Še na jvečje presenečenje je bilo, da je bil rezultati (81,1 %) pri ženski govorki (napovedovalka Tanja C egnar) v povprečju znatno boljši od r ezultatov moških g ovorcev (od 54% do 70%). Navadno se namreč pri p reizkusih samodejnih r azpoznavalnikov govora izkaže ravno obratno in imajo ti večje težave z ženskimi govorkami.

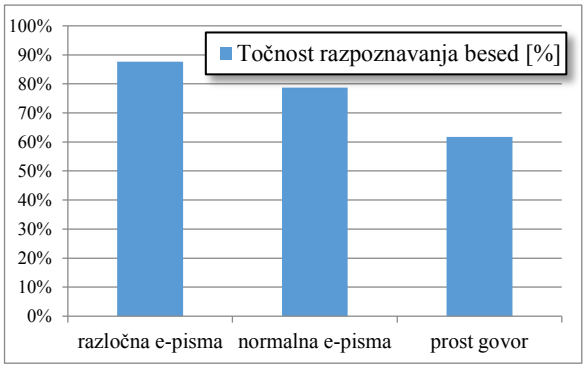
4.3 Rezultati pri narekovanju e-pisem

Zadnji p reizkus, k i s mo ga iz vedli v o kviru te raziskave, pa se je nanašal na razvoj uporabne aplikacije govornega narekovanja elektronskih pisem oz. sporočil. Za pr eizkus smo pr idobili pos netke e nega moškega govorca, ki je prebiral elektronska sporočila v skupnem obsegu pr ibližno 100 be sed in i majo obliko pr itožbe a li povabila. Besedila imajo na začetku in na koncu preprost pozdrav, sama vsebina pa je sestavljena iz stavkov in fraz, ki s o pog osta pr i pol judnem i n pr ostem i zražanju. Elektronska pisma namreč navadno ne vs ebujejo zapletenih strokovnih izrazov, preizkus pa je bil namenjen ugotavljanju p rimernosti r azpoznavalnika z a up orabo p ri narekovanju vsakdanjih preprostih nestrokovnih besedil.

Govorec je elektronska sporočila izgovarjal na dva načina. V prvem načinu je bilo izgovarjanje počasno in zelo razločno. V drugem primeru pa so bili isti primeri sporočil prebrani normalno hitro in nekaj manj razločno.

Pri t retjem p reizkusu smo uporabili š e t ri posnetke prostega go vora, kjer ni bilo branja b esedila, ampak si j e govorec b esedilo s proti iz mišljeval kar med s amim govorom. Tak način govora je bil neenakomeren in je imel precej prekinitev i n t udi ne kaj go vornih mašil oziroma medmetov. Vsebina govora se je nanašala večinoma na splošen opi s dn eva, k ot s o r azmere n a ces ti al i i zvedena opravila. B esedila s o bi la dolga ok oli 150 besed i n pr av tako k ot pr i br anju e lektronskih pisem niso vs ebovala

strokovnih al i žar gonskih b esed. R ezultat p reizkusa j e podan na Sliki 4.



Slika 4: Ugotovljena točnost razpoznavanja treh dodatnih vrst besedil.

Iz rezultatov je razvidno, da malo počasnejša i n razločna izgovarjava znatno pripomore k višji točnosti razpoznavanja in da v tem primeru postane Googlov razpoznavalnik govora že u poraben za m arsikatero aplikacijo

5 Zaključek

V članku so poda ni rezultati i zvedenih pr eizkusov Googlovega razpoznavalnika pri razpoznavanju govorjene slovenščine. Rezultati se odvisno od govorca in podatkovne z birke gibljejo n ekje od 40 d o 12 odstotne ocenjene napake razpoznavanja besed. Ta rezultat je nekaj slabši od rezultata približno 8 odstotkov napačno razpoznanih b esed, kot ga je za govorjeno angleščino objavil sam Google, v endar zar adi zel o obsežnega splošnega j ezikovnega modela i n z ahtevnih pos ebnosti slovenščine, ki otežuje zanesljivost samodejnega razpoznavanja, rezultat presega pričakovanja. Poleg tega pa se je izkazalo, da se rezultat sčasoma izboljšuje, saj je bil ugo tovljen napredek p ri p reizkusih na i stih p osnetkih, ki so bili izvedeni v različnih časovnih obdobjih. Ugotovljeni napredek pripisujemo dejstvu, da se akustične posnetke, k i se poš iljajo v obde lavo programskemu vmesniku v Googlovem računalniškem oblaku, zelo verjetno Google zbira in uporablja za prilagajanje in izboljševanje njegovih akustičnih in jezikovnih modelov.

6 Literatura

Ossama A bdel-Hamid, A bdel-Rahman M ohamed, H ui Jiang, L i D eng, G erald P enn in Dong Y u. 2014. Convolutional N eural N etworks for S peech Recognition. *IEEE/ACM Transactions On Audio, Speech, And Language Processing,* 22(10): 1533-1545.

Li D eng, Geoffrey Hinton, Brian K ingsbury. 2013. New types o f d eep n eural network l earning f or s peech recognition a nd r elated a pplications: An o verview. V: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP-2013*, str. 8599-8603, Vancouver, BC, Kanada. IEEE – Institute of Electrical and Electronics Engineers.

Li D eng. 2016. Industrial T echnology Advances: D eep learning --- from speech r ecognition to la nguage a nd multimodal pr ocessing. V : *APSIPA Transactions on Signal and Information Processing*, Cambridge University Press.

Simon Dobrišek, Boštjan Vesnicer, Jerneja Žganec G ros in France Mihelič. 2006. Uporaba kanoničnega govornega akustičnega modela za prilagajanje prostora govornih akustičnih značilk. V: *Tomaž Erjavec (ur.), Jerneja Žganec Gros (ur.). Jezikovne tehnologije : zbornik 9. mednarodne multikonference Informacijska družba IS 2006*, str. 89-92, Ljubljana, Slovenija. Institut Jožef Stefan.

Geoffrey Hinton, L i D eng, Dong Y u, G eorge E . D ahl, Abdel-rahman Mohamed, N avdeep J aitly, Andrew Senior, V incent V anhoucke, P atrick N guyen, T ara N . Sainath in Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in s peech r ecognition: The shared v iews o f four r esearch g roups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Biing-Hwang J uang, L awrence Rabiner. 2005. *Automatic speech recognition–a brief history of the technology development,* Georgia Institute of T echnology. Atlanta Rutgers University a nd t he University of C alifornia, Santa Barbara.

France Mi helič, Jerneja Žganec Gros, Simon Dobrišek, Janez Žibert, Nikola Pavešić. 2003. Spoken l anguage resources at LUKS o f t he University o f Ljubljana. *International Journal of Speech Technology*, 6(3): 221-232.

Travis P ayton. 2014. *Google Speech Api Information and Guidelines*. Google.

Suman Ravuri in Andreas Stolcke. 2015. Recurrent Neural Network an d LSTM Mo dels f or L exical U tterance Classification. V : *Proceedings of the Annual Conference of the International Speech Communication Association – INTERSPEECH 2015*, str. 1 35-139, Dresden, Germany. ISCA - International S peech Communication Association.

Tara N . S ainath, Abdel-rahman Mohamed, B rian Kingsbury in Bhuvana Ramabhadran. 2013. D eep Convolutional N eural N etworks for L VCSR. V : *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP-2013*, str. 8614-8618, Vancouver, BC, Kanada. IEEE – Institute of Electrical and Electronics Engineers.

Tara N . S ainath, O riol V inyals, A ndrew S enior in Hasim Sak. 2015. C onvolutional, Long S hort-Term M emory, Fully Connected D eep N eural N etworks. V : *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2015)*, s tr. 4 580-4584, Brisbane, A ustralija. IE EE – Institute of Electrical and Electronics Engineers.

Sabato Marco Siniscalchi, Dong Yu, Li Deng in Chin-Hui Lee. 2013. E xploiting de ep n eural n etworks for detection-based speech r ecognition. *Neurocomputing*, 106 (2013): 148–157.

Sabato Marco Siniscalchi, Dong Yu, Li Deng in Chin-Hui Lee. 2013 . Speech R ecognition U sing Long-Span Temporal P atterns i n a D eep N etwork M odel, *IEEE Signal Processing Letters*, 20 (3): 201-204.

Dong Y u in Li D eng. 2015. *Automatic Speech Recognition: A Deep Learning Approach*. S pringer-Verlag, London.

Janez Ž ibert i n F rance Mihelič. 2000. Govorna z birka vremenskih na povedi. V : *Tomaž Erjavec (ur.), Jerneja Žganec Gros (ur.). Jezikovne tehnologije : zbornik konference*, s tr. 1 08-111, Ljubljana, S lovenija. I nstitut Jožef Stefan.

Povezljivost pravopisnih pravil in slovarja: sanje pravopiscev 20. stoletja

Helena Dobrovoljc

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

Povzetek

Prispevek predstavlja proces oblikovanja novega pravopisnega priročnika, v katerem se povezujeta priprava pravopisnega slovarja in pravil. Slovaropisno orodje iLex, v katerem slovar izdelujemo, omogoča poleg priprave slovarskih sestavkov tudi zasnovo zbirke problemskih sklopov, prek katere se slovarski sestavki povezujejo s pravopisnimi pravili. Vsako pravilo je rezultat preučitve gradiva: (stara pravila, pripombe kritikov, jezikovna svetovalnica, korpusno gradivo). Uporabnik prek spleta lahko pregleduje slovar in se poveže s problemskim sklopom, po javni objavi novih pravil pa bo ob vsakem pravopisnem pravilu mogoče preveriti slovarski prikaz.

The Link Between Orthographic Rules and Dictionary Entries: 20th Century Orthographer's Dream

This paper presents the formation of the new Slovene orthographic guide which includes both a dictionary and orthographic rules. The dictionary is built in iLex, a lexicographic tool which, in addition to lexicographic work, also allows us to design a collection of orthographic categories through which the dictionary entries are connected to normative rules. Every rule is a result of relevant linguistic material examination (previous rules, reviews, language counselling, corpora). Language users can browse through the dictionary online and view the desired orthographic categories. After publication of the new rules, every rule will give the user the possibility to click on appropriate dictionary entries.

1 Slovenski pravopis kot priročnik

Pravopisni priročniki so se v raziskavah, ki so jih evropski leksikografi (Nerius, 1989: 1298) predstavili v drugi polovici 20. stoletja, uvrstil med najpogosteje uporabljene in najbolj razširjene jezikovne priročnike, s katerimi se srečuje laični uporabnik. Pri večini jezikov sestoji iz pravil in slovarja, pri čemer slovar gradivsko nadgrajuje in razširja pravila (Verovnik, 2004: 254). Medtem ko pravila predstavljajo bolj ali manj sistematičen nabor osnovnih zapisovalnih pravil na ravni glas – črka ter dogovornih norm, ki določajo rabo malih in velikih črk, pisanja skupaj ali narazen, status prevzetih jezikovnih prvin in interpunkcije, pa so slovarji lahko zelo različni.

Predvsem v slovanskem svetu in pri Francozih ter Nemcih in Dancih ter Švedih (Lorentzen, 2010) sta se oblikovali dve različni zasnovi: (a) slovarji, v katerih je poleg zapisa besede podana pravorečna, oblikoslovna in besedotvorna norma le sporadično; (b) slovarji, v katerih so vse besede pregibane in spregane, v posameznih primerih pa so predstavljene tudi pomenske uvrstitve, še zlasti, če gre za kontrastivne prikaze.

Slovenski pravopisni slovar je v 20. stoletju korenito spreminjal svoj značaj (prim. Dobrovoljc in Bizjak Končar, 2011). Če se osredinimo le na drugo polovico stoletja, pa je treba poudariti predvsem nezmerno širjenje obvestilnosti, ki se povezuje s pomanjkanjem enojezičnega razlagalnega slovarja. Tako je v slovarju iz leta 1962 mogoče najti tudi pomenske razlage, v pravopisnem slovarju iz leta 2001 pa so pomenske razlage zamenjala identifikacijske umestitve, razširila pa se je slovnična in normativna obvestilnost pri vrednotenju posebnih skladenjskih položajev in zvez, saj je občnoimenski del slovarja nastal kot normativno izostren derivat *Slovarja slovenskega knjižnega jezika* (dalje SSKJ).

Odzivi javnosti po izidu *Slovenskega pravopisa* 2001 (dalje SP 2001) niso bili usmerjeni le v kritiko vsebinskih rešitev, temveč predvsem v gradivsko zastarelost in posledično takšno vrednotenje gradiva, ki je odražalo neseznanjenost z dejansko jezikovno rabo in dojemanjem aktualne jezikovne razčlenjenosti v spremenjeni družbi na pragu 21. stoletja (Vidovič Muha, 2003). Precej nedoslednosti pa je pokazala tudi primerjava pravil in slovarja. Tu so kritiki (Kocjan Barle, 2002; Lenarčič, 2004; Bajt, 2002; Weiss, 2003; Šeruga Prek, 2002 idr.) opozarjali predvsem na neneenotne rešitve v okviru istih problemskih kategorij (npr. imena iz klasičnih jezikov: *Horacij* proti *Ovid*; neenotno uresničevanje preglasa v oblikoslovju in besedotvorju *trio*, *s triom – radio*, *z radiom*/*radiem*), v posameznih kategorijah pa tudi na razhajanja med pravili in slovarjem (npr. *volivec – volivec*/*volivec*), zlasti pri uzakonjanju dvojnic.

Elektronska (2003) in spletna (2010, 2014) izdaja tega pravopisa sta prinesli le odpravo tehničnih napak in vsebinskih lapsusov (prim. Dobrovoljc in Bizjak Končar, 2015: 36–37), razkorak med pravopisnimi pravili in pripadajočim slovarjem ter dejansko jezikovno prakso pa se je ob elektronski »hiperprodukciji« besedil in možnosti hitre javne objave še povečal.

Podobno kot po drugi svetovni vojni, ko se je s tehnološko in medijsko revolucijo »zgodila« ekspanzija nove leksike in so predvojni jezikoslovci, učenci Breznikove šole, »le strahoma sprejemali ugotovitve modernih smeri v obravnavanju sodobnega knjižnega jezika, boječ se anarhije« (Bajec, 1968: 72), je po letu 2000 postalo očitno, da se je z elektronsko revolucijo zaradi nepredvidljivih sprememb v doslej razmeroma obvladljiv jezikovni sistem, kakršnega so podajali temeljni priročniki – SSKJ, Toporišičeva slovnica (1976, 2000) in akademijski pravopis – zamajal. Za uspešno in učinkovito stabiliziranje

tega tristebrnega sistema je pomembno, da načrtujemo sočasno prenovo vseh treh inherentno povezanih priročnikov, hkrati pa izdelamo koncept prenove vsakega posebej.

V prispevku predstavljamo vlogo digitalnega okolja pri načrtovanju prenove pravopisnega priročnika.

2 Namen članka

Medtem ko je večina pravopisnih pravil nastajala vzporedno s slovarjem (1899, 1920, 1935, 1937, 1950, 1962), so pravila zadnjega pravopisa (1990) izšla desetletje pred slovarjem (2001), ko so bila le gradivsko in v malenkostih prilagojena slovarju. V tem, tj. zadnjem pravopisu je bodisi zaradi časovne distance med nastajanjem pravil in slovarja bodisi zaradi deloma spremenjene avtorske skupine pri pripravi enega ali drugega dela pravopisa pojavljalo tudi nezaželeno razhajanje med pravili in slovarjem. Pojavu, ki ga imenujemo asinhrona kodifikacija, bi se bilo mogoče izogniti s povezovanjem organizacijskih in izvedbenih faz v kodifikacijskem procesu.

Zato želimo v nadaljevanju predstaviti, da in kako je mogoče v digitalnem okolju preseči že omenjeno neskladje.

3 Sprememba koncepta priročnika v elektronski dobi

Po odločitvi, da je namesto skrajšane in prirejene različice SP 2001 na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU (dalje ISJFR) pripravimo raje novo zasnovo priročnika (Dobrovoljc in Jakop 2008), v katerem se bomo usmerili v prikaz pravopisno, oblikoslovno, besedotvorno relevantnih informacij, brez normativnih oz. slogovnih napotkov, ki niso preverjeni v sodobni rabi in z rabo že modificiranem jezikovnem sistemu, ter zvrstnega kvalificiranja, je bila dejavnost usmerjena v ugotavljanje, katere so za uporabnika pomembne informacije v pravopisu oziroma katere so normativne zadrege, na katere pričakuje konkretne odgovore. V nadaljevanju pa tudi v spremenjeno zasnovo, ki temelji na izrabi spletnih možnosti.

Raziskave jezikovne rabe so metodološko raznovrstne, v pričujočem prispevku pa predstavljamo povezavo med zbirko uporabniških vprašanj in odgovorov ter temeljnimi normativnimi vprašanji, na katere moramo jezikoslovci odgovoriti v sklopu evidentiranja oz. opisovanja jezikovne rabe. Z učinkovito izrabo vprašanj, ki jih uporabniki zastavljajo na spletu v (**3.1.1**) *Jezikovni svetovalnici*,[1] lahko dopolnjujemo obe spletno dostopni pravopisni zbirki: (**3.2.2**) pravopisni slovar,[2] ki ga objavljamo sukcesivno na portalu Fran pod imenom rastoči *Slovar pravopisnih težav* (2014–) in izdelujemo vzporedno s pravopisnimi pravili; (**3.2.1**) zbirko problemskih sklopov,[3] objavljeno na portalu Fran pod imenom *Pravopisne kategorije Slovarja pravopisnih težav* (2015–), v katerih so opisane okoliščine priprave slovarja v povezavi s pravili in preteklo kodifikacijo.

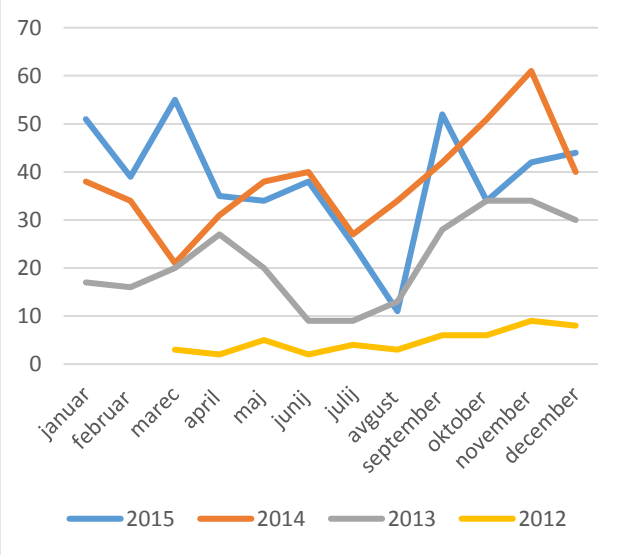
Ločeno nastaja tudi zbirka pravopisnih pravil, ki trenutno še ni objavljena na spletu.

3.1 Ugotavljanje normativnih zadreg

Veliko normativnih zadreg je bilo ugotovljenih že z evidentiranjem jezikovnih dvojnic (pisnih, oblikoslovnih, besedotvornih in skladenjskih) v okviru knjižnega jezika, ki jih prikazujejo zlasti slovarji (npr. *gredo*/*grejo*; *z Miho*/*Mihom*) in gradivsko potrjujejo besedilni korpusi, nov izziv pa ponuja jezikovni razvoj, saj se precej jezikovnih inovacij širi v jezikovni sistem iz pogovornega jezika oziroma drugih nestandardiziranih zvrsti (nadregionalni pogovorni sistem; spletni jezik). Pri tem uporabljamo različne načine – tudi razčlenitev kritik obstoječega pravopisa in preverbo aktualne variantnosti po korpusih in drugem gradivu, kar sodi med vire, ki so neodvisni od aktivnosti uporabnika, saj spremljajo njegove jezikovne izbire brez uporabnikovega angažmaja.

3.1.1 Jezikovno svetovanje

Jezikovno svetovanje spada med tiste avtoreflesivne vire normativnih zadreg (Dobrovoljc in Krek 2011: 43), kjer je uporabnik aktiven pri iskanju rešitve za svojo težavo, saj je spraševalec in pobudnik. Na ISJFR že od leta 2012 deluje spletna *Jezikovna svetovalnica* za jezik, prek katere se uporabniku ni treba identificirati, objavljenih pa je že približno 1400 vprašanj in odgovorov.



Slika 1: Število obravnavanih vprašanj v jezikovni svetovalnici v letih 2012–2015.

Začetna dinamika obravnavanih vprašanj (do 10 vprašanj mesečno) se je v letu 2013 povečala na povprečno 20 (prikaz na sliki 1). V vseh letih opažamo počitniški »upad« v juliju in zlasti avgustu. V letu 2015 (april) s eje število zastavljenih vprašanj ustalilo pri 40 mesečno, saj smo bili prisiljeni odzivni čas podaljšati na tri tedne od prejema odgovora, uvedli pa smo tudi uredniški odbor, ki se mora z odgovorom večinsko strinjati. Načeloma odgovarjamo konsenzualno, odzive na odgovore ali dopolnitve odgovorov pa objavljamo pod odgovorom post festum z opozorilom, da gre za dodatek.

Pri objavi odgovora vprašanja kategoriziramo v tristopenjskem sistemu:

[1] URL: http://isjfr.zrc-sazu.si/sl/svetovalnica

[2] URL: http://www.fran.si/135/spt-slovar-pravopisnih-tezav

[3] URL: http://www.fran.si/spt-kategorije

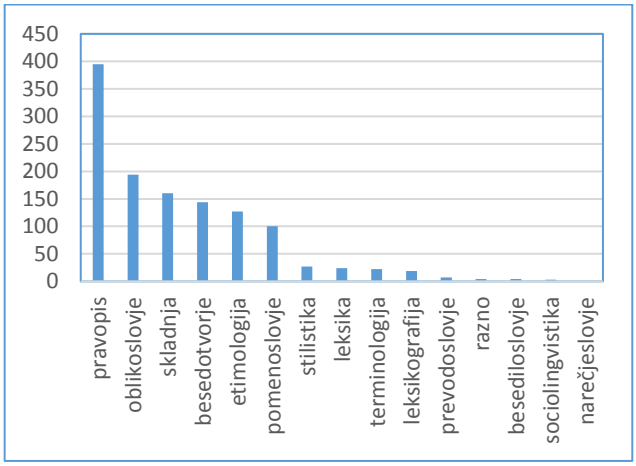
(a) področje (eno do dve),
(b) jezikovna kategorija (lahko jih je več),
(c) zgled (konkretna oblika/beseda/zveza, ki zanima uporabnika, ki je pri nekaterih kategorijah tudi ne navajamo, npr. pri vprašanjih ujemanja, vejice in podobno).

Ponazoritev tristopenjskega sistema navajanja ključnih besed (a – b – c):

- OBLIKOSLOVJE – sklanjanje lastnih imen – *Vike*
- PRAVOPIS – prevzete besede in besedne zveze – *standup*
- PRAVOPIS – ločila: vejica
- PRAVOPIS – mala/velika začetnica – imena jedi
- GLASOSLOVJE – naglas – *cveten*

Tako področij kot jezikovnih kategorij ob posameznem vprašanju navadno ni več kot tri, npr.:

- BESEDOTVORJE in OBLIKOSLOVJE – svojilni pridevnik, sklanjanje lastnih imen – Itten
- PRAVOPIS – navajanje letnic in desetletij, pisanje števk z besedami
- SKLADNJA – ujemanje, besedni red
- POMENOSLOVJE – pomenska razlaga – *razlog*, *vzrok*
- LEKSIKOGRAFIJA – vključitev besede v slovar,[4] pomenska razlaga – *izčlanitev*



Slika 2: Zastopanost vprašanj po področjih v letih 2012–2015.

Številčna zastopanost vprašanj po področjih (ročno prešteto, gl. sliko 2) je pričakovana: največ je vprašanj s področja pravopisa (raba začetnice – 98 vprašanj, pisanje skupaj in narazen – 14 vprašanj, vejica (ločila) – 57 vprašanj, druženje črk in števk, prevzemanje – 57 vprašanj), sledijo naslednja področja s pestro notranjo členjenostjo, npr.:

- OBLIKOSLOVJE (sklanjanje lastnih imen – 143 vprašanj, določni in nedoločni pridevnik, oblikoslovna dvojnica)
- SKLADNJA (ujemanje – 28 vprašanj, vikanje, vezljivost, trpnik, besedni red)
- BESEDOTVORJE (zloženke, priredne zloženke, sestavljenke, prebivalska imena)

Podoben, a zaključen nabor jezikovnih zadreg (s končnimi 620 enotami) je bil pripravljen leta 2010 v okviru projekta »Sporazumevanje v slovenskem jeziku« na osnovi 2500 vprašanj iz različnih spletnih svetovalnic (ŠUSS – Študentska skrb za slovenščino) oz. forumov (med.over.net) iz let 2001–2005 (Bizjak Končar idr. 2011; Krek idr. 2013). Gre za sistematično tematsko urejen večstopenjski nabor vprašanj s primeri, npr.:

- OBLIKOSLOVJE – samostalniki / moški samostalniki / samostalniki na samoglasnik / imena na -y – *Broadway Broadwayja* ali *Broadwaya*, *Disney Disneya* ali *Disneyja*, *Sarkozy – Sarkozyja* ali *Sarkozya*.

Pri prenovi pravopisnih pravil skušamo s pomočjo lastnega izpopolnjenega nabora, ki ga gradimo, npr. zgoraj navedene kategorije imen na *-y*, dopolniti z novimi zgledi in kategorijami iz jezikovne svetovalnice, iz katerih je mogoče izluščiti naslednje skupine, relevantne za prenovo pravopisnih pravil:
(I) imena, ki ne podaljšaujejo osnove z *j* zaradi izglasnega j: *Broadway – Broadwaya* [bródvej bródveja];
(II) imena, ki podaljšujejo osnovo z j: *Disney – Disneyja* [dízni díznija]
(III) imena s pridevniško osnovo na [ski]: slovanska, po pridevniški sklanjatvi *Dobrovsky – Dobrovskega*
(IV) imena na [ski]: neslovanska *McClosky – McCloskyja*.

Jezikovna svetovalnica je zaprt spletni sistem, vprašanj pa so dosegljiva tudi prek slovarskega portala Fran. V letu 2016 bo svetovalnica prenovljena in nadgrajena, saj ima trenutno značaj navadne spletne strani in ne omogoča uspešnega iskanja po zbirki že rešenih vprašanj, ki jo dobimo v formatu .html, poleg tega uporabniku ne nudi dovolj velike podpore pri zastavljanju vprašanj, kar povzroča tudi njihovo ponavljanje. Načrtujemo oblikovanje sistema tipskih vprašanj, ki bi po želji olajšal tako zastavljanje vprašanj kot tudi svetovalno delo; aplikacija bo na željo uporabnika tega prek vstopne točke usmerila na iskano problematiko (npr. kako besedo zapišemo, kako jo sklanjamo ipd.).

Načrtna in sistematična izraba naborov uporabniških vprašanj je inovacija v slovenskem normativnem jezikoslovju. Medtem ko je prenova dosedanjih pravopisnih pravil izhajala iz praktičnih izkušenj sodelujočih jezikoslovcev ali paberkovalnega nabiranja po normativnih priročnikih preteklih obdobij, danes poteka prek sistematičnih raziskav uporabniških zadreg, kakor jih prikazujejo jezikovne svetovalnice v obdobju 2000–2016 in seveda jezikovna raba v gradivskih korpusih ter drugih virih.

3.2 Pravopisna pravila in slovar

Pri posodobitvi pravopisnih pravil in slovarja izhajamo iz obstoječega fonda pravil, ki ga preverjamo glede na že evidentirana odstopanja od rabe (Dobrovoljc in Jakop, 2011), kritike SP 2001 (gl. zgoraj) in empirično pridobljene nabore normativnih in slovničnih zadreg oz. jezikovnih vprašanj. Predloge za spremembo, posodobitev oziroma dopolnitev pravopisnih pravil pripravlja pravopisna

[4] Mišljena so vprašanja, zakaj besede ni v slovarju ali v katerem slovarju bi našli besedo. S portalom Fran (www.fran.si), ki je zbirna točka vseh digitaliziranih ali za digitalno okolje pripravljenih slovarjev, se je število tovrstnih vprašanj zmanjšalo.

skupina na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU v povezavi s Pravopisno komisijo pri SAZU in ZRC SAZU.

Tako sta bili npr. do junija 2016 posodobljeni poglavji o pisnih znamenjih in rabi male in velike začetnice. V okviru posodobitve so bila strukturno preurejena poglavja glede na možnost spletne objave (vsak sklop pravila uvaja naslov), dodana pa so bila poglavja, ki so se glede na nabor normativnih zadreg iz jezikovne svetovalnice ali kot odraz neustaljene rabe izkazala kot nezadovoljivo rešena.

Ob vsakem pravilu je pripravljen širši geslovnik iztočnic pravopisnega slovarja, ki bodo v spletni različici pravil dosegljiva s klikom, v knjižni različici pa bodo slovarske sestavke spremljale posebne številske oznake. Vsak problemski sklop *Slovarja pravopisnih težav* je zaokrožena celota, ki jo opisuje zbirka *Problemski sklopi Slovarja pravopisnih težav*.

Predstavitev metodologije procesa prenove pravil je bila podana v Dobrovoljc in Verovnik (2015), v pričujočem prispevku pa bomo izpostavili predvsem povezljivost slovarja z opisnimi podatki o nastajanju slovarske informacije.

3.2.1 Pravopisni slovar: *Slovar pravopisnih težav* (2014–)

Slovar pravopisnih težav (dalje SPT) je rastoči slovar, ki ga od leta 2014 objavljamo na spletnem slovarskem portalu Fran. Organizacijsko slovar nastaja ob prenovi pravopisnih pravil, in sicer je zasnovan primarno za digitalno okolje.

Pri zasnovi pravopisnega slovarja smo upoštevali možnost neokrajšanega podajanja slovarskih metapodatkov (npr. m = samostalnik moškega spola), kar slovar oddaljuje od klasičnega knjižnega koncepta. Da bi presegli omejitve krajšanega slovarskega zapisa, smo se odločili za izrecno poimenovanje posameznih elementov slovarske zgradbe, ki so v spletni različici prikazani kot posebni uvajalni elementi v slovarskem sestavku, ki ga gradijo štiri osnovne enote:

1. slovnični podatki (= besednovrstna umestitev in pravorečni podatki)
2. oblikoslovni podatki (= pregibnostno-naglasni vzorci za vsa števila in spole)
3. zgledi in opozorila (= izbrani podatki, ki izhajajo iz normativnih zadreg jezikovnih uporabnikov; normativna opozorila; ponazorila rabe v neznačilnih položajih ipd., ki so bili v SP 2001 pri večini iztočnic, pri lastnih imenih pa sistematično izpuščeni)
4. povezane iztočnice (= povezava z iztočnicami, ki so bodisi normativno bodisi stvarno ustreznejše, dodajanje za pravopisje pomembnih podatkov o besedotvorni motiviranosti posameznih besed (npr. *ničejanstvo* < *Nietzsche*).

Slovarske rešitve so do potrditve v pravopisni komisiji označene z oznako »predlog«.

Pri oblikovanju notranje strukture slovarja smo upoštevali tudi ugotovitve novejših raziskav, ki poudarjajo, da sodobni uporabnik v digitalnem svetu »ne ločuje več strogo med različnimi podatkovnimi viri« (Gorjanc, 2013). Pri tem se zlasti v specializiranem slovaropisju hitro približujemo opaznim konceptualnim spremembam, ki kažejo na hibridni značaj sodobnih slovarjev. Ti so prilagojeni uporabniškim zahtevam (Bergenholtz in Tarp, 1995) in verjetno zato (raziskava te povezave še ni bila opravljena) vse pogosteje prinašajo tudi enciklopedične informacije in semantične podatke (Lorentzen, 2010: 666).

Problemski pristop omogoča po eni strani postopno dodajanje novih redakcij na splet, po drugi strani pa zahteva dinamičen pristop k slovarskemu konceptu. Posamezne iztočnice SPT se bodo povezovale z različnimi problemskimi sklopi oziroma v kasnejši fazi – z različnimi pravopisnimi pravili.

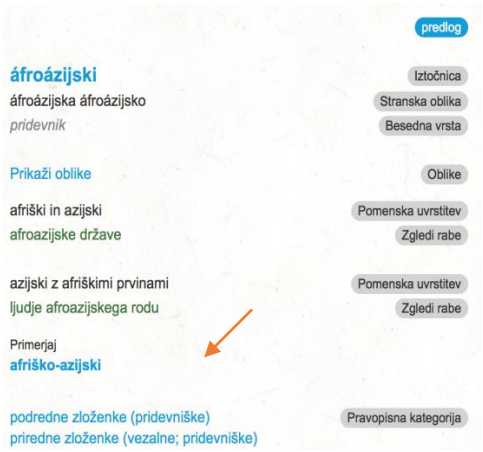
Slovarske sestavke pripravljamo v slovaropisnem orodju iLex, ki omogoča poleg priprave slovarskih sestavkov tudi zasnovo zbirke problemskih sklopov, prek katere se slovarski sestavki povezujejo s pravopisnimi pravili. Ker je ob eni iztočnici mogoče predvideti tudi več problemskih sklopov, smo v okviru iLexa zasnovali dodatne atribute. Tako se npr. moško ime *Ivo* povezuje z dvema problemskima sklopoma: (1) samostalnik moškega spola / pogovorna podaljšava s *-t*; (2) lastna imena / moška. Samostalnik *škrat* je trenutno povezan s tremi problemskimi sklopi: (1) poimenovanja bajeslovnih bitij; (2) samostalnik moškega spola / dvojnica v imenovalniku množine (*škrati* in *škratje*); (3) domišljijska imena / v stalnih besednih zvezah (*škrat Kuzma*) (gl. sliko 3).

Slika 3: Prikaz vpisa v iLex: iztočnica *škrat.*

Zaradi nepredvidljivosti števila problemskih sklopov, ki lahko nastopajo ob eni iztočnici, smo se odločili za možnost neomejenega dodajanja polj, v okviru katerih je dvojna hierarhija: odločitev za osnovni in neosnovni problemski sklop in odločitev za višji in nižji nivo v okviru problemskega sklopa: npr. domišljijsko ime (= višji nivo) in ime v stalni besedni zvezi (= nižji nivo).

Ob pregledovanju slovarskih redakcij na portalu Fran je mogoče preverjati tudi povezanost s problemskimi sklopi. V SPT na Franu trenutno prenavljamo obliko in notranjo organiziranost redakcij, kljub temu pa je tudi v delovni različici mogoče s klikom na enote, ki jih uvaja razdelek »Pravopisna kategorija« (slika 4), priklicati opis

problemskega sklopa iz zbirke *Problemski sklopi Slovarja pravopisnih težav*.



Slika 4: Pravopisne kategorije v *Slovarju pravopisnih težav*.

3.2.2 Problemski sklopi *Slovarja pravopisnih težav* (2015–)

Za premostitev časovnega zamika in boljšo obveščenost uporabnikov glede vsebinskih sprememb v procesu kodifikacije je bila zasnovana zbirka *Pravopisne kategorije: Problemski sklopi Slovarja pravopisnih težav*. Zbirka nastaja vzporedno ob redigiranju slovarskih sestavkov SPT (2014–), in sicer vsak problemski sklop prinaša podatke:

1. o številu iztočnic, vključenih v izbrani problemski sklop, in o kriterijih za njihov izbor;
2. o novostih in spremembah glede na pravopisna pravila in slovar v SP 2001, npr. podatke o slovničnih posebnostih, ki jih prinaša raba iztočnice v specifični skladenjski rabi; o stičnosti in sklonljivosti – če se ta razlikuje od lastnosti, pripisanih iztočnici;
3. o normativnih spremembah glede na SP 2001 (npr. *čehoslovaški* : *češko-slovaški*);
4. sklic na pravilo v SP 2001, ki obravnava izbrano vprašanje;
5. datum zadnje spremembe.

V letu ustanovitve zbirke (2015) smo vanjo vključili 16 problemskih sklopov, med katerimi so nekateri še podrobneje členjeni (okrajšave, simboli, psevdonimi ...), in ti problemski sklopi vključujejo 1580 slovarskih sestavkov. Skupaj z opisi bo po prenovi poleti 2016 posamezna enota zbirke (torej pravopisna kategorija) omogočala tudi pregled vseh slovarskih sestavkov, ki so bili pripravljeni za ponazoritev sklopa.

4 Sklep

Tehnološki napredek v elektronski dobi nam omogoča ugotavljanje aktualnih normativnih zadreg jezikovnih uporabnikov (Dobrovoljc in Krek, 2011) ter pridobivanje nabora problematičnih mest oziroma vrzeli v jezikovnem sistemu in tudi standardizacijskih priročnikih s pomočjo spletnega svetovalnega mesta. Tako pridobljenega nabora vprašanj ni mogoče uporabiti le kot ogrodja spletnega priročnika za laične uporabnike, temveč tudi usmerjevalo pri prenovi aktualnega pravopisa temeljne referenčne točke za vsa izvedena dela normativnega značaja.

Elementi povezovanja »pravopisno pravilo – slovarski sestavek – problemski sklop« so bili v prispevku predstavljeni kot stopnje kodifikacijskega postopka in kot samostojne entitete. Njihova povezljivost je presegla problem slovenskih pravopiscev več generacij: vse od izida prvega slovenskega pravopisnega priročnika, ki ga je za izdajo v letu 1899 pripravil Fran Levec, so pravila in slovar ostajala v odvisnem razmerju, praktično pa jih ni bilo mogoče povezati. To sta na začetku 21. stoletja omogočila šele elektronski medij in splet.

5 Literatura

Anton Bajec, 1968: Slovenski knjižni jezik. *Jezik in slovstvo* 13. Št. 3. Str. 69–74.

Drago Bajt, 2002: Zadnji pravopis – poslednji pravopis?. *Nova revija 21*. Forum. Št. 239/240 (mar.–apr. 2002). Str. 2–10.

Henning Bergenholtz, Sven Tarp (ur.), 1995: *Manual of Specialised Lexicography: The preparation of specialised dictionaries*. John Benjamins Publishing.

Aleksandra Bizjak Končar, Helena Dobrovoljc, Kaja Dobrovoljc, Nataša Logar, Polonca Kocjančič, Simon Krek, Tadeja Rozman, 2011: *Slogovni priročnik: sporazumevanje v slovenskem jeziku : kazalnik 17 – Standard za korpusno analizo težav pri tvorbi besedil*. http://www.slovenscina.eu/Media/Kazalniki/Kazalnik17/Kazalnik_17_Slogovni_prirocnik_SSJ.pdf.

Helena Dobrovoljc in Aleksandra Bizjak Končar, 2011: Vprašanja obvestilnosti sodobnega pravopisnega slovarja. V: Jesenšek, Marko (ur.). *Izzivi sodobnega slovenskega slovaropisja*, (Mednarodna knjižna zbirka Zora, 75). Maribor: Mednarodna založba Oddelka za slovanske jezike in književnosti, Filozofska fakulteta, 2011. Str. 86–109.

Helena Dobrovoljc in Nataša Jakop, 2011: *Sodobni pravopisni priročnik med normo in predpisom*. Založba ZRC.

Helena Dobrovoljc in Simon Krek, 2011: Normativne zadrege – empirični pristop. V: Kranjc, Simona (ur.). *Meddisciplinarnost v slovenistiki*, (Obdobja 30). Str. 89–97. Znanstvena založba Filozofske fakultete, Ljubljana.

Helena Dobrovoljc in Aleksandra Bizjak Končar, 2015: Pravopisno slovaropisje na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU. *Slavia Centralis* 8. Št. 1. Str. 34–50.

Helena Dobrovoljc in Tina Lengar Verovnik, 2015: Slovar pravopisnih težav kot sopotnik pravopisnih pravil in pomenskorazlagalnega slovarja. V: M. Smolej (ur.). *Slovnica in slovar – aktualni jezikovni opis* (Obdobja 34). Str. 163–171. Znanstvena založba Filozofske fakultete, Ljubljana.

Vojko Gorjanc, 2013: Slovar slovenskega jezika v digitalni dobi. *Posvet o novem slovarju slovenskega jezika*, Ministrstvo za kulturo 12. 2. 2014 http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/E_zbornik/1-_Vojko_Gorjanc_-_Slovar_MK_tekst_FINAL.pdf

Marta Kocjan - Barle, 2002: Slovenski pravopis 2001 med znanostjo in (ne)uporabnostjo. *Nova revija 21*. Forum. Št. 239/240. Str. 11–25

Simon Krek, Helena Dobrovoljc, Kaja Dobrovoljc, Damjan

Popič, 2013: Online style guide for Slovene as a language resources hub. V: Kosem, Iztok (ur.), *Electronic lexicography in the 21st century : thinking outside the paper : proceedings of eLex 2013 Conference. Tallinn, Estonia*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Tallinn: Eesti Keele Instituut, 2013. Str. 379–391. http://eki.ee/elex2013/proceedings/eLex2013_26_Krek+etal.pdf.

Simon Lenarčič, 2004: *Popravopis: kaj je narobe in kaj manjka v novem Slovenskem pravopisu?* Ljubljana: Samozaložba.

Henrik Lorentzen, 2004: Orthographical Dictionaries: How Much Can You Expect? The Danish Spelling Dictionary Revis(it)ed. Proceedings of the XIV Euralex International Congress (Leeuwarden, 6-10 July 2010). Ur. Anne Dykstra in Tanneke Schoonheim. Fryske Akademy. Str. 664–670

Dieter Nerius, 1989: Das Orthographiewörterbuch. *Wörterbücher. Ein internationales Handbuch zur Lexikographie.* Zweiter Teilband. Berlin, New York: Walter de Gruyter. Str. 1297–1304.

SSKJ 1970–1991 = *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.

SP 2001 (^{2}2003, ^{E}2014) = *Slovenski pravopis.* Ljubljana: SAZU – ZRC SAZU – Založba ZRC.

Cvetka Šeruga Prek, 2002: Izgovorni in naglasni problemi Slovenskega pravopisa 2001. Nova revija 21. Forum. Št. 239/240. Str. 26–30.

Jože Toporišič, 2000: *Slovenska slovnica*. Maribor: Založba Obzorja.

Tina Verovnik, 2004: Norma knjižne slovenščine med kodifikacijo in jezikovno rabo v obdobju 1950–2001. *Družboslovne razprave*, 2004, letn. 20. Str. 241–258.

Ada Vidovič Muha, 2003: Kaj je novega v knjižnem jeziku? – Ob izidu Slovenskega pravopisa. *SRL.* Str. 117–122.

Peter Weiss, 2003: Slovenski pravopis 2003 – priročnik na stranpoteh slovenskega jezika. Okrogla miza: Aktualna vprašanja ob novem slovenskem pravopisu. Perspektive slovenistike ob vključevanju v Evropsko zvezo. *Zbornik Slavističnega društva Slovenije*. Ur. Marko Jesenšek. Ljubljana: Slavistično društvo Slovenije. Str. 201–206.

Slovenska akademska besedila: prototipni korpus in načrt analiz

Tomaž Erjavec,[*] Darja Fišer,[†*] Nikola Ljubešić,[*♣] Nataša Logar,[‡] Milan Ojsteršek[♠]

* Odsek za tehnologije znanja, Institut »Jožef Stefan«, Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

† Oddelek za prevajalstvo, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si

♣ Filozofska fakulteta, Univerza v Zagrebu, Ivana Lučića 3, 10000 Zagreb
nikola.ljubesic@ffzg.hr

‡ Fakulteta za družbene vede, Univerza v Ljubljani, Kardeljeva ploščad 5, 1000 Ljubljana
natasa.logar@fdv.uni-lj.si

♠ Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru, Smetanova ulica 17, 2000 Maribor
milan.ojstersek@um.si

Povzetek

Razvitost jezika, ki se rabi v akademskem okolju, je pomemben kazalnik njegove vitalnosti. V prispevku prikažemo izdelavo sodobnega referenčnega vira za ta del slovenščine in podamo okvirni nabor nadaljnjih na njem osnovanih raziskav. Prototipni korpus KAS vsebuje besedila, zajeta z Nacionalnega portala odprte znanosti, ter vključuje več kot 50.000 znanstvenih in strokovnih besedil s preko milijardo pojavnic. Njegov nastanek je med drugim vključeval postopke zajema, filtriranja, čiščenja in jezikoslovnega označevanja, na njem osnovane raziskave pa bodo dale rezultate na področju klasifikacije besedil, razvoja orodij in zbirk za terminološko delo ter opisa tega dela sodobnega slovenskega jezika.

Slovene Academic Texts: Prototype Corpus and Research Plan

The development of the academic part of any language is an important indicator of its vitality. The paper presents the construction of a contemporary academic language resource for Slovene and provides a framework for further research based on it. The KAS prototype corpus contains texts harvested from the Open Science portal of Slovenia and contains about 50,000 scientific texts with over one billion tokens. Its compilation included the collection, filtering, cleaning and linguistic annotation of its texts, while KAS corpus research will give results in the fields of text classification, terminological tool and database development and in description of contemporary academic Slovene.

1 Uvod

Razvoj in uporaba slovenskega jezika v visokem šolstvu ter znanosti je zadnja leta eno osrednjih vprašanj slovenske jezikovne politike (Kalin Golob et al., 2014; gl. še druge vire v Logar, 2013a: 247). Problem nepripravljenosti slovenskega jezika za digitalno dobo na več mestih izpostavlja tudi *Resolucija o Nacionalnem programu za jezikovno politiko 2014–2018*, še izraziteje pa iz resolucije izhajajoča *Akcijski načrt za jezikovno izobraževanje* in *Akcijski načrt za jezikovno opremljenost* (oba 2015). Namen akcijskih načrtov je bil konkretizirati »izzive, ki so potrebni hitrega in učinkovitega ukrepanja«. V prvem akcijskem načrtu sta tako kot dva od štirih ciljev v zvezi s slovenščino v visokem šolstvu in znanosti navedena prav »razvijanje sporazumevalne zmožnosti v /slovenskem/ strokovnem jeziku« ter »izboljšanje položaja slovenščine kot jezika znanosti«. V drugem akcijskem načrtu je na slovenščino kot jezik znanosti vezanih 8 izmed 47 ciljev.

Leta 2016 se je začel izvajati triletni temeljni raziskovalni projekt »Slovenska znanstvena besedila: viri in opis«. Namen projekta je uresničiti del zgornjih izzivov, podatkovni temelj zanj pa predstavlja obsežen korpus pisnih besedil akademske slovenščine.[1] V prispevku bomo predstavili, kako je potekalo pridobivanje besedil skupaj z metapodatki, kako so bila besedila pretvorjena in kakšna sta korpusov zapis ter trenutna zgradba. Predstavitvi korpusa bo v drugem delu prispevka sledil še kratek oris ključnih korpusnih analiz, ki bodo zajemale klasifikacijo besedil, luščenje terminologij in izgradnjo terminološke baze ter jezikovnoopisne študije.

2 Nacionalni portal odprte znanosti

V zadnjem času so slovenske univerze in druge raziskovalne institucije začele vzpostavljati digitalne zbirke svojih publikacij, ki vsebujejo raznorodna besedila, od diplomskih, magistrskih in doktorskih del do znanstvenih ter strokovnih prispevkov. Pomemben mejnik je pri tem leta 2013 vzpostavljeni Nacionalni portal odprte znanosti, ki agregira vsebine iz repozitorijev slovenskih univerz, slovenskih raziskovalnih organizacij in drugih zbirk (dLib, DKMORS, VideoLectures.NET, repozitorij ScieVie, CLARIN.SI, arhiv ADP) za potrebe skupnega iskalnika, priporočilnega sistema in detektorja podobnih vsebin (Ojsteršek et al., 2014). Repozitoriji omogočajo izvoz metapodatkov v imenike odprtega dostopa (OpenDOAR, ROAR, BASE, DART-Europe itd.), Google Scholar in OpenAIRE. Nacionalni portal in repozitoriji so povezani s slovenskim bibliografskim katalogom COBISS.SI. Če je vir iz nacionalne infrastrukture zaveden v COBISS, se njegovi metapodatki dopolnijo z metapodatki, ki so jih knjižničarji vnesli v COBISS. Na ta način se bistveno izboljša kakovost metapodatkov vstavljenih gradiv. Portal že ponuja dostop do prek 124.000 slovenskih objav s širokega nabora strokovnih

[1] Poimenovanje akademska slovenščina kot nadpomenka za zelo različne žanre, ki nastajajo v akademskem okolju (od izvirnih znanstvenih člankov do študentskih poročil), izhaja iz žanrske teorije (npr. Bhatia, 1993; Swales, 2000). Ker bo več nadaljnjih analiz v projektu teoretično in metodološko izhajalo iz tega pristopa, smo se pri poimenovanju korpusa odločili za ta termin in ga posledično večkrat uporabljamo tudi v tukajšnjem nadaljnjem besedilu, gre pa za – če uporabimo v slovenskem prostoru bolj razširjeno poimenovanje iz zvrstne teorije – strokovno-znanstvena besedila.

področij. Ta dela so izjemno dragocen, a zaenkrat še pomanjkljivo izkoriščen vir podatkov o akademski slovenščini, kot tudi bogat vir terminologije.

3 KAS-proto

Na osnovi gradiv iz podatkovne baze Nacionalnega portala odprte znanosti smo v začetku leta 2016 izdelali prvo različico korpusa slovenskih akademskih besedil, korpus KAS-proto.

3.1 Izvoz podatkov za prototipni korpus

Podatkovna baza, ki smo jo izvozili, je za vsako besedilo obsegala metapodatke, datoteko z izvornim formatom besedila in iz njega izluščeno besedilo. Omejili smo se samo na del metapodatkov (naslov, avtorji, povzetek, univerza, fakulteta, ključne besede, leto nastanka gradiva, vrstilci UDK, COBISS id vira in podatki, ki so potrebni za citiranje vira). Za preslikavo numerično zapisanega UDK v ključne besede področij smo uporabili odprte povezane podatke konzorcija UDK.[2]

3.2 Pretvorba korpusa

Pri pretvorbi zajetih podatkov smo v KAS-proto vključili samo besedila, ki so imela:

- izvorni zapis v PDF, saj je ključno, da je poleg besedila v korpusu raziskovalcem dostopen tudi vpogled v izvirnik, ki poleg besedila ponuja tudi njegovo oblikovanje in slike;
- s tehničnega vidika razmeroma kakovostno besedilo, saj PDF ni idealen format za luščenje besedila in je veliko besedil pokvarjenih do te mere, da so za raziskave neuporabna;
- pripisane vsaj minimalne metapodatke (podatki o avtorjih, naslovu, letnici, univerzi in fakulteti ter zvrsti po tipologiji COBISS), saj je brez teh podatkov nemogoče korektno citiranje, otežene pa so tudi analize korpusa;
- leto izdaje 2000 ali mlajše, saj je bilo starejših besedil zelo malo, zaradi česar korpus ne bi bil reprezentativen za starejša obdobja;
- dovolj velik delež slovenskega besedila, saj so bila v izvozu tudi besedila, ki so v celoti ali pretežno v angleščini.

Po filtriranju smo korpus pretvorili v format XML po shemi, ki smo jo izdelali zanj. Pretvorba je vsebovala naslednje korake:

- popravljanje najpogostsejših napak kodiranja znakov, saj ima veliko avtomatsko izluščenih besedil sistematične napake v kodiranju (*ž* je npr. zapisan kot *ţ*, *µz*, *z˘*, *œ*, *ſ* itd.);
- brisanje slabih znakov, ki bodisi niso veljaven UTF-8 ali pa so v področju zasebne uporabe (PUA);
- odstranjevanje glav in nog strani, ki vsebujejo številko strani, naslov dela, fakulteto itd. in bi sicer zelo izkrivili jezikovno podobo besedil;
- hevristično določanje mej med odstavki, saj so ti osnovna enota diskurza, so pa v neposredno izluščenem besedilu pogosto napačno identificirani;
- odstranjevanje odstavkov, ki so bodisi prazni ali vsebujejo samo ločila;
- avtomatsko določanje jezika posameznega odstavka, bodisi slovenščina ali angleščina, saj za večino analiz potrebujemo samo slovenske dele besedila, obenem pa je koristno ohraniti tudi angleške dele;
- zapis v skladu s shemo XML korpusa.

3.3 Jezikoslovno označevanje

V naslednji fazi je bilo besedilo vsakega dokumenta razčlenjeno na stavke in pojavnice (tokenizirano), oblikoskladenjsko označeno ter lematizirano, za kar smo uporabili nov označevalnik (Ljubešić in Erjavec, 2016), ki deluje na osnovi pogojnih naključnih polj in je bil modela jezika naučen na korpusu ssj500k (Krek et al., 2015) ter leksikonu Sloleks (Dobrovoljc et al., 2015). Označevalnik je sicer počasen, daje pa kakovostne rezultate. Evalvacija je namreč pokazala, da doseže na testni množici iz ssj500k 94,27-odstotno natančnost, kar je signifikantno več v primerjavi z 92,49-odstotno natančnostjo označevalnika Obeliks (Grčar et al., 2012), ki je do sedaj veljal za najboljši oblikoskladenjski označevalnik za slovenščino.

3.4 Zapis korpusa

Kot kaže slika 1, je vsako besedilo v korpusu zapisano kot svoj dokument XML. Korenski element `document` ima pripisane že omenjene metapodatkovne atribute, kamor spadajo tudi URL izvornega dokumenta v repozitoriju svoje univerze in URL (z geslom zaščitene) lokalne kopije PDF celotnega dela.

Dokument je sestavljen iz posameznih strani (`page`) in odstavkov (`p`) znotraj njih; če je bil prelom strani znotraj odstavka, je v XML premaknjen na njegov začetek, posamezna stran pa je lahko tudi prazna. Atributi strani so njena zaporedna številka in kazalka na lokalno kopijo strani v izvirniku, medtem ko ima odstavek pripisano kodo jezika vsebovanega besedila (`sl` ali `en`). Vsi elementi so opremljeni tudi z identifikatorjem (`@xml:id`).

Odstavki nato vsebujejo jezikoslovno označeno besedilo, in sicer povedi (`s`), znotraj njih pa besede (`w`), ločila (`pc`) in presledke (`c`), pri čemer sta prvima dvema pripisani še lema (`@lema`) in oblikoskladenjska oznaka (`@ana`) po priporočilih JOS.

Kot omenjeno, so izvirniki besedil dostopni tudi na strežniku projekta, nanje pa kažemo s kazalci URL iz elementa `document` in na določeno stran iz elementa `page`. Ta pristop je problematičen za javno uporabo korpusa, saj so celotni dokumenti PDF dostopni prek repozitorijev posameznih univerz, ki rade prepovejo njihovo nadaljnje razširjanje, dostop pa omogočijo šele, ko se uporabnik strinja s pogoji uporabe. Zato smo PDF razdelili na strani in te shranili na strežniku kot grafične datoteke PNG, pri čemer je vsaki dodan ključ (gl. `page/@facs_url` na sliki 1). S tem dobimo možnost uporabniku korpusa pokazati nekaj strani besedila, ob tem pa ne omogočamo prevzema celotnega izvirnika, podobno kot to dela spletna storitev Google Books.

3.5 Korpus na spletu

Korpusna besedila smo iz izvornih dokumentov XML pretvorili v t. i. vertikalni format, primeren za uvoz v konkordančnik, in ga vključili v lokalno instalacijo orodja noSketch Engine (Rychlý, 2007; Erjavec, 2013), s čimer dobimo možnost raznovrstnih analiz korpusnih podatkov.

[2] Gl. več na: http://udcdata.info/.

```
<document xml:id="kas-10000" doc_id="10000" text_id="16514" cobiss_id="7078419"
 title="Uravnoteženi sistem kazalnikov v poslovni banki X"
 author="Aver, Goran" supervisor="Bernik, Mojca" year="2012"
 publisher_abbr="UM FOV" publisher="Fakulteta za organizacijske vede" place="Kranj"
 url="http://dkum.uni-mb.si/Dokument.php?id=28143"
 type="Diplomsko delo" udc="005" udc_desc="Menedžment"
 pdf_url="http://nl.ijs.si/project/kas/pdf/000/kas-10000.pdf">
  <page xml:id="pb1" n="1"
   pdf_url="http://nl.ijs.si/project/kas/pdf/000/kas-10000.pdf#page=1"
   facs_url="http://nl.ijs.si/kas/facs/000/kas-10000/p0001-Pr4U.png">
    <p xml:id="pb1.p1" xml:lang="sl">
     <s>
      <w lemma="diplomski" ana="jos:Agpnsn">Diplomsko</w>
      <c> </c>
      <w lemma="delo" ana="jos:Ncnsn">delo</w>
        ...
```

Slika 1: Zapis korpusa v XML.

Zvrst	*besedil*	*%*	*strani*	*%*	*sl. besed*	*%*	*besed*	*%*	*pojavnic*
KAS-proto	50.793	100	3.796.957	100	952.172.179	100	992.429.078	100	1.189.100.226
Diplomska	41.212	81,14	2.819.462	74,26	686.276.048	72,07	711.048.854	71,65	850.937.549
Magistrska	6.401	12,60	704.960	18,57	192.438.734	20,21	200.965.696	20,25	240.145.441
Doktorska	700	1,38	147.049	3,87	38.208.949	4,01	42.872.974	4,32	52.874.876
Specialistična	573	1,13	50.144	1,32	14.153.388	1,49	14.474.521	1,46	17.068.921
Znanstvena	782	1,54	29.635	0,78	9.206.780	0,97	10.475.965	1,06	12.737.433
Strokovna	393	0,77	9.568	0,25	3.730.797	0,39	3.977.127	0,40	4.694.058
Ostalo	732	1,44	36.139	0,95	8.157.483	0,86	8.613.941	0,87	10.641.948

Tabela 1: Velikost korpusa KAS-proto po zvrsteh besedil.

4 Zgradba korpusa

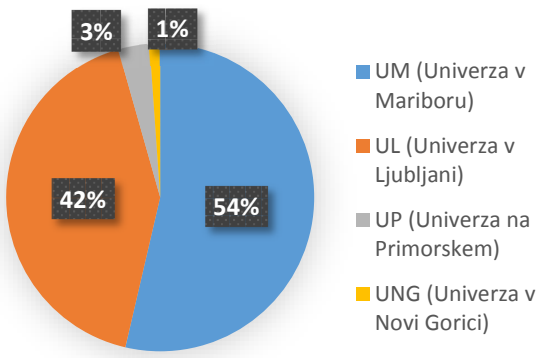
4.1 Zvrst

Tabela 1 vsebuje podatke o velikosti korpusa in njegovih posameznih zvrsteh po besedilih, straneh, številu besed v odstavkih, ki so bili identificirani kot slovenski in v celoti, ter po pojavnicah. Korpus vsebuje skoraj 1,2 milijarde pojavnic, s čimer je po velikosti primerljiv s trenutno največjim korpusom slovenskega jezika Gigafido (Logar Berginc et al., 2012).

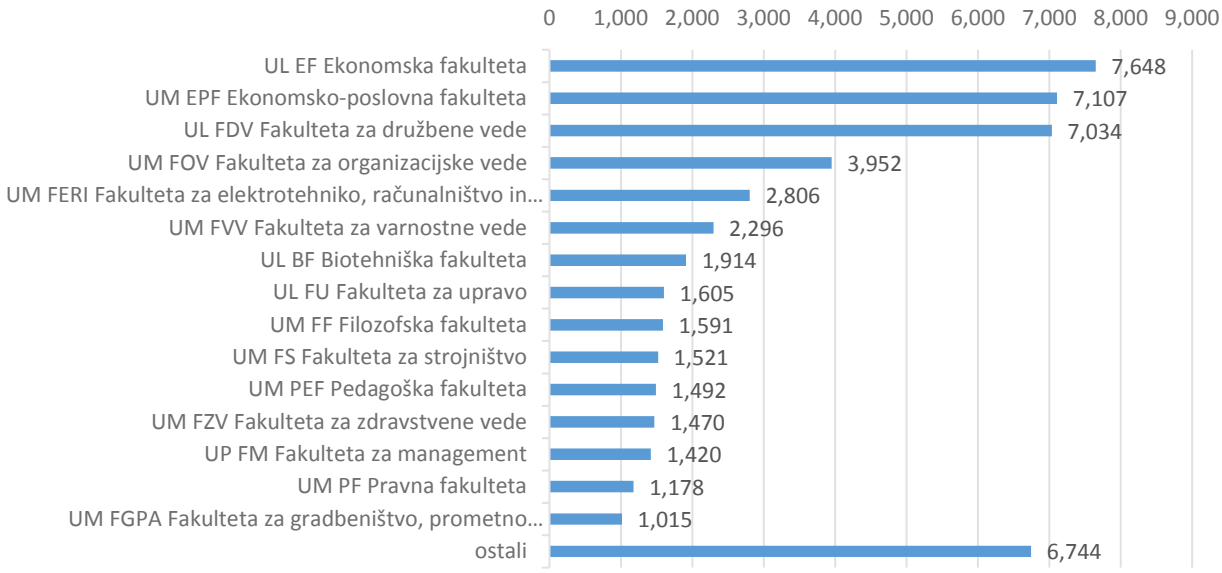
Daleč največji del korpusa predstavljajo diplomska dela, saj zajemajo dobre štiri petine vseh del oz. v korpus prinašajo skoraj tri četrtine strani ali 72 % vseh slovenskih besed. Sledijo magistrska dela in doktorske disertacije, iz katerih je v korpus prišla skoraj četrtina slovenskih besed.

Pri zvrsteh besedil, navedenih v zadnjih treh vrsticah, smo združili več tipov objav: pod znanstvena dela spadajo znanstveni prispevki na konferencah, izvirni znanstveni članki itd., pod strokovna dela strokovni članki, strokovne monografije itd., pod ostala dela pa npr. predgovori, spremna besedila in učna gradiva. Kljub temu, da vsa ta dela skupaj predstavljajo samo nekaj odstotkov korpusa, gre še vedno za razmeroma velike podkorpuse: nezaključna znanstvena dela npr. vsebujejo skoraj deset milijonov besed.

Zanimiv je tudi podatek, kolikšen delež korpusa predstavljajo angleški deli besedil, saj jih je, po eni strani, treba za enojezične raziskave filtrirati, po drugi pa so ti deli dragoceni kot neke vrste vzporedni ali vsaj primerljivi



Slika 2: Število besedil po univerzah.

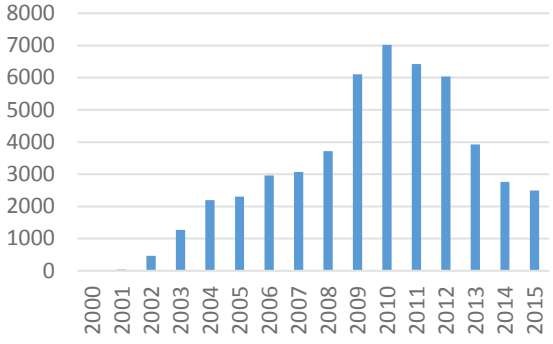


Slika 3: Število besedil po fakultetah.

podkorpus. V korpusu je približno 4 % angleškega besedila: večino prinašajo diplomska dela (3,5 %), po posameznih zvrsteh pa je najvišji delež angleških besedil značilen za doktorske disertacije (skoraj 11 %) in nezaključna znanstvena dela (12 %)

4.2 Vir

Slika 2 kaže, kolikšen delež besedil so v korpus prispevale posamezne slovenske univerze. Tu predvsem preseneča, da prihaja več kot polovica besedil z Univerze v Mariboru (UM), saj ima Univerza v Ljubljani (UL) vsaj dvakrat več študentov. Vzrok je v tem, da je UM repozitorij vzpostavila že leta 2008, UL pa šele leta 2013 (pred tem so na UL obstajale samo podatkovne zbirke posameznih fakultet, npr. Ekonomske fakultete, Fakultete za družbene vede in še nekaterih). Enako velja tudi za ostali dve slovenski univerzi, Univerzo na Primorskem (UP) in Univerzo v Novi Gorici (UNG), ki sta repozitorija prav tako vzpostavili šele leta 2013. Če v korpusu pogledamo število besed po univerzah, je sicer delež obeh največjih (UM in UL) podoben: dela z UM v KAS-proto prinašajo 49 % besed, dela z UL pa 47 % besed.



Slika 5: Število besedil po letih.

Z vzpostavitvijo repozitorijev je povezano tudi to, koliko besedil so v korpus prispevale posamezne fakultete. Na sliki 3, ki prikazuje 15 fakultet z največjim deležem besedil v korpusu KAS-proto (od skupno 55), je razvidno, da kar polovico vseh besedil prispevajo samo štiri visokošolske ustanove, in to vse družboslovne. Tudi sicer je del s tehničnih oz. naravoslovnih fakultet, vključenih v Nacionalni portal odprte znanosti (in posledično v KAS-proto), manj, še posebej pa zaostajajo humanistične vede. Tako se Filozofska fakulteta UM po obsegu del v korpusu še uvrsti na seznam prvih petnajstih, Filozofska fakulteta UL pa je šele na 18. mestu z le 847 deli.

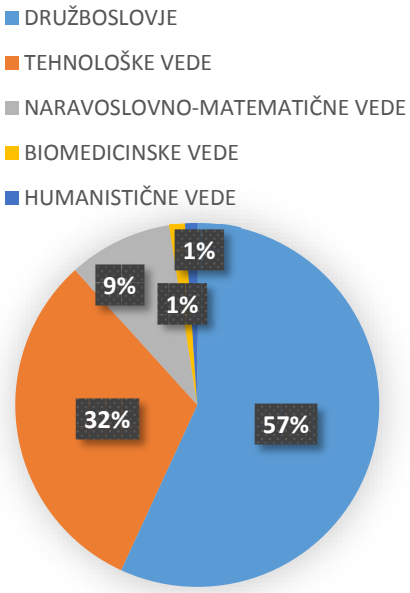
Zanimiva je tudi razporeditev del v korpusu po vedah. Prvi (tipično najpomembnejši) vrstilec UDK smo preslikali v vede taksonomije CERIF (Common European Research Information Format) in vsakemu delu v korpusu pripisali njegovo (glavno) vedo. Slika 4 podaja razmerja med posameznimi vedami.

Vede v veliki meri odslikavajo razdelitev del po fakultetah, saj tudi tu vodijo družboslovne vede, tem pa sledijo tehnološke vede z malo manj kot tretjino vseh del. Če je naravoslovno-matematičnih ved še za slabo desetino, je tako biomedicinskih kot humanističnih ved vsega skupaj 1 % vseh besedil.

4.3 Leto

Kot rečeno, smo iz zajetih besedil izločili dela, starejša od leta 2000, zajem pa je bil narejen v začetku leta 2016. Slika 5 podaja razporeditev števila besedil po letih objave. Iz leta 2000 jih je samo 5, iz naslednjega leta 39, iz leta 2002 pa že 470 in iz vseh kasnejših let po več kot tisoč. Presenetljivo je, da začne po letu 2010, še bolj pa po letu 2012, število besedil v repozitoriju strmo padati. Vzrok je v tem, da so univerze začele po letu 2012 vzpostavljati procese vstavljanja zaključnih del in pravne podlage za podporo tem procesom, kar je upočasnilo dodajanje novih del. Na UL bodo procesi dokončno vzpostavljeni šele do konca septembra 2016. Zaradi tega nekatere fakultete dela svojih študentov samo arhivirajo in jih ne odlagajo v

imenike. Drugi vzrok je v tem, da se število študentov na slovenskih univerzah vsako leto nekoliko zmanjša.



Slika 4: Število besedil po področjih.

5 Načrtovane raziskave in uporaba

Korpus KAS-proto bo omogočil izvedbo več jezikoslovnih in bibliotekarskih (ter v povezavi z obojimi tehnoloških) raziskav slovenskega akademskega jezika.

5.1 Klasifikacija besedil

Na podlagi korpusa KAS-proto in njegovih nadaljnjih različic bomo razvili metode za klasifikacijo besedil in luščenje ključnih besednih zvez, ki bodo izboljšale uporabnost portala odprte znanosti s tem, da bo z njimi omogočeno bolj kompleksno iskanje po vsebinah (Bordea et al., 2015; Fišer et al., 2010; Siddiqi in Aditi, 2015). S priporočili ključnih besednih zvez bo nadgrajen tudi vmesnik za knjižničarje, ki v univerzitetne repozitorije vnašajo nova besedila.

Pri tem bomo uporabili prosto dostopni (vrhnji) del klasifikacije UDK (UDC Linked Data), klasifikaciji ARRS in CERIF, za posamezna področja pa tudi MeSH, Eurovoc in Agrovoc. Zadnje tri klasifikacije so dostopne v XML ali SKOS/RDF, medtem ko bomo ostale za lažje procesiranje in izmenjavo v ta format pretvorili sami.

Kot glavno taksonomijo bomo uporabili UDK, saj je z njo že opremljena večina dokumentov v repozitorijih, poleg tega pa je taksonomiji ARRS in CERIF mogoče enostavno povezati s področji UDK. Taksonomija UDK ni zgolj šifrant, temveč je klasifikacijski sistem s pravili, ki omogočajo relacije med razredi, priredno in zaporedno razširitev področja, enostavne relacije ter podrobne delitve. Razvili bomo razčlenjevalnik za dekompozicijo razredov, ki bodo nato služili kot osnova za večlabelno klasifikacijo dokumentov.

Za učno množico bomo uporabili besedila v korpusu, ki že vsebujejo oznako UDK. Za vsak dokument bomo izluščili besede skupaj z njihovo utežjo TF-IDF. Nato bomo na podatkovni množici preizkusili več metod strojnega učenja. Za razvrščanje dokumentov po podobnosti bomo uporabili rangirno funkcijo BM25, za nadzorovano učenje klasifikacije pa bomo preizkusili različne metode, ki si dostopne v okviru knjižnice Scikit-learn.

V naslednjem koraku bomo izvedli preizkuse s spreminjanjem oz. razširitvijo podatkov ter značilk z:

- uporabo kombinacije določenih delov besedila (naslov, ključne besede, kazalo ipd.) namesto celotnega besedila;
- uporabo lem namesto besednih oblik iz besedila;
- uporabo pogostih n-gramov (enostavnih večbesednih enot) namesto posameznih besed;
- uporabo identificiranih terminov namesto splošnih večbesednih enot;
- dodatno uporabo nadpomenk ali pomenskih relacij pri terminih, ki so povezani z zunanjimi terminološkimi zbirkami;
- dodatno uporabo definicij pri povezanih terminih.

5.2 Razvoj orodij za delo s terminologijo

Luščenje terminologije poteka v treh korakih (Heyle in De Hertog, 2015): zaznavanje jezikovnih prvin, ki sestavljajo večbesedno enoto (zaznavanje enotskosti), razvrščanje po verjetnosti, da so izluščeni termini z določenega področja (zaznavanje terminološkosti) ter združevanje pomensko in konceptualno povezanih terminov (zaznavanje variantnosti), kar je pomemben korak, saj sem sodi kar 15–35 % izluščenih terminoloških kandidatov (Daille, 2005). Četrti, opcijski korak je identifikacija prevodnih ustreznic terminov v drugem jeziku (Daille et al., 1994).

5.2.1 Luščenje terminologije

Osrednji cilj je nadgradnja in evalvacija orodja CollTerm (Pinnis et al., 2012), ki je bil razvit v naših preteklih raziskavah. Trenutna različica terminološke kandidate izlušči s pomočjo oblikoskladenjskih vzorcev in več statistik za sopojavitev besed oz. besednih zvez, kot izhod pa ponudi urejen seznam terminoloških kandidatov, po en seznam za vsako stopnjo n-gramov, pri čemer je n tipično 1 – 6. Nadgradnja bo orodju dodala modul za nadzorovano učenje, ki bo filtriral in rangiral vsak element seznamov, s čimer bomo kot izhod dobili en sam seznam rangiranih terminoloških kandidatov.

Za delovanje modula bomo potrebovali korpus, ročno označen s termini, kar bomo izvedli s pomočjo spletne platforme za označevanje korpusov WebAnno (Eckart de Castilho et al., 2014).

V sklopu projekta bomo razvili tudi sistem za identifikacijo angleških prevodnih ustreznic, ki so na voljo v izvornih dokumentih v obliki dvojezičnih seznamov ključnih besed, tabelaričnih dvojezičnih glosarjev, dvojezičnih izvlečkov in povzetkov ali kot prevodne ustreznice v oklepajih.

5.2.2 Zaznavanje terminoloških variant

Drugi cilj je nadgradnja sistema z identifikacijo različnih poimenovanj istega pojma. Do terminološke variantnosti ne prihaja le zaradi terminološke večpomenskosti, ki je pogost meddisciplinarni pojav, temveč tudi zaradi stilističnih načel tvorjenja besedil in jezikovne gospodarnosti, s katero se izogibamo prekomernemu ponavljanju, zlasti v primeru daljših

terminov. Pričakujemo še, da bodo raznorodne rešitve uporabljali različni avtorji in v različnih časovnih obdobjih na področjih, na katerih se terminologija šele uveljavlja.

5.2.3 Analiza rabe terminov

Na podlagi identificiranih terminoloških kandidatov in njihovih variant bomo izvedli analize, ki bodo prvič omogočile celosten in dragocen vpogled v stanje ter trende terminološke rabe na različnih raziskovalnih področjih v Sloveniji. Za različne vrste akademskih besedil, strokovna področja in časovna obdobja bomo izmerili terminološko gostoto, stopnjo terminološke variantnosti in stopnjo terminološke interdisciplinarnosti.

5.2.4 Izgradnja terminoloških zbirk

Izluščeni terminološki kandidati bodo objavljeni v prosto dostopnem spletnem slovarskem urejevalniku, ki bo slovenskim znanstvenim in strokovnim skupnostim omogočal upravljanje s terminologijo lastnih področij. V izbranih skupnostih bomo pridobili tudi odziv na terminološko zbirko, ki jo bomo zanje pripravili v projektu.

5.3 Korpusni podatki za opis slovenskega akademskega jezika

Podatki iz obsežnega in področno raznolikega korpusa KAS nam bodo omogočili pripravo med vedami ter področji primerjalnega in različnim besedilnim žanrom prilagojenega opisa sodobnega slovenskega jezika, kakršen se rabi v akademskem okolju. Opis bo nastal s sintezo rezultatov treh vrst analiz:

- leksikalne analize,[3]
- besediloslovne in slovnične analize ter
- stilne analize.

5.3.1 Leksikalna analiza

Z metodo frekvenčnega profila (Rayson in Garside, 2000) bomo med drugim analizirali za akademsko pisanje značilno področno nespecializirano leksiko, za katero bi lahko rekli, da je del splošnega strokovnega jezika, npr. *definirati*, *določiti*, *analizirati*, *ključna beseda*, *metoda*, *vzorec*. Funkcija Besedne skice v leksikografskem orodju Sketch Engine (Kilgarriff et al., 2004) nam bo omogočila podrobnejšo analizo tipičnega besedilnega okolja tega besedišča (Logar, 2013b: 115–124), rezultate katere bomo nato ročno pregledali in jih enako kot zgoraj terminološke kandidate vnesli na prosto dostopen spletni portal.

5.3.2 Besediloslovna in slovnična analiza

Besediloslovna in slovnična analiza bo najobširnejša, saj nam bo korpus omogočal npr. analizo skladenjske zapletenosti povedi akademskega pisanja ter značilnih in izstopajočih skladenjskopomenskih kategorij, ki pripomorejo k temu, da je akademsko (znotraj njega zlasti znanstveno) pisanje natančno, jasno ter zgoščeno (Skubic, 1994/95). Posamezne slovnične kategorije (npr. trpnik, gl. Toporišič, 2000: 27–30) bomo opazovali primerjalno med posameznimi področji in primerjalno z leposlovnim delom korpusa ccGigafida (Logar Berginc et al., 2012). Korpusno bomo razčlenili tudi medbesedilnost (Hyland, 2004), ki je blizu področju plagiatorstva (Chandrasoma et al., 2004), in pregledali pojavljanje metabesedilnosti (Williams, 1981) kot dela retoričnih konvencij, ki se nanašajo na besedilo samo ali na odnos med tvorcem in naslovnikom.

5.3.3 Stilna analiza

Izbrani podkorpusi besedil različnih področij nam bodo dali vpogled v prvine (ne)osebnega stila akademskega pisanja in omogočili ugotovitev, ali poudarjena intelektualizacijska vloga (Južnič, 1992; Skubic, 2005) res v celoti izključuje avtorjevo prisotnost v besedilu. Preverili bomo tudi, kakšna je norma slovenskega akademskega pisanja, ter to, ali je znana tipa pisanja – germanski in anglosaksonski (Kalin Golob, 2008: 88–89) – na slovenskih besedilih mogoče (še) prepoznati ter ločiti.[4]

6 Zaključek

V prispevku smo predstavili korpus KAS-proto, njegovo izdelavo in kvantitativni vpogled v njegovo sestavo ter pregled načrtovanih raziskav.

Pred tremi leti smo v zvezi z aktualnimi terminološkimi opisi in njihovo dostopnostjo razmišljali takole: »V času nujnosti internacionalizacije strok in mednarodnega odpiranja njenih nosilcev je za polno funkcionalnost nacionalnega jezika na področju strokovnega jezika mogoče poskrbeti predvsem tako, da ga digitalno celostno podpremo ter si pri tem pomagamo prav z orodji, ki jih je prinesla digitalizacija« (Logar, 2013a: 251). V projektu »Slovenska znanstvena besedila« si bomo prizadevali to podporo še okrepiti in dokazati, da je pravzaprav šele s pomočjo takih orodij ter virov mogoče slovenščino v različnih žanrih akademskega diskurza zares dobro opisati; pri čemer ni zanemarljiva vloga še enega dejavnika: odprtosti tujih – in v prihodnje tudi naših – znanstvenih rezultatov.

Zahvala

Avtorji se zahvaljujejo anonimnima recenzentoma za koristne pripombe. Raziskavo, opisano v prispevku, je podprl projekt ARRS J6-7094 »Slovenska znanstvena besedila: viri in opis«.

7 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.

Akcijski načrt za jezikovno opremljenost. 2015, http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/raziskave-analize/slovenski_jezik/Akcijska_nacrta/ANJO.pdf.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, Roberto Navigli. 2015. Semeval-2015 task 17: Taxonomy Extraction Evaluation (TexEval). *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. 902–910, Denver, Colorado, June 4–5, 2015. ACL.

[3] V tem delu tiste leksike, ki jo lahko (oz. kolikor jo lahko) ločimo od terminologije.

[4] O drugem gl. še npr. Lengar Verovnik, Logar, Kalin Golob (2013: 29–49).

Ranamukalage Chandrasoma, Celia Thompson, Alastair Pennycook. 2004. Beyond Plagiarism: Transgressive and Nontransgressive Intertextuality. *Journal of Language, Identity, and Education* 3, 171–193.

Béatrice Daille. 2005. Variations and application-oriented terminology engineering. *Terminology*, 11/1, 181–197.

Béatrice Daille, Éric Gaussier, Jean-Marc Langé. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology. *Proceedings of the 15th conference on Computational linguistics, Volume 1*, Kyoto, Japan, 515–521.

Vijay K. Bhatia. 1993. *Analysing genre: Language use in professional settings*. London, New York: Longman.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih. 2015. Oblikoskladenjski leksikon Sloleks 1.2, *Slovenian Language Resource Repository CLARIN.SI*, http://hdl.handle.net/11356/1039.

Richard Eckart de Castilho, Chris Biemann, Iryna Gurevych in Sied Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands.

Tomaž Erjavec. 2013. Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0* 1/1, 24–49, http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf.

Tomaž Erjavec. 2009. Odprtost jezikovnih virov za slovenščino. V: M. Stabej, ur.: *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: ZIFF.

Tomaž Erjavec, Jan Jona Javoršek, Simon Krek. 2014. Raziskovalna infrastruktura CLARIN.SI. *Zbornik 9. konference Jezikovne tehnologije.* Ljubljana: IJS. 19–24.

Andrej Ermenc Skubic. 2005. *Obrazi jezika*. Ljubljana: Študentska založba.

Darja Fišer, Senja Pollak, Špela Vintar. 2010. Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources. *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*.

Ken Hayland. 2004. *Disciplinary Discourses: Social Interactions in Academic Writing*. Michigan: University of Michigan.

Kris Heylen, Dirk De Hertog. 2015. Automatic Term Extraction. *Handbook of Terminology, Volume 1*. John Benjamins Publishing Company, 203–221.

Stane Južnič. 1992. *Diplomska naloga: napotki za izdelavo*. Ljubljana: Amalietti.

Monika Kalin Golob. 2008. *Jezikovnokulturni pristop h knjižni slovenščini*. Ljubljana: FDV.

Monika Kalin Golob, Marko Stabej, Mojca Stritar Kučuk, Gaja Červ, Samo Kropivnik. 2014. *Jezikovna politika in jeziki visokega šolstva v Sloveniji*. Ljubljana: Založba FDV.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, David Tugwell. 2004. The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*. Lorient: Universite de Bretagne-Sud, 105–116.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz. 2015. Učni korpus ssj500k 1.4, *Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1052.

Tina Lengar Verovnik, Nataša, Logar, Monika Kalin Golob. 2013. *Slovenščina kot strokovni jezik na slovenskih univerzah: pregled stanja ter razčlenitev pomena, načina in možnosti njene večje vključitve*, http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/raziskave-analize/slovenski_jezik/Slovenscina_kot_strokovni_jezik_na_slovenskih_univerzah_01.pdf.

Nikola Ljubešić, Tomaž Erjavec. 2016. Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene. *Proceedings of 10th Language Resources and Evaluation Conference (LREC 2016)*.

Nataša Logar. 2013a. Aktualni terminološki opisi in njihova dostopnost. V: A. Žele, ur.: *Družbene funkcijskost jezika (vidiki, merila, opredelitve)*. Ljubljana: ZIFF. 247–253.

Nataša Logar. 2013b. *Korpusna terminologija: primer odnosov z javnostmi*. Ljubljana: Trojina: zavod za uporabno slovenistiko; Založba FDV.

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

Milan Ojsteršek, Mojca Kotar, Marko Ferme, Goran Hrovat, Mladen Borovič, Albin Bregant, Jan Bezget, Janez Brezovnik. 2014. Vzpostavitev repozitorijev slovenskih univerz in nacionalnega portala odprte znanosti. *Knjižnica* 58/3, 15–39, http://knjiznica.zbds-zveza.si/index.php/knjiznica/article/view/499.

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 the 10th Conference on Terminology and Knowledge Engineering*, Madrid, Spain, 193–208.

Paul Rayson, Roger Garside. 2000. Comparing Corpora Using Frequency Profiling. *Proceedings of the ACL Workshop on Comparing Corpora*. Hong Kong, 1–6.

Resolucija o Nacionalnem programu za jezikovno politiko 2014–2018. 2013, http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/Zakonodaja/2013/Resolucija_-_sprejeto_besedilo__15.7.2013_.pdf.

Pavel Rychlý. 2007. Manatee/Bonito – A Modular Corpus Manager. *Proceedings of the Workshop on Recent Advances in Slavonic Natural Language Processing.* Brno: Masaryk University, 65–70.

Sifatullah Siddiq, Aditi Sharan. 2015. Keyword and Keyphrase Extraction Techniques: A Literature Review. *International Journal of Computer Applications*, 109. 2.

Andrej Skubic. 1994/95. Klasifikacija funkcijske zvrstnosti in pragmatična definicija funkcije. *Jezik in slovstvo* 5, 155–168.

John M. Swales. 2000. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Jože Toporišič. 2000. *Slovenska slovnica*. Maribor: Obzorja.

Joseph M. Williams. 1981. *Style: Ten Lessons in Clarity & Grace*. Glenview, IL: Scott, Foresman and Company.

Sentiment Annotation of Slovene User-Generated Content

Darja Fišer,[†*] Jasmina Smailović,[*] Tomaž Erjavec,[*] Igor Mozetič,[*] Miha Grčar[*]

[†]Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si
[*]Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
jasmina.smailovic@ijs.si, tomaz.erjavec@ijs.si,
igor.mozetic@ijs.si, miha.grcar@ijs.si

Abstract

The Janes corpus contains posts from five different platforms (tweets, forums, blogs, comments on news articles and on Wikipedia) containing 167 million words of Slovene user-generated content. We have annotated the texts in the corpus with their sentiment, using a SVM-based sentiment classifier trained on a large collection of Slovene tweets. The paper introduces the classifier and its model for Slovene, gives an evaluation of the assigned scores and an analysis of the sentiment scores assigned to the text types of Janes.

1. Introduction

Sentiment analysis (also referred to as opinion mining) is a type of text analysis which detects opinions, sentiments and emotions about different entities. It can be applied in various scenarios, e.g., analysing public opinion about companies and products, voters' comments and debates regarding political parties, or investors' expectations about stocks, as well as analyses of the prevailing sentiment in written communication (Dodds et al., 2015). The first approaches to sentiment analysis emerged at the beginning of the century, and since then, it has gained increasing attention, esp. due to the massive usage of on-line platforms, such as blogs, forums and social networking services, where people regularly express their emotions about various topics (Liu, 2012; Liu, 2015).

We report on applying a pre-trained sentiment labelling system to a corpus of Slovene user-generated content (UCG), e.g. *ma nimam besed. Dost mam teh slinastih farjev,ki glumjo sirote,dnarja pa ko toče . Da ne pomislim na Zvon,Betnavo.. Plačajo naj p...! // I'm lost for words. I've had it with these sleazy loaded guys playing the broke card. I don't even want to think about the Zvon, Betnava.. Them f... should pay up!*. The goal of the paper is threefold: a) to evaluate the performance of the system on a collection of fairly heterogenous Slovene texts, b) to perform an analysis of the sentiment characteristics and distribution across different types of Slovene UCG and, last but not least, c) to add valuable metadata to the texts contained in the corpus.

2. Automatic Sentiment Labeling

In this section we describe the sentiment classifier that was used to automatically label the Janes corpus. In particular, we briefly outline the SVM-based algorithm for training the sentiment classifier, the manually labeled training data, and the data preprocessing steps.

2.1. Sentiment Classification Algorithm

We employ the Support Vector Machine (SVM) algorithm (Vapnik, 1995) to train a sentiment classification model. More precisely, we use the TwoPlaneSVMbin implementation, which is a three-class extension of the basic two-class SVM and is introduced in Mozetič et al. (2016). The three-class extension is needed to categorize texts into three sentiment classes: negative, neutral and positive.

TwoPlaneSVMbin is a combination of two binary SVM models, where one model separates the negative examples from the neutrals-or-positives while the other separates the positives from the neutrals-or-negatives. Therefore, two SVM hyperplanes are constructed. Additionally, the vector space is partitioned into bins (in our experiments the bin width is 0.2 of the SVM margin), and for each bin the information about the label distribution of the training examples is calculated. In the classification phase, a new example is projected into the vector space and a corresponding bin is determined. If the number of training examples in the bin is equal or higher than 5 and the distance from the hyperplanes is less than two margins, the class of the example is assigned based on the majority class in the bin. Otherwise, the class is determined based on the sides and distances from both hyperplanes in which the example resides.

2.2. Training Dataset

We acquired a large collection of Slovenian tweets during a joint project with Gama System[1]. The tweets were collected through the Twitter API, assisted by the PerceptionAnalytics platform[2]. The collected tweets are general (they do not discuss any particular entity) and were posted from January 2014 to February 2015.

The Twitter data was manually labeled by seven annotators. The resulting dataset contains 112,832 tweets labeled as negative (34,164), neutral (48,458) or positive (30,210). The dataset was used to train the sentiment classification model and to assess the classifier's performance by performing 10-fold cross validation. The details on evaluation are provided in Mozetič et al. (2016). It should be noted that this manually labeled dataset is not publicly available:

[1] http://www.gama-system.si
[2] http://www.perceptionanalytics.net

while it has been used to train the sentiment classifier and evaluate it, we cannot use it for further experiments.

2.3. Data Preprocessing

Before the training and the classification phase, the data is prepared by applying Twitter-specific and standard preprocessing techniques.

The Twitter-specific preprocessing includes: replacing URLs, hashtags, happy emoticons, sad emoticons, different combinations of punctuation marks, and mentions of Twitter users with common tokens; appending common tokens, which reflect the tweet length or provide information that a tweet contains a stock symbol or a term in uppercase; removing repetitive letters and appending a common token, which represent that a term contained repetitive letters; and normalizing diacritical characters.

The standard text preprocessing techniques consist of performing tokenization, lemmatization, unigram and bigram construction, removing terms which appear less than 5 times in the dataset, and constructing the normalized Delta TF-IDF (Martineau and Finin, 2009) feature vectors.

3. The Sentiment of Janes

The Janes corpus (Erjavec et al., 2015) is the first large (215 million tokens) corpus of Slovene user-generated content. While the corpus is still under construction, the current version of the corpus, Janes v0.4, already contains almost all the texts of the planned final version. The corpus is composed of the following text types:

- **BLOGp**: Blog posts from two popular platforms in Slovenia (*www.rtvslo.si* and *www.publishwall.si*);
- **BLOGc**: Comments on posts in BLOGc;
- **FORUM**: Posts on three popular Slovenian forums (*www.avtomobilizem.com* discussing cars, *med.over.net* on medical and related questions, and *forum.kvarkadabra.net* on scientific topics);
- **NEWS**: Comments on news articles in three popular Slovenian news sites (*www.rtvslo.si*, the portal of the national TV and radio, *www.mladina.si*, the main left-wing weekly magazine, and *www.reporter.si*, the main right-wing weekly magazine);
- **TWEET**: Tweets of 8,749 Slovene users in the period July 2013 – December 2015;
- **WIKI**: Pagetalk and usertalk pages from the Slovene Wikipedia.

Each text in the corpus is richly anntotated with metadata (e.g. author, title, time of post and, of course, sentiment score). Its content has also been linguistically annotated with a tool-chain that consists of rediacritisation, word-form normalisation, part-of-speech tagging and lemmatisation.

The texts in the Janes corpus were automatically annotated also for sentiment, using the SVM model as described in Section 2. As noted, the SVM training set is unavailable and is distinct from the Janes TWEET corpus, and we were interested how the system performs on our data, be it Tweets or other text types contained in Janes.

3.1. Sentiment by Text Type

In order to gain insight into the sentiment-annotated corpus, we created 18 subcorpora with texts of negative, positive, or neutral sentiment for each text type. As Table 1 shows, the largest of these subcorpora is the corpus of tweets with neutral sentiment. At the other end of the spectrum is the almost thirty times smaller subcorpus of wiki posts with positive sentiment. In all text types apart from tweets and wiki posts, negative content dominates. The smallest amount of positive as well as neutral content is in blog and news comments. Positive content prevails only in wiki posts while tweets are predominantly neutral.

Subcorpus	Senti	Tokens	%
BLOGp	neg	12,758,383	**72**
	neut	3,172,827	18
	pos	1,889,522	11
	total	17,820,732	100
BLOGc	neg	11,071,184	**69**
	neut	2,602,217	16
	pos	2,335,223	15
	total	16,008,624	100
FORUM	neg	25,529,662	**55**
	neut	12,715,683	27
	pos	8,053,284	17
	total	46,298,629	100
NEWS	neg	10,765,972	**74**
	neut	2,295,678	16
	pos	1,570,667	11
	total	14,632,317	100
TWEET	neg	32,493,298	34
	neut	36,092,424	**38**
	pos	26,202,339	28
	total	94,788,061	100
WIKI	neg	1,304,319	17
	neut	1,745,448	23
	pos	4,536,936	**60**
	total	7,586,703	100

Table 1: Sizes of the created subcorpora.

These results reflect the differences in the communicative role and nature of the various social platforms. While bloggers and commentators mostly use these on-line channels to express their opinions, disagreement and frustration with the daily politics and other events, forum members and Twitter users focus more on sharing information, news and knowledge, and Wikipedia editors prioritise community building efforts with supportive, encouraging and inclusive communication.

3.2. Sentiment by Key Words

The top key words reflect the domain of the focus corpus very well and can be used to explore differences between corpora (Kilgarriff, 2012). This is why we performed an analysis of 100 top-ranking key lemmas wrt. the complete corpus of that text type. The keyness score of a word is calculated according to the following formula:

$$\frac{fpm_f + n}{fpm_r + n}$$

where fpm_f is the normalized (per million) frequency of the word in the focus corpus, fpm_r is the normalized (per million) frequency of the word in the reference corpus, and n is a smoothing constant, with $n = 1$ the default value.

The key lemmas were manually classified — not taking into account the context they appear in — as positive, negative or neutral. Since they can be used either positively or negatively, proper names, place names and usernames were annotated as neutral lexical items. Mistokenised or mislemmatised words for which it was not possible to determine what they refer to out of context as well as noise in the form of URLs and foreign words were assigned an "other" tag. Intuitively, most keywords from a subcorpus of texts with negative sentiment would be expected to be negative, etc.

The confusion matrix presented in Table 2 show that subcorpora of tweets best follow this premise as the predominant category of key words is of appropriate sentiment in each subcorpus. The results are very good for all subcorpora of news and blog comments as well. Negative and neutral forum posts behave very well too while top-ranking keywords in the subcorpus of positive comments contain a little more out-of-context neutral words than positive ones. The positive supcorpus of wiki posts is slightly biased towards neutral key words while blog posts display the heaviest bias towards neutral expressions in both the negative and the positive subcorpus, suggesting our automatic sentiment analysis to be the least reliable for this text type.

		L_{neg}	L_{neut}	L_{pos}	Other
BLOGp	neg	21	**77**	1	1
	neut	4	**92**	4	0
	pos	0	**94**	6	0
BLOGc	neg	**64**	32	0	4
	neut	7	**77**	12	4
	pos	0	40	**57**	3
FORUM	neg	**98**	1	0	1
	neut	0	**97**	0	3
	pos	0	**59**	39	2
NEWS	neg	**92**	8	0	0
	neut	1	**75**	8	16
	pos	0	43	**57**	0
TWEET	neg	**99**	1	0	0
	neut	2	**89**	7	2
	pos	0	26	**74**	0
WIKI	neg	**58**	36	4	2
	neut	8	**84**	3	5
	pos	3	**83**	6	8

Table 2: Classification of 100 top-ranking positive, negative and neutral key lemmas for each subcorpus.

The fact that the results for Twitter subcorpora are the best is not surprising, given that the model for sentiment annotation was trained on tweets. News and blog comments which perform second best are not very different from tweets in terms of their length and usage and therefore seem to be almost equally reliably annotated with sentiment. Forum posts and wiki comments are longer but also deal with more specialized topics and have a different communicative purpose and target audience, which is why they are probably harder to annotate with a model trained on twitter data. The biggest outlier in terms of the results but also the most different as a text genre are blog posts.

		N	V	Adj	Adv	F	Np	Oth
BLOGp	neg	**51**	16	16	8	0	5	4
	neut	**38**	0	29	1	0	30	2
	pos	**66**	17	0	17	0	0	0
BLOGc	neg	**33**	**33**	24	10	0	0	0
	neut	32	3	16	1	4	**39**	5
	pos	**27**	4	16	26	**27**	0	0
FORUM	neg	**49**	22	25	4	0	0	0
	neut	**57**	1	31	1	1	7	2
	pos	18	10	**31**	21	20	0	0
NEWS	neg	**45**	27	16	8	3	1	0
	neut	29	1	12	0	3	**40**	15
	pos	**30**	7	18	15	**30**	0	0
TWEET	neg	**37**	21	35	5	2	0	0
	neut	**47**	8	18	6	1	18	2
	pos	30	8	**42**	12	8	0	0
WIKI	neg	**60**	16	19	3	2	0	0
	neut	**51**	0	19	2	0	22	6
	pos	**66**	17	17	0	0	0	0

Table 3: Distribution of keywords per part of speech.

3.3. Sentiment by Part of Speech

For a more detailed understanding of the linguistic nature of the most characteristic vocabulary of positively or negatively charged or neutral texts we performed a part of speech analysis of the 100 top-ranking key lemmas in all the 18 subcorpora. Part of speech assignment was manual. The part of speech was assigned even in cases of lemmatization or tokenization errors. In ambiguous cases, the most common part of speech was assigned. Apart from the main parts of speech, in particular nouns, verbs, adjectives, and adverbs, proper nouns (Np) were considered as a separate category because they were very prominent in certain subcorpora and called for a more detailed treatment. Non-Slovene words were annotated with a "foreign" (F) tag. Pronouns, conjunctions, abbreviations and interjections were also annotated, but were so infrequent in all subcorpora that they were subsequently merged into a single category called "other".

As can be seen in Table 3, nouns, with a 18-66% share, are the most prevalent overall. It is interesting, however, that different parts of speech are most indicative for different sentiments. All negative sentiment subcorpora display the highest proportion of top-ranking key nouns and a much higher proportion of verbs than subcorpora of positive or neutral sentiment. While positive sentiment subcorpora too contain a high proportion of nouns, the most prevalent part of speech in tweet and forum positive sentiment subcorpora are adjectives. Adverbs have the largest share in positively charged news and blog comments while blog comments and forum posts also contain a significant number of adverbs. Proper nouns figure by far the most frequently in neutral sentiment subcorpora, especially in neutral news comments where at 40% they are the most frequent cate-

gory. Neutral news comments are the only category with a significant share of abbreviations (15%) while foreign words, conjunctions and pronouns were all very rare this high on the key word lists for all subcorpora.

These results suggest that we use different linguistic means for communicating different sentiment. Negatively charged messages will be expressed directly, with nouns and verbs, while positive messages will be delivered descriptively, through adjectives and adverbs. Neutral, factual an informative content is characterized by frequent mentions of persons and their titles.

3.4. Sentiment Lexica

Finally, the 100 top-ranking key lemma lists from all positive and negative sentiment subcorpora were used to build sentiment lexica. Only the key lemmas that were manually annotated as negative or as positive were taken into account. All such lemmas were collected from all 5 subcorpora for each sentiment and added to the lexica. The negative sentiment lexicon created in this way contains 263 different words, 36 (14%) of which appear in three or more subcorpora. 44% of the lemmas in the lexicon are nouns, 25% each are adjectives and verbs, and 6% adverbs. The only two words that appeared in all five negative sentiment subcorpora are the verb *sovražiti* (hate) and the adverb *brezveze* (nonsense).

It is interesting to note that despite the fact that the keywords in subcorpora with positive sentiment showed a much greater variety in terms of their part of speech than their negative counterparts, the positive sentiment lexicon built in the same way contains only half as many words (146) as the negative one. 12% of these appear in at least three corpora, which is similar to the results in the negative sentiment lexicon. Here too the most frequent category are nouns (40%), followed by adjectives (29%) and adverbs (14%). Unlike in the negative sentiment lexicon where there are not found at all, interjections (9%) are an important part of the positive sentiment lexicon while verbs (7%) are barely present. The only word that appears in all five positive sentiment subcorpora is the interjection *bravo* (well done).

Sentiment lexica with lemmas that appear among the 100 top-ranking key lemmas in at least three subcorpora are listed, together with their translation into English, in Table 4 for negative sentiment and in Table 5 for positive sentiment. As can be seen from the tables, a major part of the negative vocabulary expresses personal stance, discontent with the political situation and the governing elites who are seen as corrupt and incompetent as well as the negative emotions authors of these message experience in response to unfavourable political and economic circumstances. The list also contains offensive and discriminatory words that indicate intolerance towards certain social groups. Positive vocabulary, on the other hand, is distinctly interactive and phatic, suggesting that the main communicative function of messages with positive sentiment is relationship and community building with positive feedback, such as praise, congratulations, thanking and good wishes.

Keyword	English	PoS	Subcorpora
baraba	bastard	N	4
bedarija	rubbish	N	4
drek	shit	N	4
lopov	crook	N	4
sranje	crap	N	4
svinjarija	bullshit	N	4
bruhanje	vomit	N	3
cigan	gypsy	N	3
gnoj	bullshit	N	3
kreten	idiot	N	3
kriminalec	criminal	N	3
laž	lie	N	3
sram	shame	N	3
sramota	shame	N	3
butast	stupid	Adj	4
žalosten	sad	Adj	4
beden	pathetic	Adj	3
bolan	sick	Adj	3
glup	stupid	Adj	3
kriv	guilty	Adj	3
nesposoben	incompetent	Adj	3
obupen	terrible	Adj	3
ogaben	disgusting	Adj	3
pokvarjen	corrupt	Adj	3
zmešan	crazy	Adj	3
sovražiti	**hate**	**V**	**5**
groziti	threaten	V	3
jebati	fuck	V	3
krasti	steal	V	3
nakladati	yack	V	3
pljuvati	spit	V	3
pobijati	kill	V	3
smrdeti	smell	V	3
ubiti	kill	V	3
brezveze	**nonsense**	**Adv**	**5**
žalostno	sad	Adv	3

Table 4: Negative sentiment lexica with keywords from at least three subcorpora.

4. Evaluation of the Sentiment Scores

We have performed a manual evaluation of the automatically assigned sentiment scores on a sample of the corpus. The sample contained random 600 texts, 120 from each text type, except form BLOGc, for which we obtained results consistent which news comments in keyword analysis and therefore did not include them in the dataset as we presumed that here too comments on blogs would be very similar to news comments. In addition, we balanced the number of texts from the sources of particular text types, e.g. for NEWS there are 40 texts from each of *www.rtvslo.si*, *www.mladina.si*, and *www.reporter.si*. This was done to arrive at a more diverse sample, as otherwise the much larger sources would swamp the smaller ones, e.g. the number of comments from *www.reporter.si* is only 5% of the those from *www.rtvslo.si*. It should also be noted that the length of an individual text varies widely between the text types. The shortest are tweets, with an average of 12 words per

Keyword	English	PoS	Subcorpora
čestitka	congratulations	N	4
pohvala	praise	N	3
poklon	bow	N	3
carski	great	Adj	3
dobrodošel	welcome	Adj	3
lep	nice	Adj	3
odličen	excellent	Adj	3
super	super	Adj	3
odlično	excellent	Adv	3
pohvalno	deserving compliment	Adv	3
srečno	good luck	Adv	3
bravo	**well done**	**Adv**	**5**
hvala	thank you	Adv	4
tooo	yesss	Adv	3
tnx	tnx	Adv	3
čestitati	congratulate	V	4
polepšati	make (sbd's day)	V	4

Table 5: Positive sentiment lexica with keywords from at least three subcorpora.

text, followed by news comments (42 words), Wikipedia (51), with blogs being the longest (71).

Each text was manually assigned a sentiment score by three annotators, where the annotators also had the option of marking individual texts as out of scope, as they were in a foreign language or contained e.g. adverts and were thus not user-generated, resulting in the final evaluation sample of 555 texts.

The manually assigned scores were compared to each other while the automatically assigned ones were compared to the majority vote (i.e. the label assigned by the most annotators). The agreement results in terms of Krippendorff's Alpha (Krippendorff, 2012) are given in Table 6. Perfect agreement is reached when $Alpha = 1$, while $Alpha = 0$ indicates agreement by chance. Acceptable inter-annotator agreement for this type of task is estimated at $Alpha > 0.4$ (Mozetič et al., 2016).

	All	Wiki	News	Blog	Forum	Tweet
Humans	0.563	0.464	0.513	0.594	0.464	0.547
Auto-major	0.432	0.402	0.394	0.446	0.245	0.372
n	555	107	115	115	119	99

Table 6: The agreement measures in terms of Krippendorff's Alpha for different sub-samples of the corpus which contain n texts.

The table confirms that assigning sentiment scores is a very subjective task and difficult to perform automatically. All the interannotator agreements are below 0.6 Alpha, which, while acceptable, is far from perfect agreement. The automatic assignment of sentiment labels is, of course, worse than the agreement between humans; while it is, overall, above the acceptability threshold, it is slightly below it for three out of five text types. However, it should be noted that the evaluation of the automatic system was quite strict, as it was compared to the majority class of the human annotators, i.e. even in cases the humans did not agree on the score, the system was penalised when it disagreed with the majority vote.

Blogs seem to be the easiest to assign a sentiment to, as both humans and the automatic assignment achieve here the highest score. This is most likely due to the length of the text, where it becomes clear which overall sentiment is expressed by the author. For humans, the second easiest are tweets, whereas the automatic system preforms worse on them than on News and Wiki. This is especially interesting as the automatic system was trained on tweets and would therefore be expected to perform best on the same type of texts. An explanation could be the short length of tweets, which does not give the system enough data to correctly determine the sentiment. Furthermore, it is likely that Twitter is often less straightforwardly opinionated than other types of text, i.e. it contains more ironic posts, which are hard for the automatic system to detect.

In general, the evaluation shows that it might help giving the annotators more precise instructions — preferably in line with those for annotating the training data — with which we would increase the interannotator agreement, while it is less clear on how to improve the quality of the automatic labelling. Here, providing additional training data from the text type that performed the worst, namely Forums, might be of help.

5. Conclusions

The paper presented a sentiment classification system trained on Slovene tweets and its application on the Janes corpus of Slovene user-generated content. The analysis of sentiment-specific keywords gives interesting insight into the vocabulary that is typically used to express different sentiment. Evaluation results show that automatic sentiment classification is consistent with human judgements and that there are considerable differences among the performance of the system across genres. Although the sentiment annotation accuracy could still be significantly improved, the current annotation of the Janes corpus is already useful for e.g., selecting only those texts that have predominantly negative, neutral or positive sentiment and performing on them targeted linguistic analyses.

A detailed analysis of disagreement among the annotators and an error analysis of incorrectly classified texts is planned in the future, which will reveal the outlying problem areas as well as provide clues for further refinements of the algorithm. Another venue of future work is to tackle irony and identify more fine-grained sentiment at the paragraph or even sentence level.

Acknowledgements

The authors thank the two anonymous reviewers for their useful comments and suggestions. We also thank Mojca Mikac for computing Alpha for the manual evaluation, and Sašo Rutar who implemented several classification algorithms and evaluation procedures in the LATINO library for text mining (https://github.com/latinolib). We acknowledge Gama System (http://www.gama-system.si) who collected the tweets for training the sentiment classifier. We thank Sowa Labs (http://www.sowalabs.com) for providing the Goldfinch platform for manual sentiment annotations.

This work was supported in part by the European Union project DOLFINS (no. 640772), by the Slovenian ARRS research project Resources, Tools and Methods for the Research of Nonstandard Internet Slovene (no. J6-6842), and by the ARRS programme Knowledge Technologies (no. P2-103).

6. References

Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdoomian, Matthew T. McMahon, Brian F. Tivnan, and Christopher M. Danforth. 2015. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2015. Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes. In *Zbornik konference Slovenščina na spletu in v novih medijih*, pages 20–26, Ljubljana. Znanstvena založba Filozofske fakultete.

Adam Kilgarriff, 2012. *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings*, chapter Getting to Know Your Corpus, pages 3–15. Springer Berlin Heidelberg, Berlin, Heidelberg.

Klaus Krippendorff. 2012. *Content Analysis, An Introduction to Its Methodology*. Sage Publications, Thousand Oaks, CA, 3rd edition.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Justin Martineau and Tim Finin. 2009. Delta TFIDF: An improved feature space for sentiment analysis. In *Proc. 3rd AAAI Intl. Conf. on Weblogs and Social Media (ICWSM)*, pages 258–261.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual Twitter sentiment classification: The role of human annotators. *PLoS ONE*, 11(5):e0155036.

Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.

Slovar tviterščine

Polona Gantar,[*] Iza Škrjanec,[‡] Darja Fišer, [*†] Tomaž Erjavec[†]

* Oddelek za prevajalstvo, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@ff.uni-lj.si,
darja.fiser@ff.uni-lj.si
‡ Ljubljana
skrjanec.iza@gmail.com
† Odsek za tehnologije znanja, Institut »Jožef Stefan«, Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

V prispevku opišemo postopek izdelave Slovarja tviterščine, ki je osnovan na korpusu uporabniško generirane slovenščine Janes. Najprej opišemo postopek luščenja korpusnih podatkov ter izdelavo geslovnika in se osredotočimo na kategorizacijo tviterske leksike z vidika standardizacije in stopnje podomačenosti. Nato predstavimo zgradbo slovarskega gesla in tip v slovar vključenih podatkov ter način urejanja v spletnem orodju za izdelavo slovarjev Lexonomy. Prispevek zaključimo z idejami za nadaljnje leksikalne analize tviterske leksike.

Dictionary of Slovene Twitterese

The paper describes the creation of the Slovene Twitterese Dictionary which is based on the corpus of Slovene user-generated texts Janes. First, the procedure of extracting corpus data and headword list creation are described, focusing on the categorisation of the Twitterese vocabulary with respect to standard Slovene and the levels of lexical adoption. Then, the structure of the dictionary entry and the types of lexicographic information included in the dictionary are described, along with the dictionary writing, editing and browsing platform Lexonomy. The paper concludes with plans for future lexical analysis of Slovene Twitterese.

1 Uvod

V prispevku predstavimo proces izdelave slovarja tviterščine na podlagi korpusa Janes Tviti v0.3.4 (Erjavec et al., 2015), ki poteka v okviru projekta »Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine«.[1] Slovarček, ki naj bi ob koncu projekta vseboval pribl. 800 gesel, bo predstavljal besedišče, ki se v okviru širše pojmovane računalniško posredovane komunikacije pojavlja v tvitih, tj. kratkih, na 140 znakov omejenih objavah (s fotografijami, videi ali brez) na družbenem omrežju Twitter.

Kot bomo podrobneje opisali v nadaljevanju prispevka, je besedišče, ki smo ga vključili v slovar, vezano na slovenske tvite, in sicer take, ki po avtomatski oceni (Ljubešić et al., 2015) vsebujejo pretežno nestandardno besedilo. V prispevku podrobneje pojasnimo kategorizacijo tviterske leksike, trenutno stanje, v katerem jo je mogoče pregledovati v spletnem slovarčku, ter orodje za izdelavo spletnega slovarčka.

2 Tviterščina v okviru računalniško posredovane komunikacije

V širšem kontekstu pojmovanja računalniško posredovane komunikacije velja, da je za leksiko na družbenih omrežjih značilen nestandarden, pogosto fonetiziran zapis besed, številne specifične okrajšave in veliko tujejezičnih elementov (Crystal, 2001; Baron, 2003).

Za angleščino so leksiko računalniško posredovane komunikacije začeli zbirati že zelo zgodaj, med najvidnejšimi pionirskimi zbirkami je *The Internet Dictionary* (Crumlish et al., 1995), eden najobsežnejših pa *A glossary of netspeak and textspeak* (Crystal, 2004). Sorodna zbirka je tudi glosar tipičnih okrajšav iz SMS-sporočil (Anjaneyulu, 2013). Medtem ko za spletno slovenščino nekaj parcialnih raziskav leksike že obstaja (Kalin Golob, 2008; Šabec, 2009; Erjavec in Fišer, 2013; Michelizza, 2015), celovitih popisov ali slovarskih zbirk zanjo še nimamo, kot tudi ne opisov leksike, specifične za družbena omrežja, kot sta denimo Twitter in Facebook.

Jezikoslovne analize različnih jezikov na družbenem omrežju Twitter od angleščine (Giai, 2013) in španščine (Álvarez et al., 2012), pa vse do indonezijščine (Bruggman in Conners, 2016) in malezijščine (Isa, 2014) sistematično kažejo, da je jezik v tvitih izrazito dinamičen, se hitro spreminja in prilagaja razvoju platforme ter je močno odvisen od namena in okoliščin komuniciranja in tako včasih bolj podoben javnemu pisnemu komuniciranju v klasičnih medijih, kot so novice ali blogi, drugič pa bolj zasebnim SMS-sporočilom oziroma pogovorom v spletnih klepetalnicah. Hu et al. (2013) so pokazali, da je jezik tvitov bistveno bolj konservativen in manj neformalen kot v SMS-sporočilih in spletnih klepetalnicah. Facchinetti (2015) ugotavlja, da v tvitih zaradi omejitve dolžine posameznega tvita na 140 znakov ne izstopa nobena jezikovna značilnost, kot so npr. strategije krajšanja sporočil, saj je povprečna dolžina tvitov v njenem korpusu zgolj polovična dovoljene, krajšave, nadomeščanje besed z nealfabetičnimi simboli ali črkami pa redke. Pogosto prisoten fonetiziran in prozodičen zapis besed ter fragmentirana skladnja sta značilna tudi za ostale oblike računalniško posredovane komunikacije (Herring, 2012), prve raziskave slovenskih tvitov v primerjavi s sorodnimi jeziki, kot sta srbščina in hrvaščina, pa poleg izjemno pogostega fonetiziranega zapisa kažejo na izrazito veliko prisotnost tujejezičnih prvin in različnih stopenj prevzetosti besed (Fišer et al., 2015).

[1] Spletna stran projekta: http://nl.ijs.si/janes/.

3 Namen in metoda izdelave slovarja

Namen slovarčka je izdelati nabor za slovenske tvite najbolj značilne leksike, pri čemer smo »značilnost« določali na podlagi različnih statističnih parametrov in jezikoslovnih kategorij. Posledično bodo informacije in način njihovega prikaza v slovarju namenjene seznanjanju potencialnih uporabnikov s tipično tvitersko leksiko, možnostmi zapisa posamezne besede in načinom vključevanja v slovenski jezik tako z vidika stopnje jezikovne podomačenosti kot z vidika pomena in rabe – zadnje predvsem v smislu izbire registra in govornega položaja (nestandardna beseda; kletvica, žaljivka ipd.).

Slovar oz. slovarska baza bo namenjena tudi leksikalnim analizam tviterskega žanra v širšem kontekstu leksike spletnih uporabniških vsebin za slovenščino, analizam jezikovnega (ne)standarda, načina podomačevanja tujih besed ter možnostim vključevanja v bodoče slovarske priročnike za slovenščino.

Pri oblikovanju geslovnika smo izhajali iz korpusa Janes Tviti v0.3.4 (Erjavec et al., 2015) in podatke o pogostnosti primerjali s korpusom Gigafida v1.0 ter s preostalim delom korpusa Janes, torej s forumi, komentarji in blogi (prav tam). Na ta način smo želeli izluščiti leksiko, ki je za tvite (a) bodisi bolj specifična in se v drugih žanrih računalniško posredovane komunikacije ne pojavlja oz. se pojavlja razmeroma redko (b) bodisi se v tvitih pojavlja opazno pogosteje kot v preostalih žanrih računalniško posredovane komunikacije. Pričakovali smo tudi, da je leksika, značilna za računalniško posredovano komunikacijo, prisotna tudi zunaj specializiranih spletnih žanrov. V ta namen smo frekvenčne podatke primerjali s korpusom Gigafida, hkrati pa upoštevali še vključenost in opis tviterske leksike v obstoječih enojezičnih slovarjih za slovenščino (Slovar slovenskega knjižnega jezika 2, Slovar novejšega besedja slovenskega jezika) ter novosti v pomenu in rabi.

Pri luščenju podatkov smo poleg frekvence upoštevali še razmerje med pojavnicami in lemami ter oblikoskladenjske oznake v korpusu. Na ta način je bilo mogoče prepoznati variantnost posameznih zapisov ter na podlagi ugotovitev analizirati proces vključevanja tujejezične leksike v slovenski jezik. Variantnost smo v slovarčku prikazali z nizom konkurenčnih variantnih zapisov, stopnjo podomačenosti pa z navedbo leksikalne kategorije (nova beseda) in stopnje podomačenosti (tuje podomačeno ali nepodomačeno).

3.1 Luščenje podatkov

Kot rečeno, je bil osnova za izdelavo geslovnika korpus tvitov, ki je bil zajet z namenskim orodjem TweetCat (Ljubešić et al., 2014b), s katerim smo identificirali uporabnike, ki tvitajo pretežno v slovenščini, in zajeli njihove tvite. Poleg besedila smo zajeli tudi metapodatke, kot so uporabniško ime avtorja, datum in čas pošiljanja ter število posredovanj (ang. *retweets*) in všečkov (ang. *favourites*) zajetega tvita. Korpus Janes Tviti v0.3.4 vsebuje več kot 56 milijonov besed oz. 4 milijone tvitov, ki jih je napisalo okoli 7.600 avtorjev. Metapodatke smo obogatili še z ročno določenimi podatki o lastnostih avtorja (tip računa, spol) ter avtomatsko pripisanimi podatki o sentimentu tvita.

Razvili smo tudi metodo, ki vsakemu tvitu v korpusu avtomatsko pripiše stopnjo jezikovne (zapis besed, raba slengizmov, narečnih in tujejezičnih besed, besedni red, slovnica) in tehnične standardnosti (raba ločil, presledkov, velikih/malih tiskanih črk). Ti dve meri (imenovani L in T) sta pripisani vsakemu tvitu v korpusu, in sicer z ocenami 1 (zelo standardno) – 3 (zelo nestandardno) (Ljubešić et al., 2015). Za namene slovarja smo upoštevali le tvite, ki so zapisani v jezikovno nestandardni slovenščini (L2 in L3) ter tako dobili podkorpus, ki vsebuje okoli četrtino celotnega korpusa, kar je milijon tvitov oz. 14 milijonov besed.

Na podlagi korpusa smo nato naredili frekvenčna leksikona lem in pojavnic. Prvi seznam je za izdelavo geslovnika sicer bolj uporaben, vendar so avtomatsko pripisane leme, posebej za besede v nestandardnem zapisu, velikokrat napačne, zato smo v oporo izdelali še drugega, kar nam je omogočilo tudi identifikacijo variantnih zapisov posamezne leme. Oba seznama smo z metodo frekvenčnega profila (Rayson in Garside, 2000) primerjali z lemami oz. pojavnicami celotnega korpusa Janes v0.3 ter s korpusom Gigafida. Tako smo dobili seznama ključnih besed in besednih oblik tviterščine v obeh virih, kjer vsak izpostavi njene specifične lastnosti, prvi glede na druge spletne uporabniške vsebine, drugi pa glede na splošnejše besedišče. Seznama obeh virov smo združili, v informacijo pa označili tudi tiste leme in oblike, ki se pojavljajo v Slovarju slovenskega knjižnega jezika (SSKJ). Vsaka vrstica v izdelanih seznamih tako vsebuje lemo oz. obliko, pogostost pojavitve na tisoč besed v korpusu Janes Tviti v0.3.4, frekvenco in ključnost glede na korpus Janes v0.3 in Gigafido ter informacijo, ali je lema oz. oblika zastopana v SSKJ.

3.2 Kategorizacija in izbor leksike

Seznama ključnih lem in besednih oblik smo prekrižali tako, da smo identificirali medsebojno povezane oblike, kar nam je predstavljalo izhodišče, t. i. širši geslovnik za jezikoslovno kategorizacijo. S posameznimi leksikalnimi kategorijami, ki jih prikazujemo tudi uporabniku, smo želeli opredeliti način vključevanja prevzete leksike v slovenski oblikoskladenjski sistem (krajšava; nova beseda; stopnja podomačenosti) in izpostaviti pomenske lastnosti, zlasti pomenske premike, ter značilnosti rabe, kot je npr. izbira registra (kletvica, žaljivka) in nestandardnosti. V nadaljevanju opišemo merila za določitev posamezne kategorije in podkategorije ter prenos informacije v spletni slovarček.

3.2.1 Nestandardni zapis in nestandardne besede

Z izrazoma nestandardna beseda in nestandardni zapis besede razumemo besede in zapise besed, ki jih ni mogoče pričakovati v besedilih, ki predstavljajo gradivno osnovo za standardizacijski opis jezika. Gre za besedila, ki sodijo v sfero javne pisne rabe, zlasti s področja javne uprave, izobraževanja in komunikacijskih medijev, ter znanstvena besedila (Skubic, 2005; Frawley, 2003). Za razumevanje pojma nestandardna leksika v slovarju tviterščine je zato treba najprej izpostaviti dejstvo, da je bil izbor tvitov v prvi vrsti vezan na pripis tehnične in jezikovne (ne)standardnosti (gl. poglavje o luščenju podatkov ter Ljubešić et al., 2015), kjer smo se namenoma odločili le za tvite z oznako L2 in L3, tj. za zgolj nestandardne tvite. Prvi kriterij nestandardnosti je torej statistični in ustreza predpostavki, da je določitev jezikovnega standarda lahko vezana le na razmeroma ozek segment jezika, kjer je kodifikacija zaželena in konsenzualno sprejeta (Krek,

2015).[2] Z namenom, da bi besedišče, ki se postopno vključuje v slovenščino zlasti prek angleščine (npr. *frendica*, *čekirati*, *luzer*, *lider*, *biznis*, *kul*) ločevali od že relativno ustaljenih prevzetih besed (npr. *cajteng*, *fršlok*, *kofe*), smo ta segment tviterske leksike dodatno opredelili kot nestandarden, seveda ob zavedanju, da je večina novejše prevzetih besed, ki v slovarju nimajo te opredelitve, prav tako nestandardnih v že zgoraj opredeljenem pomenu te besede.

Pri izdelavi ožjega geslovnika smo iz širšega geslovnika najprej izločili besede, ki so se znašle na seznamu ključnih lem in pojavnic zaradi nestandardnega, pogosto govornega zapisa besede, kar je glede na upoštevano stopnjo nestandardnosti že na ravni korpusa pričakovano. Te besede oz. oblike same na sebi niso posebej značilne za tvite, saj enakovredno pripadajo splošnemu besedišču, npr. *saj, jaz, kar*, *sem* ipd. Razlog, da so bile v postopku luščenja zajete v širši geslovnik, je njihov specifičen zapis, npr. *sj*, *js*, *kr*, *sm* ipd. Te besede nas z vidika pomena in rabe v slovarju tviterščine niso zanimale.

Nestandardne zapise besed smo ločevali od t. i. nestandardnih besed. S to kategorijo smo označevali (a) besede, ki imajo v jeziku prepoznavno standardno različico, npr. *bajta – hiša, crkavati – umirati, izležavati – lenariti*; tudi na ravni izbire registra, npr. *govno – sranje*, in so navadno prevzete iz tujega jezika, zlasti iz srbohrvaščine ali nemščine, npr. *švoh*, *cajteng*, *kao*, *čelav*. Novejših besed, prevzetih iz angleščine in nemščine, načeloma nismo opredeljevali s kategorijo standardnosti, čeprav imajo nekatere prav tako standardno ustreznico, npr. *hengati – družiti se; invajtati – povabiti*, ampak zgolj glede na stopnjo podomačenosti, o čemer več v razdelku 3.2.2. V nekaterih primerih smo kot nestandardne označili tudi (b) besede, ki nimajo ustrezne standardne različice, posledično pa so lahko prisotne tudi v standardnih besedilih, npr. *afnati se*, *pofočkati*, *štopati*, kar je povezano z njihovimi specifičnimi pomenskimi lastnostmi in izbiro registra, kot se kaže npr. v visoki stopnji pozitivnega ali negativnega vrednotenja. Kot nestandardne smo določili tudi (c) besede ki imajo ob standardni ustreznici tudi nestandardno obliko, ki je nastala bodisi kot posledica krnjenja, npr. *depresija → depra*; združevanja, npr. iz besedne zveze: *pornografski film → pornič*, ali izbire obrazila, npr. *penzionist → penzič*, *profesionalec → profič*.

Kot dodatno merilo smo za prepoznavanje nestandardnosti upoštevali lastnosti rabe, kot je denimo (d) izbira registra, npr. *govno, jeben, komunajzar*, *nadrkan*, in (e) pomenske lastnosti besede, npr. *hud* v pomenu '*lep, kakovosten*', *pičiti* v pomenu '*oditi, hitro iti*'. Zadnjo kategorijo bi bilo zato morda ustrezneje poimenovati »nestandardni pomen«. Kot zanimivost je mogoče dodati, da imajo besede, ki smo jih v širšem geslovniku označili s kategorijo »nestandardna«, če so vključene v SSKJ, v njem navadno kvalifikator *pogovorno* in *zlasti v sproščenem ožjem krogu* ter *nižje pogovorno*, *nizko*, *vulgarno*, *slabšalno*, *ekspresivno*, v nekaterih primerih tudi *zastarelo* (npr. *šetati*, *pušiti*, kjer gre tudi za pomenski premik) ali *starinsko* (npr. *glupost*).

3.2.2 Nove besede

Kategorija ***nova beseda*** je besedam v geslovniku pripisana na podlagi podkategorij, ki določajo stopnjo podomačenosti, zato oznake *nova beseda* slovarskim uporabnikom nismo prikazovali (zastopana je v slovarski bazi), je pa na dejstvo, da gre za besedo, ki se postopoma integrira v slovenski jezik, mogoče sklepati iz njene stopnje podomačenosti. Čeprav gre, kot rečeno, v večini primerov za nestandardne besede, ki imajo bodisi standardne ustreznice (npr. *browser – iskalnik*, *comp – računalnik, čekirati – preveriti; prijaviti se*) bodisi se zunaj korpusa tvitov skoraj ne pojavljajo (npr. *dejtati*, *folovati*, *ritvitati*) oz. se pojavljajo zelo redko (npr. *guglati, dron, logirati se*), jim oznake nestandardnosti v slovarju nismo eksplicitno pripisovali. Deloma zato, ker je njihova nestandardnost določena že z izborom besedil, deloma pa zato, ker zaradi relativno kratke prisotnosti v slovenščini njihovega standardizacijskega statusa še ni mogoče predvideti (prim. zlasti besede, kot so *bizarka*, *internetiti*, *odslediti*, *virtualka*, *tiskovka* ipd.).

Kategorijo ***tuje podomačeno*** smo pripisovali besedam, ki poleg enega ali več podomačenih zapisov ohranjajo tudi zapis v izvirniku, npr. *follower – folover*, in glede na to, ali je zapis oz. kateri od variantnih zapisov pisno in/ali glasovno podomačen, npr. *happy – hepi, cute – kjut*. Upoštevali smo tudi pregibanje po slovenskem oblikoslovnem vzorcu, kjer smo bili pozorni na prekrivnost osnovne oblike, kjer glasovna in pisna podomačitev ni potrebna, se pa beseda pregiba po slovenskem sistemu v neimenovalniških sklonih, npr. *link – linka; bed – biti v bedu*, včasih tudi prek postopne pisne in glasovne podomačitve, npr. *junk, džank – junka, džanka*. Kot stopnjo podomačenosti smo upoštevali tudi sposobnost tvorbe novih podomačenih oblik, ki se lahko uveljavljajo postopoma prek pisnega in glasovnega podomačevanja, npr. *follower*, *follover, folower, folover – followat* (najpogosteje), *follovat, folovat – pofollowat* (najpogosteje), *pofollovat pofolovat*. Pri izboru za ožji geslovnik smo upoštevali tudi, ali je katera od variant tipična predvsem za tvite in ali je že opisana v obstoječih slovarjih.

V slovarčku so posamezne variante, če so v korpusu tvitov izkazane vsaj trikrat, predstavljene znotraj variantnega niza (gl. sliko 4) in hkrati kot iztočnice. Na ta način je vsaki varianti na ravni iztočnice pripisana ustrezna kategorija podomačenosti, npr. *kjut* – tuje podomačeno; *cute* – tuje nepodomačeno, hkrati pa so na vsako varianto vezani tudi drugi podatki v slovarskem geslu, npr. potencialne kolokacije in korpusni zgledi. Oznaka ***tuje nepodomačeno*** je tako rezervirana za tiste variantne oblike, ki so glede na podomačeno obliko dovolj pogosto zastopane ali celo prevladujejo, npr. *annoying – anojing, deal – dil*, bodisi podomačene oblike (še) niso razvile, se pa v korpusu pojavljajo razmeroma pogosto, in sicer v slovenskem kontekstu, npr. *hardcore, multitasking* ipd. V to skupino sodijo tudi frazeološke enote, npr. *pitaj boga, lagano sportski, kein problem*.

3.2.3 Novi pomeni

Kategorijo ***nov pomen*** smo uporabljali za označevanje besed, ki so prišle v ožji geslovnik zaradi izkazanega pomenskega premika glede na obstoječi opis v SSKJ ali SNB, npr. *štekati*, *sledilec*, *koma*, *pičiti*. Te informacije v

[2] Tvite je pri izbiri standardizacijsko primernih besedil sicer smiselno upoštevati glede na dejstvo, da se je z razmahom spleta in s prehodom s papirja na zaslon možnost javne objave in dostopa do besedil bistveno povečala ter da je veliko še do nedavnega zasebnih žanrov prešlo v javno sfero (forumi, klepetalnice, družbena omrežja itn.) (Gorjanc et al., 2015a).

slovarskem geslu ne prikazujemo eksplicitno v obliki leksikalne kategorije, pač pa na ravni pomenskega opisa in kolokacij, če so izkazane, ter s tipičnimi korpusnimi zgledi, kar ilustrira slika 1.

štekati
SSKJ: *pogovorno*: razumeti, razumeti se
Novo: občasno prenehati delovati za krajši čas, predvsem v zvezi z elektronskimi napravami in programsko opremo (*a še komu Firefox zadnje čase šteka za popizdit*).

Slika 1: Primer obravnave novega pomena obstoječe besede.

3.2.4 Kratice, krajšave in alfanumerični znaki

Med kraticami in krajšavami ter alfanumeričnimi znaki,[3] kot npr. *gr8, ju3* ipd., smo v ožji geslovnik sprejeli zgolj tiste, ki se ne nanašajo na lastna imena, npr. izdelkov, in sicer ne glede na stopnjo podomačenosti: *iPhone, ajfon, ajfoun*, politična telesa, družbe in podjetja: *DZ, RKC, nyt* (New York Times), na zapise datotečnih formatov ter splošno rabljene kratice in krajšave, npr. *mr., cca, ipd.* itd. Enotna kategorija a ***krajšava***,[4] je tako v slovarskem geslu pripisana tistim občnoimenskim kraticam, okrajšanim besedam in zvezam, ki so za tviterski žanr tipične, npr. *app/ep; omg/omb/omajgad; tnx/thanks/tenks/thnx; bd/bday/rd*, in drugim. Tujejezične okrajšave s podomačenim zapisom imajo v slovarskem geslu v pomenskem razdelku vedno naveden tujejezični ustreznik in slovenski prevod, npr. *gr8/grejt* – krasno (great); *bdw/btW* – mimogrede (by the way).

4 Zgradba gesla in vrsta slovarskih podatkov

Tip slovarske informacije in zgradba gesla so navadno pogojene s preučitvijo uporabnikov in njihovih slovarskih potreb, vendar pa razmisleki v zvezi z naborom slovarskih informacij v slovarju tviterščine zaradi omejenosti projektne aktivnosti ne temeljijo na konkretnih uporabniških izkušnjah ali empiričnih raziskavah. Ob popisu tipične leksike, variant in pomenskih lastnosti smo si prizadevali izpostaviti predvsem tiste lastnosti tviterske leksike, ki jih je bilo mogoče v čim večji meri formalno in objektivno prepoznati v korpusu. Sem sodi zlasti variantnost oblik in stopnja podomačenosti v pisni in glasovni podobi. Dodano vrednost slovarja tako predstavlja kategorizacija na ravni standardnosti (nestandardna beseda) ter stopnje podomačenosti (tuje podomačeno/nepodomačeno), poleg tega pa še opis pomena in rabe, navadno v obliki kratkega pojasnila, ki je pri tujejezičnih besedah največkrat slovenski (*tenks – hvala*), pri nestandardnih pa standard(izira)ni sinonim (*komad – pesem*), ter navedba potencialnih kolokacij, zapis v izvornem jeziku in korpusni zgledi. Za pridobitev najmanj treh korpusnih zgledov smo uporabili aplikacijo GDEX v orodju Sketch Engine, ki je že bila preizkušena pri avtomatskem luščenju leksikografskih podatkov iz korpusa Gigafida (Kosem et al., 2013). Elemente, ki jih predvideva polni geselski članek v slovarju tviterščine so podani v sliki 2.

5 Orodje za izdelavo spletnega slovarja in objava slovarja na spletu

Pri izbiri slovarskega vmesnika smo upoštevali prosto dostopnost, čim večjo neodvisnost pri vnosu podatkov v shemo XML in možnost prenosa slovarskih gesel na lastni strežnik. Ključna je bila tudi fleksibilna nastavitev elementov v zgradbi slovarskega gesla, ki mora omogočati hierarhično ureditev in prilagoditev vizualizacije, saj se zastopanost elementov geselske zgradbe med posameznimi gesli razlikuje. Med možnimi platformami (Termania, Razvezani jezik, Wiktionary, DEBWrite in Lexonomy) smo se odločili za Lexonomy (Měchura, 2012), ki ga podrobneje opišemo v nadaljevanju, hkrati pa smo podatke prenesli tudi v program za izdelavo slovarjev iLex (Erlandsen, 2004), kjer bomo urejali slovarsko bazo.

iztočnica	osnovna oblika
kategorija	krajšava/nestandardna beseda/nova beseda
podkategorija	tuje podomačeno/tuje nepodomačeno
variantni zapisi	osnovne oblike var. zapisov
pomen	
slovenski del	
pomenski opis	standardni sinonim; kratek pomenski opis ali opis rabe; slovenski prevod
kolokacije	
tujejezični del	
izvirni zapis	zapis v jeziku izvirnika
zgledi	
zgled	korpusni zgled

Slika 2: Elementi predvideni v geselskem članku.

Lexonomy [5] je prosto dostopno spletno orodje za izdelavo slovarjev in njihovo neposredno objavo na spletu. Slovarski vmesnik deluje v spletnem okolju, zato namestitev programa na lastni računalnik ni potrebna. Uporabniki lahko pregledujejo objavljene slovarje v iskalniku ali pa si ustvarijo lastni račun in kreirajo svojo bazo podatkov (slika 3). Program omogoča preprosto izdelavo zgradbe slovarske baze, ki jo je mogoče v procesu izdelave slovarja enostavno spreminjati, in podpira skupinsko delo, saj lahko isti slovar ureja več uporabnikov.

Uporabnik lahko izbere že izdelano predlogo za kreiranje preprostega enojezičnega slovarja, mogoče pa si je ustvariti lastno predlogo in podatke kadarkoli objaviti na spletu. Podatke je v formatu XML mogoče izvoziti ali uvoziti, kar omogoča obdelavo in nadgradnjo v drugih programskih orodjih ter združevanje z drugimi podatkovnimi bazami.

Slika 4 prikazuje slovarsko geslo za iztočnico *dafaq*, ki vključuje opredelitev leksikalne kategorije in stopnjo podomačenosti. Sledi niz variantnih zapisov ter pomeni. Vsak registriran pomen lahko vsebuje pomenski opis,

[3] Za zadnje se uveljavlja tudi izraz *kratkopisne kratice* (Logar, 2004).

[4] Za nadpomenko krajšava smo se odločili zato, ker se formalne lastnosti, kot je npr. pika pri okrajšavah, in zapis s samimi velikimi črkami pri kraticah bodisi ne uporablja ali pa se uporablja nerazlikovalno.

[5] Lexonomy: http://www.lexonomy.eu/_en/.

kolokacije ter slovenski prevod ali pa zapis v izvirnem jeziku. Kot rečeno, so korpusni zgledi iz korpusa izluščeni s pomočjo aplikacije GDEX v orodju Sketch Engine. Namen korpusnega zgleda je potrditi registriran pomen in prikazati njegovo rabo ter tipično besedilno okolje na čim bolj avtentičen način.

Trenutni slovarček [6] vsebuje 21 testnih gesel, v nadaljevanju pa nameravamo v program uvoziti približno 1000 iztočnic, skupaj s pripisano leksikalno kategorijo in stopnjo podomačenosti. V program bomo avtomatsko uvozili tudi variantne zapise, pri čemer bo vsak variantni zapis, če je v korpusu dovolj pogost in je njegova raba razpršena med različnimi uporabniki Twitterja, v slovarju predstavljen kot samostojno geslo s podatki, vezanimi zgolj na konkretno varianto.

Slika 3: Izdelava slovarske baze v programu Lexonomy.

Slika 4: Prikaz gesla v Slovarju tviterščine na spletu v programu Lexonomy.

[6] Slovarček je prosto dostopen na http://lexonomy.cjvt.si/slovar-tviterscine/.

Čeprav je ena od osnovnih prednosti programa Lexonomy možnost neposredne objave slovarskega gesla na spletu, je njegova pomanjkljivost predvsem v tem, da ni mogoče vzdrževati razlike med slovarsko bazo, kamor želimo vključiti tudi korpusne metapodatke, kot so npr. tip uporabnika, spol, sentiment, standardnost, regija ter različne statistične vrednosti za posamezno lemo, vendar jih hkrati (še) ne želimo prikazovati navzven oz. jih želimo uporabnikom prikazovati na načine, ki jih v trenutni obliki Lexonomy ne omogoča, npr. v obliki grafov, preglednic ipd. Prav tako je naš namen čim več podatkov v slovarju neposredno povezati s korpusnimi viri in obstoječimi slovarji, ki so dostopni na spletu. Zaradi tega smo se odločili, da bomo slovar tviterščine kot bazo hranili tudi v programu iLex, kamor je mogoče za potrebe združevanja in medsebojnega povezovanja leksikalnih baz vključevati tudi podatke, ki jih ne želimo prikazovati navzven, so pa za leksikalne analize in nadaljnje nadgradnje koristni.

6 Zaključek in prihodnje delo

V nadaljevanju bomo slovarsko bazo nadgradili s podatki slovarskega tipa, kamor sodi oblikovanje pomenskih opisov, izbor relevantnih kolokacij, dodajanje tujejezičnih elementov pri razvezavi kratic in okrajšav ter izbor dobrih zgledov. Hkrati razmišljamo tudi o vključitvi podatkov leksikonskega in slovničnega tipa ter o izboljšavah avtomatskega luščenja podatkov iz korpusa. Pri gradnji slovarske baze tviterske leksike imamo ves čas v mislih možnost integracije v druge slovarske baze, npr. v slovarsko bazo za izdelavo Slovarja sodobne slovenščine (Gorjanc et al., 2015).

Obstoječo slovarsko bazo bomo v prihodnje izkoristili za nadaljnje raziskave tviterske leksike, zlasti za ugotavljanje načina integracije tujejezičnih elementov v slovenski jezik, kjer se uveljavljajo različne možnosti tako na ravni zapisa in morfologije kot tudi na ravni besedotvorja in skladnje. V nadaljnje analize želimo vključiti tudi podatke o tipu uporabnika in njegovi regijski pripadnosti in nenazadnje tudi podatke o analizi sentimenta in druge korpusne metapodatke. Predvidevamo, da bo na tej podlagi mogoče spremljati trend podomačevanja in določiti leksiko, ki se postopno vklaplja v slovenski leksikalni fond.

7 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.

8 Literatura

Irina Álvarez Argüelles in Alfonso Muñoz Muñoz. 2012. An insight into Twitter: a corpus based contrastive study in English and Spanish. *Revista de Lingüística y Lenguas Aplicadas*, 7.1: 37–50.

Thotapally Anjaneyulu. 2013. A glossary: usage abbreviations of mobile phone SMS. *et Cetera*, 70.2: 141.

Naomi S. Baron. 2003. Language of the Internet. Ali Farghali (ur.): *The Stanford Handbook for Language Engineers*. Stanford: CSLI Publications. 59–127.

Claudia Brugman in Thomas Conners. 2016. Comparative study of register specific properties of Indonesian SMS

and Twitter: implications for NLP. *Winter Storm*. College Park, Maryland.

Christian Crumlish et al. 1995. *The Internet Dictionary: The Essential Guide to Netspeak*. SYBEX Inc.

David Crystal. 2001. *Language and the Internet*. Cambridge: University Press.

David Crystal. 2004. *A glossary of netspeak and textspeak*. Capstone.

Tomaž Erjavec in Darja Fišer. 2013. Jezik slovenskih tvitov: korpusna raziskava. *Družbena funkcijskost jezika: (vidiki, merila, opredelitve), Obdobja 32*. Ljubljana: Znanstvena založba Filozofske fakultete, 109–116.

Tomaž Erjavec, Darja Fišer in Nikola Ljubešić. 2015. Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes. *Zbornik konference Slovenščina na spletu in v novih medijih*, Ljubljana, 25.–27. november 2015. Ljubljana: Znanstvena založba Filozofske fakultete, 20–26, http://nl.ijs.si/janes/wpcontent/uploads/2015/11/Konferenca2015.pdf.

Jens Erlandsen. 2004. iLex – new DWS. *Third International Workshop on Dictionary Writing systems (DWS 2004)*. Brno, 6. – 7. September 2004.

Roberta Facchinetti. 2015. English in social media: A linguistic analysis of tweets. *XIV Simposio Internacional de Comunicación Social. Santiago de Cuba*. 19-23.

Darja Fišer, Tomaž, Erjavec, Nikola Ljubešić in Maja Miličević. 2015. Comparing the nonstandard language of Slovene, Croatian and Serbian tweets. Smolej, M. (ur.). *OBDOBJA 34: Slovnica in slovar – aktualni jezikovni opis*. Ljubljana: Znanstvena založba Filozofske fakultete, 225–231.

William J. Frawley (ur.). 2003. *International Encyclopedia of Linguistics.* Oxford: Oxford University.

Enrico Giai. (2013). *Twenglish: A New Variety of English? A quantitative analysis of a Twitter based corpus,* http://www.tesionline.com/intl/thesis.jsp?idt=48368.

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek (ur.). 2015. *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete.

Vojko Gorjanc, Simon Krek in Damjan Popič. 2015a. *Med ideologijo knjižnega in standardnega jezika*. Ljubljana: Znanstvena založba Filozofske fakultete. 32–48.

Yuheng Hu, Kartik Talamadupula in Subbarao Kambhampati. 2013. Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language. *Zbornik ICWSM* 2013.

Monika Kalin Golob. 2008. SMS-sporočila treh generacij. Miran Košuta (ur.): *Slovenščina med kulturami, Zbornik slavističnega društva Slovenije 19*. Celovec, Ljubljana: Slavistično društvo Slovenije. 283–294.

Iztok Kosem, Polona Gantar in Simon Krek. 2013. Avtomatizacija leksikografskih postopkov. *Slovenščina 2.0*, 1(2): 139–164. http://slovenscina2.0.trojina.si/arhiv/2013/2/Slo2.0_2013_2_07.pdf.

Simon Krek. 2015. Standardni in knjižni jezik – drugi poskus. Smolej, M. (ur.). *Obdobja 34: Slovnica in slovar – aktualni jezikovni opis*. Ljubljana: Znanstvena založba Filozofske fakultete, 401–407.

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2015. Predicting the level of text standardness in usergenerated content. *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference*, 7–9. September 2015, Hissar, Bulgaria. Hissar: 371–378.

Nataša Logar. 2004. Nove tehnologije in nekateri nesistemski besedotvorni postopki. Kržišnik, E. (ur.) *Obdobja 22: Aktualizacija jezikovnozvrstne teorije na Slovenskem – členitev jezikovne resničnosti.* Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 121-132.

Michal Boleslav Měchura. 2012. Léacslann: A platform for building dictionary writing systems. V Ruth Vatvedt Fjeld and Julie Matilde Torjusen (ur.). *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012.* 855-861. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Mija Michelizza. 2015. Spletna besedila in jezik na spletu: primer blogov in Wikipedije v slovenščini. *Zbirka Lingua Slovenica, 6*. Ljubljana: Založba ZRC, ZRC SAZU.

Isa Na. 2014. *Language Use On Twitter Among Malaysian L2 Speakers*. Doktorska disertacija, University of Malaya Kuala Lumpur.

Paul Rayson in Roger Garside. 2000. Comparing Corpora Using Frequency Profiling. *Zbornik ACL Workshop on Comparing Corpora*. Hong Kong. 1–6. .

Andrej E. Skubic. 2005. *Obrazi jezika*. Ljubljana: Študentska založba.

Nada Šabec. 2011. The Globalizing Effect of English on the Language of the Slovene Media. V Vukanovic, Marija Brala; Krstanovic, Irena Vodopija (ur.). *The Global and Local Dimensions of English: Exploring Issues of Language and Culture*. Berlin: Dr. W. Hopf, 133–126.

Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter

Teja Goli,* Eneja Osrajnik,† Darja Fišer‡+

* Kropa 48a, 4245 Kropa
teja.goli@gmail.com
† Šerugova 10, 2000 Maribor
eneja.osrajnik@gmail.com
‡ Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
darja.fiser@ff.uni-lj.si
+ Odsek za tehnologije znanja, Institut Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana

Povzetek

V pričujočem prispevku obravnavamo pogostost in načine krajšanja slovenskih sporočil na družbenem omrežju Twitter. Za analizo smo uporabili vzorčni podkorpus 800 tvitov z različno stopnjo tehnične in jezikovne standardnosti ter v njem označili pojave krajšanja na ravneh zapisa, leksike in skladnje. V njih smo zabeležili skupno 3464 pojavov krajšanja, ki so bili uvrščeni v 32 različnih kategorij. V nestandardnih tvitih je krajšanja bistveno več kot v standardnih, količinsko in tipološko največ se jih pojavlja na ortografski, najmanj pa na skladenjski ravni.

Shortening Phenomena in Slovene Tweets

This paper addresses the frequency and types of shortening phenomena in Slovene tweets. The analysis was carried out on a subcorpus of 800 tweets of various levels of technical and linguistic standardness in which instances of shortening strategies were observed on the levels of spelling, lexis and syntax. We have registered 3,464 instances of shortening strategies that belong to 32 different categories. The results show that shortening strategies are much more common in non-standard tweets and that the highest number and widest range of shortening strategies arise on the orthographic level, whereas they are the lowest on the syntactic level.

1 Uvod

Družbeno omrežje Twitter postaja vse popularnejše tudi v Sloveniji. Ena njegovih glavnih značilnosti, po kateri se razlikuje od ostalih družbenih omrežij, je omejena dolžina posameznega sporočila na 140 znakov. Ta omejitev je prevzeta iz protokola SMS-sporočil, ki so omejeni na 160 znakov, pri čemer so pri Twitterju 20 znakov obdržali za zapis uporabniškega imena avtorja tvita (Rogers, 2014).

Med jezikoslovci, kot tudi novinarji in laično javnostjo, že od samih začetkov komunikacije preko SMS-sporočil velja, da je zanje značilno pogosto krajšanje sporočil, kot glavna razloga za to pa navajajo omejitev dolžine sporočila in neergonomske tipkovnice na mobilnih telefonih (Crystal, 2001), podobno naj bi veljalo tudi za visoko interaktivne oblike računalniško posredovane komunikacije, med katere sodijo tviti. Vendar tovrstne raziskave za večino jezikov še niso bile opravljene, tiste, ki so bile izvedene, pa kažejo, da o krajšanju sporočil kot univerzalnem pojavu tovrstne komunikacije ne moremo govoriti, saj se tako količina kot načini krajšanja sporočil od jezika do jezika močno razlikujejo (Bieswanger, 2007).

Po drugi strani je zanimiva ugotovitev, da je krajšanje pogosto tudi v sporočilih, ki ne dosegajo največje dovoljene dolžine (Thurlow, 2003), kar nakazuje na to, da načelo ekonomičnosti zaradi omejene dolžine ni avtorjev edini motiv za krajšanje sporočil, temveč ima tovrstna jezikovna raba tudi družbeno in čustveno razsežnost, s katero izražamo osebno identiteto in izkazujemo pripadnost neki družbeni skupini (Sveningsson, 2001), pa tudi ustvarjamo spontano, lahkotno vzdušje v komunikaciji (Androutsopoulos in Schmidt, 2001).

Cilj pričujočega prispevka je ugotoviti, ali, v kolikšni meri ter na kakšen način svoja sporočila krajšajo slovenski uporabniki Twitterja, ne glede na to, ali je krajšanje namerno/zavestno, posledica oz. kombinacija drugih dejavnikov ali povsem nenamerno/nezavestno. V naslednjem razdelku predstavimo sorodne raziskave, v 3. razdelku opišemo zasnovo pričujoče raziskave, v 4. razdelku podamo analizo rezultatov, nato pa prispevek sklenemo z zaključki in načrti za prihodnje raziskave.

1.1 Sorodne raziskave

Fenomeni krajšanja v t. i. novih medijih so v zadnjih desetletjih postali priljubljen predmet raziskav v številnih jezikih. Pri tem prednjačijo predvsem analize SMS-sporočil, e-pošte in spletnih klepetalnic (npr. Crystal 2001, Thurlow 2003), zadnje čase pa dobivajo pozornost tudi družbena omrežja. Alis in Lim (2013) z analizo 229 milijonov ameriških objav na Twitterju med 2009 in 2012 ugotavljata, da tviti v ameriškem prostoru postajajo vse krajši in kompaktnejši tako na sintaktični kot na leksikalni ravni. Poleg krajšanja zaznata porast žargonizmov in priložnostnih tvorjenk, zaradi katerih kljub krajšanju ne pride do izgube informacij.

Gouws et al. (2011) se posvečajo analizi in avtomatskemu označevanju nestandardnega angleškega besedišča avstralskih, hongkonških, britanskih in ameriških uporabnikov Twitterja. Med najpogostejšimi pojavi krajšanja identificirajo nadomeščanje večznakovnega niza z enim znakom, transformacije s pripono, izpuščanje samoglasnikov, nadomeščanje s predpono, nadomeščanje »you« z »u«, izpuščanje zadnje črke, združevanje besed ter nadomeščanje »th« z »d«.

Beiswenger (2007) primerja leksikalne redukcije, kot so inicializmi, krnitve, skrajšane oblike glagolov, črkovno-številčni homofoni, fonetični zapisi in logogrami, v nemških in angleških SMS-sporočilih. Ugotovi, da je v nemških SMS-sporočilih šestkrat manj redukcij kot v angleških, pri čemer Nemci pogosteje uporabljajo inicializme, Angleži pa vse ostale oblike krajšav.

V slovenskem prostoru krajšanje v SMS-sporočilih opazuje Jarnovič (2007), ki s svojo anketno raziskavo ugotavlja, da je jezikovna podoba SMS-sporočil odvisna predvsem od dojemanja SMS-ov s strani uporabnikov. Rezultati njene raziskave so pokazali, da je uporaba razpeta med tehnično in tehnološko prostorsko omejenostjo na 160 znakov, razumljivostjo sporočila ter ekspresivnim izražanjem. Po njenih ugotovitvah izpuščanje presledkov ter uporaba različnih ustaljenih in spontanih krajšav služita predvsem ustvarjalnosti pri izražanju in nista zgolj pojav, ki je odvisen od prostorskih omejitev. Ta sklep utemelji z dejstvom, da anketiranci z osnovno in srednjo izobrazbo prikazujejo bolj raznoliko uporabo krajšav kot ljudje z višjo in visoko izobrazbo.

V elektronski komunikaciji se krajšanja pri proučevanju rabe ločil dotakne tudi Michelizza (2012), ki zaznava neskladenjsko rabo pike in dvopičja, ki se pojavljata pri zapisu končnic internetnih domen ali formatov datotek (.exe, de:wiki) in prehajajo tudi v pridevniško rabo. Omeni tudi pogosto rabo stičnega tropičja in zvezdic, ki nakazuje izpust dela besede in predstavlja t. i. grafični evfemizem.

2 Zasnova raziskave

V pričujoči raziskavi smo analizirali pogostost in načine krajšanja slovenskih sporočil na družbenem omrežju Twitter. Za analizo smo uporabili vzorčni podkorpus tvitov, ki je že bil ročno označen na ravni stavčne segmentacije, tokenizacije in normalizacije nestandardnih besed (Čibej et al., 2016) ter obsega 800 tvitov z različno stopnjo tehnične in jezikovne standardnosti (Ljubešić et al., 2015), kot je podano v tabeli 1, pri čemer je bil vzorec nestandardnih tvitov (L3T3) za namene zagotavljanja jasnosti in doslednosti smernic za označevanje dvojno označen. Izmed vseh pregledanih tvitov je bilo 21 tvitov označenih kot nerelevantnih za raziskavo.

Vrsta tvitov glede na stopnjo standardnosti	Število tvitov
tehnično in jezikovno standardni (T1L1)	200
tehnično standardni, jezikovno nestandardni (T1L3)	200
tehnično nestandardni, jezikovno standardni (T3L1)	200
tehnično in jezikovno nestandardni (T3L3)	200
Skupaj	800

Tabela 1: Število označenih tvitov pri posamezni stopnji tehnične in jezikovne standardnosti.

Označevanje pojavov krajšanja je potekalo v orodju WebAnno, prosto dostopni označevalski spletni platformi, ki omogoča označevanje besedila na več ravneh (Eckart de Castilho et al., 2014).

3 Tipologija

Pojave krajšanja smo opazovali na ravneh zapisa, leksike in skladnje. Vse nivoje smo razdelili na najmanj dve in največ štiri podkategorije (gl. tabele 3, 4 in 5).

Pojavi krajšanja na nivoju zapisa se v osnovi delijo na izpuščanje, opuščanje in nadomeščanje. Pri *Izpuščanju* smo se osredotočali na vsakršno krajšanje zapisa besede na ravni črk, ne glede na razlog. Pri tem ločujemo izpuste samoglasnikov, izpuste soglasnikov in kombinacije izpustov samoglasnikov in soglasnikov, ki se nadalje ločijo glede na to, ali so izpuščeni na začetku (*mam*), na koncu (*rajš*) ali sredi (*tko*) posamezne besede. Krajšanje sporočil z opuščanjem presledkov in ločil uvrščamo v ločeno kategorijo, ki jo zaradi boljše preglednosti poimenujemo *Opuščanje*, pri čemer ne želimo vzbujati vtisa, da je neuporaba presledkov in ločil zavestnejša od izpuščanja črk. Opuščanje presledkov pri ločilih in simbolih ter med besedami, tj. združevanje besed, je označeno ločeno. Pri opuščanju ločil so upoštevana končna ločila, pike za okrajšavami ter deli večdelnih ločil, tj. treh pik, oklepajev in narekovajev. Kot nadomeščanje je opredeljena uporaba drugih (krajših) nizov črk, zapis s številčnimi homofoni, substitucija domačih črk s tujejezičnimi in obratno (tu smo upoštevali le tiste črke, ki niso del slovenske abecede), zapis z logogrami ter nadomeščanje leksemov z emotikoni, emojiji ali piktogrami.

Na leksikalni ravni smo delno izhajali iz tipologije Markusa Bieswangerja (2007) in upoštevali vse pojave, pri katerih izbrana oblika obsega manj znakov kot polno razvezana beseda oz. besedna zveza. V osnovi smo krajšave razdelili na ustaljene in neustaljene. Med prve spadajo kratice (*EU*, *LOL*, *FDV*, *KPK*), sledijo jim ustaljene okrajšave s piko ali brez (*npr., slo.*), pri čemer je manjkajoča pika označena na nivoju zapisa, posebej pa so označeni tudi ustaljeni simboli, formule in krnjene besede (*$, EUR*, *kg*). Med neustaljene krajšave so uvrščene žanrskospecifične krajšave, ki so nesklonljive (*rd*, *lp*, *Lj*, *btw*), sledi jim kategorija krajšanih neologizmov (*mata*, *appi*, *Zoki*), tretja neustaljana skupina krajšav pa so priložnostne krajšave, ki jih posamezni avtorji tvitov ustvarijo ad hoc in se izgovarjajo črkovno (*VD*, *pr.*).

Zaradi pestrih možnosti interpretacije pri označevanju sintaktičnih elips smo na skladenjski ravni označevali zgolj tiste izpuste izbranih besednih vrst, ki niso posledica sicer običajnih izpustov zaradi sobesedila, temveč evidentni izpusti zaimkov, predlogov, samostalnikov in glagolov (*na trstenjakovi dobra knjiznica bojda*). Izpusti glagolov so dodatno razdeljeni na podkategorijo pomožnih in glavnih glagolov.

Na vsaki ravni smo dodali še kategorijo Drugo, če bi med označevanjem naleteli na pojave, ki ne spadajo v nobeno od navedenih kategorij. Kadar smo pri isti besedi opazili več kot en pojav krajšanja, smo označili vse. Izjema je krajšanje na leksikalni ravni, saj v takih primerih nismo označevali izpuščanja in nadomeščanja črk na nivoju zapisa.

4 Potek označevanja

Označevanje je potekalo v treh fazah. Za vzpostavitev tipologije in smernic za označevanje je bilo najprej dvojno označenih petdeset testnih tvitov, ki so bili izbrani naključno in niso bili uporabljeni za raziskavo. Zatem je bilo za testiranje izdelane tipologije in smernic dvojno

označenih 100 jezikovno in tehnično nestandardnih tvitov (T3L3) ter 100 jezikovno standardnih in tehnično nestandardnih tvitov (T3L1). Za dokončno uskladitev tipologije je sledila faza kuriranja oznak omenjenih tvitov, ki je odpravila nesoglasja med označevalkama in s katero je bilo označevanje dokončno poenoteno. Nadaljnjih 600 tvitov iz preostalih kategorij tehnične in jezikovne standardnosti (T1L1 ter T1L3) sta označevalki označili enojno in brez kuriranja, sta pa med označevanjem komunicirali in skupaj razreševali morebitna vprašanja.

Enega od najbolj problematičnih vidikov označevanja so predstavljale narečne besede, saj se je izkazalo, da je težko ločevati med rabo nestandardne leksike in krajšanjem. Nestandardnih besed (*šipa*, *ketna*, *čuza* ipd.) nismo dojemali kot pojav krajšanja, čeprav so krajše od svojih standardnih ustreznic, medtem ko smo kot krajšanje razumeli npr. nestandardne redukcije končnic besed (*jedu* namesto *jedel*).

5 Analiza rezultatov in diskusija

Tabela 2 prikazuje pojave krajšanja po analiziranih nivojih, ki smo jih identificirali v različno standardnih tvitih.

Nivo	Stopnja stand.	Št.	%
Zapis (86,92 %)	T1L1	185	5,34 %
	T1L3	781	22,55 %
	T3L1	716	20,67 %
	T3L3	**1329**	**38,37 %**
Leksika (11,61 %)	T1L1	85	2,45 %
	T1L3	114	3,29 %
	T3L1	74	2,14 %
	T3L3	**129**	**3,72 %**
Skladnja (1,47 %)	T1L1	12	0,35 %
	T1L3	12	0,35 %
	T3L1	**17**	**0,49 %**
	T3L3	10	0,29 %

Tabela 2: Število posameznih identificiranih pojavov krajšanja.

Vsaj en pojav krajšanja smo identificirali v 89,4 % analiziranih tvitov. Vsega skupaj smo zabeležili 3464 pojavov krajšanja, ki so bili uvrščeni v 32 različnih kategorij. Za tri kategorije, ki smo jih predvideli v fazi oblikovanja tipologije, v analiziranem vzorcu nismo identificirali nobenega primera. Vse tri sodijo na ortografski nivo, in sicer nadomeščanje s številčnimi homofoni, nadomeščanje z logogrami in nadomeščanje z emotikoni/emojiji/piktogrami, kar je zanimivo, saj so ravno te v medijih pogosto izpostavljene kot značilni pojavi jezika v računalniško posredovani komunikaciji. Z oznako $0, ki predstavlja nerelevantne tvite za analizo, smo označili 21 tvitov. Ti med drugim vključujejo avtomatsko generirane tvite, tujejezične tvite ipd. V enem izmed primerov smo identificirali celo tvit, ki je v celoti napisan v bohoričici.

Največ pojavov krajšanja najdemo v tehnično in jezikovno nestandardnih tvitih (T3L3), saj smo v njih identificirali kar 43 % vseh redukcij. Sledijo jim tehnično standardni in jezikovno nestandardni tviti (T1L3), kjer smo zabeležili dobro četrtino (26 %) vseh pojavov krajšanja. Tehnično nestandardni in jezikovno standardni tviti (T3L1) vsebujejo 23 % redukcij, najmanj krajašanja (8 %) pa najdemo pri jezikovno in tehnično standardnih tvitih (T1L1).

Glede na vrhnji nivo v tipologiji močno prevladujejo redukcije na ortografski ravni, saj predstavljajo 87 % vseh pojavov krajšanja. Na leksikalnem nivoju smo zabeležili dobrih 11,5 %, na skladenjskem pa nekaj manj kot 1,5 % odstotka vseh pojavov krajšanja. 722 pojavnic v vzorcu je prejelo vsaj dve oznaki, pri 21 pa smo zabeležili po tri. Maksimalno število oznak na pojavnico je bilo štiri, najdemo pa jih pri dveh primerih (*nardil* (oznake OIKV, OOLK, OISV in OOPB) in *rtvslo* (oznake LNŽ, OOLO, OOPB in LUZ)).

5.1 Ortografski nivo

Kot je razvidno iz tabele 2, so pojavi krajšanja najpogostejši na nivoju zapisa. Motivacija za izpuščanje samoglasnikov in soglasnikov je pogosto približevanje zapisu, ki sledi govorjeni obliki besede, pri opuščanju presledkov in ločil je precej tudi nenamernega izpuščanja, kot so tipkarske napake, do namernega izpuščanja pa prihaja tako zaradi osebnega sloga in izražanja identitete uporabnika kot tudi tehničnih okoliščin, pri čemer razlogi niso vedno enoznačni.

Identificirano krajšanje na ravni zapisa je podrobneje predstavljeno v tabeli 3. Najpogostejše je opuščanje presledkov pri ločilih (29,76 %), čemur sledi izpuščanje samostalnikov na koncu besede (17,77 %), opuščanje končnega ločila (14,48 %) v stavkih, izpuščanje samostalnika na sredi besede (14, 31 %) in nadomeščanje daljših nizov črk s krajšimi nizi (6,91 %). Omeniti velja tudi, da se presledek pri ločilih najpogosteje opušča za vejico, nekoliko redkeje pa za piko ali tropičjem na koncu povedi.

Nivo1	Nivo2	Nivo3	*Primer*	Pogostost
Izpušč.	začetek b.	samostalnik	*mam*	1,89 %
		soglasnik	*lej*	0,20 %
		oboje	*koj*	0,07 %
	sredina b.	samostalnik	*bedn*	14,31 %
		soglasnik	*današnega*	1,13 %
		oboje	*kera*	1,13 %
	konec b.	samostalnik	*anglesk*	17,77 %
		soglasnik	*sam*	1,13 %
		oboje	*lah*	1,93 %
Opuš.	presl.	pri ločilih	*prepričano,da*	**29,76 %**
		med besedami	*inče tood njihni*	3,49 %
	ločil	konč.	*Ti to iz rokava streseš*	14,48 %
		pike za okr.	*Slo*	1,43 %
		večbes. ločila	*Kajmak in marmelada..*	3,42 %
Nadom.	daljših nizov s krajšimi		*ponuju*	6,91 %
	s številčnimi homofoni		-	0,00 %
	domačih črk s tujejezičnimi		*explozij*	0,53 %
	tujih črk z domačimi		*Tviter*	0,37 %
	z logogrami		-	0,00 %
	z emotikoni		-	0,00 %
drugo			*EPPja, 13incni*	0,46 %

Tabela 3: Delež posameznih kategorij krajšanja na nivoju zapisa.

Kategorije nadomeščanja na nivoju zapisa so zanimive tudi zato, ker smo v tipologiji predvideli več vrst krajšanja, kot smo jih z analizo gradiva identificirali. Nadomeščanj s številčnimi homofoni, logogrami in emotikoni/emojoji/piktogrami namreč v pregledanem vzorcu tvitov nismo našli, vendar predvidevamo, da bi se te kategorije pojavile, če bi označili večji vzorec tvitov.

Identificirali pa smo tako nadomeščanje domačih črk s tujejezičnimi in tujih črk z domačimi kot tudi nadomeščanje daljših nizov črk s krajšimi, vendar se je pri nadomeščanju z domačimi črkami izkazalo, da tako prilagojen zapis v večini primerov ni krajši od izvirnega (*stori* namesto *story*, *vorld* namesto *world*). Pri nadomeščanju domačih s tujimi črkami so novi izrazi dejansko krajši od izvirnih, zanimivo pa je, da je v večini primerov (v 9 primerih od 11 identificiranih) niz črk »ks« zamenjan s tujejezično črko »x« (npr. *expert* namesto *ekspert*).

5.2 Leksikalni nivo

Na leksikalnem nivoju smo zabeležili dobrih 11 % vseh pojavov krajšanja.

Nivo1	Nivo2	*Primer*	Pogostost
ustalj. krajšave	začetnice/ kratice	*BDP*	**34,83 %**
	okrajšave	*št., slo.*	11,69 %
	simboli, formule, krnjene besede	*€, +*	14,93 %
neustal. krajšave	žanrsko-specifične	*Tw.*	12,44 %
	krajšani neologizmi	*appi, Zoki*	9,45 %
	priložnostne	*sod[nik]*	12,94 %
drugo		*e-sožalje*	3,73 %

Tabela 4: Delež posameznih kategorij krajšanja na leksikalnem nivoju.

Najpogostejša je raba ustaljenih krajšav (34, 83 %) z začetnicami ali kraticami, kamor spadajo imena strank (*SDS*, *PS*), državnih tvorb (*EU*, *ZDA*) in osebna lastna imena (*JJ*). Te zelo variirajo, saj med 140 označenimi krajšavami najdemo kar 104 različne. 27 različnih krajšav s simboli se v analiziranih besedilih pojavi 60-krat. Najpogostejša je uporaba znaka +, ki največkrat nadomešča veznik *in*, ter črke *x*, ki je uporabljena namesto besede *krat*.

Med najzanimivejše identificirane pojave krajšanja sodijo neustaljene krajšave, ki imajo izrazit strateški značaj časovnega, prostorskega oz. tehničnega krajšanja sporočil, zanimive pa so tudi z vidika kreativnosti uporabnikov in pestrosti jezika v elektronskih medijih. Tej kategoriji sledijo neustaljene priložnostne in žanrske krajšave, ki se pojavijo 50-krat. Priložnostne krajšave smo definirali kot ad hoc krajšanja, pri kateri avtor uporabi skovanko, ki jo lahko razumemo zgolj s pomočjo konteksta, v katerem se pojavi. Zabeležili smo 47 različnih priložnostnih krajšav, od tega je najpogostejši *PV* (v pomenu *predsednik vlade*), ki se trikrat. Med najzanimivejšimi so npr. *KlincaTM* (Univerzitetnega kliničnega centra), *upravič* (upravičiti) ter *odl*. (odločitev). Tudi neustaljene žanrske krajšave se pojavijo 50-krat. Kot te razumemo vse tiste krajšave, ki se izgovarjajo po posameznih črkah in so tipične za elektronska besedila. Med najbolj značilne sodijo npr. *rt*, *btw*, *fb*, *lj* in *jbt*.

S 47 pojavitvami in 22 različicami jim sledijo ustaljene okrajšave, pri čemer uporaba pike za njimi ni vplivala na njihovo (ne)ustaljenost, saj smo manjkajočo piko označili na nivoju zapisa. Med žanrske krajšave med drugim sodijo *cca*, *dr.*, *itd.* in *oz*. Na tem nivoju smo beležili še neologizme, v katere smo uvrstili tiste krajšave, ki se izgovarjajo kot besede, in jih je v analiziranem vzorcu najmanj (9,45 %). Sem spadajo tvorbe, kot so *alko*, *app* in *simultanka* ter različni vzdevki, kot denimo *Zoki* in *Bojči*. Našteli smo jih 38, med njimi je 28 različnih. Od priložnostnih krajšav se ločijo tudi po tem, da so večinoma razumljive tudi brez sobesedila.

5.3 Skladenjski nivo

Določanje izpustov oz. krajšanja na skladenjskem nivoju je že samo po sebi problematično, saj je nemalokrat težko oceniti ločnico med namernim krajšanjem in elipsami, ki pripomorejo k večji koherentnosti besedila. Iz tega razloga smo pri označevanju upoštevali le take izpuste besednih vrst, ki niso bili posledica »običajnega« izpusta zaradi sobesedila. Za boljšo predstavo v nadaljevanju navajamo primer izpusta pomožnega glagola, ki smo ga označili na besedi pred identificiranim izpustom, v tem primeru na pojavnici *včeraj:*

Včeraj **[SGG]** *bil na kuhančku v Ljubljani.*

Nivo1	Nivo2	*Primer*	Pogostost
izpust glag.	pomož.	*uf, zdej mi ze vec stvari jasnih hehe thx za info /.../*	**60,78 %**
	glavnega	*Bi blo treba tiralico?*	13,73 %
izpust zaimka		*...strinjam popolnoma...*	3,92 %
izpust predloga		*ocene /.../ se izkažejo predvsem [SD] politično motivirane*	3,92 %
izpust samostalnika		*/.../ uhhh ova je lejpa ka ma na boki [SS] od vseh knjig...*	7,84 %
drugo		*/.../ kjer lahko lajkaš to stran, ne pa vas morm dodat za frenda /.../*	9,80 %

Tabela 5: Delež posameznih kategorij krajšanja na skladenjskem nivoju.

Kot vidimo v tabeli 2, so pojavi krajšanja na skladenjskem nivoju najredkejši, najverjetneje zato, ker skladnja ne omogoča podobne kreativnosti, kot jo zaznamo na leksikalni ravni (npr. uporaba ustaljenih ali neustaljenih krajšav) ali na nivoju zapisa (npr. izpuščanje določenih črk ali nadomeščanje nizov črk), hkrati pa je bistvenega pomena za razumevanje besedila, zato lahko izpuščamo le določene skladenjske elemente.

Najpogostejši pojav krajšanja na skladenjskem nivoju (gl. tabelo 5) je izpust pomožnega glagola *biti*, ki predstavlja kar 60,78 % pojavov krajšanja na skladenjskem nivoju. Temu s 13,73 % sledi izpust glavnega glagola, kar je razumljivo, saj izpust pomožnega glagola načeloma ne onemogoča razumevanja sporočila, pri izpustu glavnega

glagola pa se lahko bistvo sporočila izgubi in je tako razumevanje oteženo. Še redkeje se pojavlja izpust samostalnika, ki že precej otežuje razumevanje besedila. Najredkeje pa se s 3,92 % pojavljata kategoriji izpusta zaimka in predloga, ki ne vplivata bistveno na razumevanje sporočila besedila.

5.4 Drugo

Med označevanjem smo naleteli tudi na nekaj primerov, ki jih nismo mogli umestiti v nobeno izmed osnovnih kategorij. Za take primere smo na vseh nivojih uporabili kategorijo *Drugo*.

Na nivoju zapisa smo sem prišteli označili tiste besede, v katerih je manjkal vezaj (npr. *euprava*, *90tih*, *rtvja*). Skupno se je nabralo 14 takih oznak.

Na leksikalnem nivoju smo pod *Drugo* uvrstili npr. besedi *Desusovec* in *ex*, v zvezi *ex politični zapornik*.

Na skladenjski ravni smo zabeležili 5 nepredvidenih pojavov krajšanja, ki so najverjetneje nenamerni izpusti oz. napake, saj zanje nismo našli nobene druge razlage. Primer:

> *@Cvetlicarna Kak FB profil mate to? A ne bi raj naredili en page, kjer lahko lajkaš to stran, ne* ***pa*** **[SD]** *vas morm dodat za frenda (kar ne dela)?*

Iz naslednjega navedenega primera lahko razberemo, da je tvit na koncu odrezan, kar smo označili kot skladenjski izpust, manjka pa tudi končno ločilo na nivoju zapisa. Tak izpust smo ugotovili pri dveh primerih:

> *punce :) vasja danc na bazenu v sgjo od 3 ure dalje :) tk da lahka pridete ko ma tako zeljo vas* ***s*** **[SX]** *http://t.co/qvk6jeMeBc*

6 Zaključek

V pričujoči raziskavi smo analizirali pogostost in načine krajšanja slovenskih sporočil na družbenem omrežju Twitter. Analiza je pokazala, da izmed vseh 800 analiziranih tvitov le nekaj nad 10 % zapisov ni vsebovalo nobenega krajšanja. Trend krajšanja je med slovenskimi uporabniki Twitterja torej zelo pogost. Največ krajšanj najdemo v tehnično in jezikovno nestandardnih tvitih (T3L3), kjer smo identificirali 43 % vseh redukcij, najmanj krajšav (8 %) pa najdemo pri jezikovno in tehnično standardnih tvitih (T1L1).

Skupno smo zabeležili 3464 pojavov krajšanja, ki so bili uvrščeni v 32 kategorij. Pri 21% pojavnic smo hkrati identificirali dve ali več pojavov krajšanja, maksimalno število oznak na pojavnico je bilo štiri, ki smo jih označili v dveh primerih. Glede na nivoje, ki smo jih določili v tipologiji, močno prevladujejo pojavi krajšanja na nivoju zapisa, ki predstavljajo kar 87 % vseh pojavov krajšanja. Na tem mestu velja omeniti tudi, da ortografski nivo vključuje največ različnih vrst krajšanja. Na leksikalnem nivoju smo zabeležili približno 11,5 %, na skladenjskem pa nekaj manj kot 1,5 % odstotka vseh krajšanj. Najpogostejši tip krajšanja je opuščanje presledkov pri ločilih, ku mu sledi izpuščanje samoglasnikov na koncu besede, najredkeje pa sta rabljena izpusta predloga in zaimka.

Z vidika kreativnosti uporabnikov so najzanimivejše neustaljene krajšave na leksikalnem nivoju, ki smo jih razdelili na priložnostne in žanrsko specifične. Pri priložnostnih krajšavah avtor uporabi lastno skovanko, ki jo razumemo zgolj s pomočjo sobesedila npr. *KlincaTM* (Univerzitetnega kliničnega centra, žanrsko specifične krajšave pa se izgovarjajo po posameznih črkah, so nesklonljive in so značilne za elektronska besedila (*jbt*, *btw*, *fb*, *lj*, *rt*). Med neustaljene krajšave spadajo tudi neologizmi, kot so *alko*, *app* in *simultanka* ter različni vzdevki, kot denimo *Zoki* in *Bojči*. Od priložnostnih krajšav se ločijo po tem, da so razumljive tudi brez sobesedila.

Z analizo smo prišli do jasne ugotovitve, da so slovenski uporabniki zelo nagnjeni h krajšanju besedil in da najpogosteje posežejo po krajšanju zapisa besedila (izpusti presledkov, ločil), manj pogosto pa posežejo po leksikalnih in skladenjskih redukcijah. To je razumljivo, saj od naštetih treh kategorij posegi na nivoju zapisa najmanj vplivajo na razumevanje besedila.

Precej manj očitna je motivacija za redukcije. Razen očitnih tehničnih bližnjic z izpuščanjem presledkov in večdelnih ločil v analiziranem vzorcu namreč ni bilo mogoče ugotoviti, ali do krajšanja v sporočilih na Twitterju prihaja zaradi zavestnih odločitev uporabnikov, ki zaradi prostorske omejitve želijo skrajšati število znakov, za osebni slog oz. nenamerne izpuste.

V prihodnje bi bilo zanimivo raziskavo dopolniti z analizo drugih žanrov v korpusu Janes in s primerjavo pojavov krajšanja slovenskih uporabnikov, kadar tvitajo v različnih jezikih. Poleg tega bi lahko še primerjali pojave krajšanja v kratkih in dolgih tvitih, s čimer bi dobili vpogled v vpliv tehničnih okoliščin komuniciranja na objavljena sporočila. Druga zanimiva razsežnost je primerjalna analiza krajšanja v sporočilih, zapisanih na mobilnih napravah, s tistimi, ki so ustvarjena na osebnih računalnikih. Prav tako bi bilo pojave krajšanja zanimivo opazovati v različnih časovnih obdobjih, s čemer bi preverili, ali krajšanje narašča oz. upada ter ali postaja bolj oz. manj homogeno. Nenazadnje pa bi lahko raziskavo razširili tako, da bi se posvetili tudi daljšanju v zapisu, ki je usklajen z glasovno realizacijo besed pri prevzemanju iz jezikov z veččrkji (na primer q – ku in x – ks nasproti ch – č).

7 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. Posebna zahvala za pomoč pri oblikovanju tipologije označevanja gre Špeli Arhar Holdt, Jaki Čibeju, Tomažu Erjavcu, Damjanu Popiču in Katji Zupan.

8 Literatura

Christian M. Alis in May T. Lim. 2013. Spatio-Temporal Variation of Conversational Utterances on Twitter http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077793#s3.

Jannis Androutsopoulos in Gurly Schmidt. 2001. SMS-Kommunikation: Ethnografische Gattungsanalyse am Beispiel einer Kleingruppe. V: Meer, D., ur., *Zeitschrift für Angewandte Linguistik*. Bd. 36, Frankfurt/Main, 49–80.

Markus Bieswanger. 2007. abbrevi8 or not 2 abbrevi8: A contrastive analysis of different space and time- saving strategies in English and German text messages. V: *Texas Linguistics Forum*, volume 50.

David Crystal. 2001. *Language and the Internet*. Cambridge: Cambridge University Press.

Jaka Čibej, Darja Fišer in Tomaž Erjavec. (V tisku). Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets.

Nicola D. ring. 2002. Kurzm. wird gesendet – Abkürzungen und Akronyme in der SMS-Kommunikation. V: *Muttersprache - Vierteljahresschrift für deutsche Sprache*, 112 (2), 97–114.

Richard Eckart de Castilho et al. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN V: *Proceedings of the CLARIN Annual Conference (CAC) 2014*. CLARIN ERIC, October 2014. https://www.clarin.eu/content/papers-posters-and-demos-cac2014.

Stephan Gouws et al. 2011. Contextual Bearing on Linguistic Variation in Social Media. V: *Proceedings of the Workshop on Language in Social Media* (LSM 2011). Portland, Oregon. 20–29. http://dl.acm.org/citation.cfm?id=2021113

Ylva Hård af Segerstad. 2002. Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication. Göteborg: Göteborg University.

Susan C. Herring. 2001. Computer-mediated discourse. V: D. Schiffrin, D. Tannen, & H. Hamilton, ur., (pp.*The Handbook of Discourse Analysis* 612-634). Oxford: Blackwell.

Urška Jarnovič. 2007. Diskurzivne značilnosti SMS-ov. *Jezik in slovstvo*, 52 (2): 61–79. Ljubljana: Slavistično društvo Slovenije.

Mojca Kompara. 2009. Prepoznavanje krajšav v besedilih. V: Peter Weiss, ur., *Jezikoslovni zapiski* 15, št. 1–2. 95–112. Založba ZRC, Ljubljana.

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. Zbornik konference RANLP 2015, 7.–9. september 2015, str. 371–378, Hissar, Bolgarija.

Mija Michelizza. 2012. Ločila in druga pisna znamenja v elektronskih besedilih. V: Nataša Jakop, Helena Dobrovoljc, ur., *Pravopisna stikanja: razprave o pravopisnih vprašanjih.* 151–162. Založba ZRC, Ljubljana.

Richard Rogers. 2014. Debanalising Twitter. The transformation of an object of Study. V: Katrin Weller et al., ur., *Twitter and Society*. IX–XXVI. Peter Lang Publishing, Inc., New York.

Peter Schlobinski et al. 2001. Simsen. Eine Pilotstudie zu sprachlichen und kommunikativen Aspekten der SMS-Kommunikation. Networx 22. Retrieved July 1, 2006, from http://www.mediensprache.net/networx/networx-22.pdf.

Malin Sveningsson. 2001. Creating a Sense of Community: Experiences from a Swedish Web Chat. The TEMA Institute, Dept. of Communication Studies. Linköping, Linköping University: 250.

Crispin Thurlow. 2003. Generation Txt? The sociolinguistics of young people's text- messaging. Discourse Analysis Online, 1 (1).

Analiza mnenj s pomočjo strojnega učenja in slovenskega leksikona sentimenta

Klemen Kadunc, Marko Robnik-Šikonja

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
klemzimeister@gmail.com, marko.robnik@fri.uni-lj.si

Povzetek

V prispevku obravnavamo področje tekstovnega rudarjenja, ki se primarno osredotoča na identifikacijo mnenj v besedilih ter njihovo opredeljevanje kot pozitivna ali negativna. V našem delu je bil uporabljen pristop z metodami strojnega učenja, ki smo ga nadgradili z ročno zgrajenim leksikonom sentimenta. Z leksikonom smo v statističnih modelih želeli zaobjeti leksikalno znanje in s tem izboljšati uspešnost klasifikacije. Z evalvacijo rezultatov na primeru spletnih komentarjev nekaterih bolj obiskanih slovenskih spletnih portalov pokažemo, da je pristop uspešen, saj smo klasifikacijo zaznavno izboljšali.

Opinion Mining Using Machine Learning and Slovene Sentiment Lexicon

The article deals with text mining and focuses on identifying opinions in texts and determining their positive or negative character. The research was based on machine learning methods that were upgraded with a manually compiled sentiment lexicon. Its purpose was to encompass lexical knowledge in statistical models and thus improve the classification. Furthermore, the analysis of online comments found on some of the most visited portals indicated that the approach is successful because it managed to significantly improve the classification.

1 Uvod

Analiza mnenj v besedilih (angl. opinion mining), v ožjem kontekstu poimenovana tudi analiza sentimenta (angl. sentiment analysis), je eno od področij tekstovnega rudarjenja (Liu in Zhang, 2012). Ukvarja se z odkrivanjem piščevega mnenja o predmetu pisanja. Večinoma se uporablja polarna analiza, torej določamo pozitivno, negativno in nevtralno mnenje. Naloga za avtomatske sisteme ni enostavna, saj je potrebno iz besedila izluščiti bistveno semantično informacijo, pri tem pa težavo predstavljajo (večkratno) zanikanje, sarkazem, dvoumnost besedila, kontekstna odvisnost rabe besed itd. Pravilno določanje sentimenta je koristno za številne namene, npr. za napovedovanje uspešnosti izdelkov, izidov volitev ali v sociooloških raziskavah. Zaradi vse širšega izražanja mnenj na internetu, preko spletnih forumov, komentarjev novic, tvitov ter recenzij izdelkov in storitev postaja avtomatska analiza mnenj nujen raziskovalni pripomoček.

Analizo sentimenta lahko izvajamo na različnih nivojih, od najbolj splošnega, tj. na nivoju celotnega besedila, do posameznih vidikov oz. značilnosti, ki se v besedilu pojavijo (npr. piščevo mnenje o porabi goriva testnega avtomobila). Pri slednjem nivoju je velika težava že identifikacija entitet in vidikov. V našem delu smo se osredotočili na nivo celotnega besedila. Takšni analizi sentimenta pravimo tudi klasifikacija sentimenta, saj podobno kot pri drugih nalogah klasifikacije besedil (npr. klasifikacija dnevnih novic pod gospodarstvo, šport ipd.), besedilo klasificiramo v eno od sentimentnih kategorij.

Pri analizi sentimenta sta se v grobem uveljavila dva pristopa, leksikalni in pristop s strojnim učenjem. Pri leksikalni metodi potrebujemo enega ali več leksikonov sentimenta, ki vključujejo besede in fraze s pozitivno in negativno konotacijo. Pristop prešteje te besede in fraze v besedilu, ki ga želimo klasificirati. Če prevladujejo besede z negativno konotacijo, besedilo označi kot negativno, sicer kot pozitivno. Težava pristopa je, da je potrebno izdelati leksikon, poleg tega pa se izrazoslovje s časom spreminja, nekatere besede pa imajo različen sentiment v različnih kontekstih (npr. beseda majhen se pri opisu zaslona mobilnega telefona šteje kot negativna oznaka, pri opisu pomnilniškega ključka pa kot pozitivna), zato je ta pristop v praksi v zadnjem času skoraj vedno združen s strojnim učenjem. Pri pristopu s strojnim učenjem s pomočjo učne množice besedil, ki jim priredimo eno od sentimentnih kategorij (npr. pozitivno ali negativno), tvorimo učno množico, na podlagi katere zgradimo klasifikacijski model. Kljub temu, da sentimentni leksikon ni nujno potreben za analizo s strojnim učenjem, so raziskave pokazale, da je dober leksikon koristen in izboljša klasifikacijsko točnost (Hu in Liu, 2004).

Za angleščino obstajajo številna orodja in besedilni korpusi namenjeni analizi sentimenta. Analiza sentimenta v slovenščini je še v povojih in manjkajo še številni viri, npr. slovenski SentWordNet - sentimentno označen WordNet (Baccianella et al., 2010), orodja za pridobivanje in označevanje spletnih virov kot so tviti in spletni komentarji, korpusi sentimentno označenih besedil pa tudi slovenski leksikon polarnih besed potrebuje dodelavo.

Po vzoru leksikona Hu in Liu (2004) smo v tem delu sestavili prosto dostopen leksikon sentimentnih besed. Predstavljamo njegovo evalvacijo v okviru klasifikacije spletnih komentarjev z metodami obdelave naravnega jezika in strojnega učenja. Širša analiza leksikona in označene podatkovne baze, ki jo uporabljamo, je predstavljena v (Kadunc, 2016).

V 2. razdelku pripravimo kratek pregled obstoječih leksikonov sentimenta v slovenskem jeziku ter podrobneje predstavimo izdelan leksikon, v 3. se osredotočimo na zbirko uporabniških komentarjev, ki smo jo sestavili za potrebe evalvacij klasifikacijskih modelov, v 4. si ogledamo dosežene rezultate, jih kritično ovrednotimo ter jih primer-

jamo z rezultati sorodnih raziskav v drugih jezikih. Prispevek zaključimo z omembo glavnih sklepov ter navedemo nekaj možnih izboljšav.

2 Sentimentni leksikon

Sentimentni leksikoni predstavljajo osnovo za sentimentno analizo z leksikalnim pristopom. V strojnem učenju leksikalno znanje običajno vključujemo v fazi priprave značilk. Leksikoni v najosnovnejši obliki sestojijo iz seznama pozitivnih in seznama negativnih besed ali fraz. Naprednejši poleg informacije o polariteti vsebujejo še uteži (npr. beseda je močno pozitivna), oblikoslovne oznake ipd. Gradnja leksikonov poteka na različne načine, od povsem ročnega (drago in zamudno), do polavtomatskega in avtomatskega. Pri naprednejših pristopih se določi začetno množico besed, ki predstavljajo semena za avtomatsko ekspanzijo s pomočjo Wordneta in drugih strukturiranih leksikalnih baz. Podrobneje so avtomatski pristopi opisani v (Potts, 2011). Predvsem za angleščino obstaja veliko število prosto dostopnih leksikonov sentimenta, od splošnih do bolj specializiranih.

Za slovenščino že obstaja nekaj manjših sentimentnih leksikonov. Martinc (2013) je na podlagi seznama AFINN-111 (Nielsen, 2011), ki vsebuje 2477 besed, sestavil seznam polarnih besed ter vsaki priredil vrednost z razponom od -5 (skrajno negativno) do +5 (skrajno pozitivno). Leksikon je uporabil za izdelavo orodja za analizo sentimenta na družbenem omrežju Twitter. Za razliko od Martinca se je Volčanšek (2015) osredotočila na sentimentno analizo bolj formalnih besedil. Za analizo novic je Volčanšek (2015) uporabila leksikalni pristop, njen leksikon pa je osnovan na angleškem slovarju General Inquirer (Stone, 1997), ki poleg seznamov besed vsebuje še dodatne metapodatke, kot je denimo označba vseh kategorij, v katerih se beseda nahaja. Prevedeni slovar teh podatkov ne zajema. Iz kategorij *Positiv* in *Negativ* je z uporabo avtomatskega in ročnega preverjanja sestavila slovenski slovar sentimenta, ki šteje 1669 pozitivnih in 1912 negativnih besed. Rezultati analize na podlagi klasifikacije 5000 novic so bili po mnenju avtorice pod pričakovanji. Kot sentimentni leksikon lahko uporabimo tudi angleški SentiWordNet (Baccianella et al., 2010), ki temelji na leksikalni bazi WordNet. SentiWordNet je povezan z WordNetom in vsakemu vnosu v WordNetu priredi tri numerične vrednosti, s katerimi meri pozitivnost, negativnost ter objektivnost oziroma nevtralnost pojmov. Za razliko od prejšnjih dveh slovarjev, SentiWordNet hrani ocene za različne pomene iste besede. Ker bi bilo brez opisa nemogoče ločevati med posameznimi pomeni, so vnosi oplemeniteni z gloso oz. kratkim opisom pomena za lažje pomensko razdvoumljanje besed. Ker obstaja leksikalna baza WordNet tudi v slovenskem jeziku pod imenom SloWNet (Fišer, 2008) in ker so vnosi povezani z WordNetom, je mogoče sentimentno informacijo iz angleščine prenesti v slovenščino. V naših poskusih smo uporabili tudi to možnost.

Naš primarni leksikon temelji na angleškem leksikonu (Hu in Liu, 2004), ki ga trenutno sestavljata seznama 2006 pozitivnih in 4783 negativnih besed. Za Hujev leksikon smo se odločili zato, ker je bil uporabljen že v vrsti raziskav iz obravnavanega področja ter se stalno posodablja. Osnova sicer izvira iz prve polovice prejšnjega desetletja, ko so se raziskovalci osredotočali predvsem na analizo sentimenta opisov filmov in raznih produktov, kar je razvidno tudi iz samih vnosov. V seznamu so namenoma vključene tudi napačno črkovane in žargonske besede. Izdelava našega leksikona je potekala tako, da smo za osnovo vzeli slovar sentimentnih besed v angleškem jeziku ter ga ročno prevedli v slovenščino s pomočjo spletnih prevajalskih orodij. Vključili smo tudi polarno obarvane sinonime in nekaj različnih oblik iste besede. V slovenščino neprevedljive besede smo izpustili. Zaradi bogate pregibnosti slovenskega jezika je priporočljiva uporaba lematizacije. Naš leksikon trenutno sestavljata seznama 2646 pozitivnih in 6689 negativnih besed. Slovar je prosto dosegljiv v obliki kompresirane datoteke ZIP, ki vsebuje seznam pozitivnih (*positive_words.txt*) ter seznam negativnih (*negative_words.txt*) besed. Posamezne besede so med seboj ločene z znakom za novo vrstico. Vsebina je shranjena v kodnem naboru UTF-8 (Unicode), tako da je uporaba mogoča na večini računalniških platform.

Tako izdelan leksikon ni brez slabosti. Poleg neupoštevanja konteksta uporabe, ki je ena od splošnih slabosti tovrstnih slovarjev, je težava tudi v tem, da smo s prevajanjem iz angleščine izpustili nekatere pogosteje uporabljane slovenske izraze, predvsem tiste, ki se pogosto uporabljajo v neformalni komunikaciji na spletnih omrežjih in za katere ne obstaja neposredni prevod. Leksikon sicer predstavlja dovolj dobro osnovo za nadaljne posodabljanje z novim izrazoslovjem. Omenili smo, da se v slovarju nahaja nekaj različnih oblik iste besede, ki smo jih pri prevajanju dodajali. Postavi se vprašanje o smotrnosti tega početja z ozirom na dejstvo, da za posamezno besedo lahko obstaja tudi več deset različnih oblik. Za določene rabe bi bilo smiselno vse besedne oblike dodati v leksikon, za druge pa vključiti le lemo vsake besede, saj je lematizator za slovenski jezik prosto dostopen z že narejeno podporo za integracijo z več programskimi jeziki (Juršič, 2007). Za uporabnika leksikona tako ne bi smelo biti težav s pretvarjanjem v osnovne oblike besed. Morda bi bilo smiselno v leksikon vključiti še določene metapodatke, denimo podatek o intenziteti pozitivnosti oziroma negativnosti vnosa (npr. z besedo *odličen* lahko boljše identificiramo besedilo kot pozitivno, kot z besedo *dober*).

Kvalitete izdelanega leksikona nismo neposredno ocenjevali ampak smo zgolj merili njegov doprinos k uspešnosti same klasifikacije (več v 4 razdelku). Pri tem smo uporabili večinsko glasovanje. Besede v besedilu smo primerjali z vnosi v leksikonu. Celotno besedilo je bilo pozitivno, v kolikor so večinsko prevladovale besede s pozitivno konotacijo in obratno, besedilo je bilo negativno, v kolikor so večinsko zastopane besede z negativno konotacijo. V kolikor s pomočjo slovarja ni bilo moč pridobiti informacije o polariteti (npr. število pozitivnih besed je enako številu negativnih besed) smo tudi to informacijo uporabili pri klasifikaciji, saj se je izkazalo, da blagodejno vpliva na uspešnost klasifikacije. Dodatnih možnosti, kot je denimo vpeljava praga, s katerim bi nastavili, koliko pozitivnih oz. negativnih besed je potrebno, da ima besedilo s stališča leksikona polariteto, nismo preizkušali.

3 Zbirka spletnih komentarjev

Analiza sentimenta je z razvojem Spleta 2.0 doživela precejšen razmah, saj so raziskovalci dobili praktično neomejene možnosti analiziranja misli uporabnikov, ki jih dnevno delijo preko objav na socialnih omrežjih, (mikro)blogih ipd. Medtem, ko so bili uporabniki na začetku predvsem pregledovalci vsebin, so z razvojem spleta postali tudi njihovi aktivni ustvarjalci. Govorimo o pojmu uporabniško generiranih vsebin (UGV). Z razmahom Spleta 2.0 se je tako fokus raziskav sentimenta iz opisov filmov, produktov ter analize novic prestavil na UGV. Za razliko od formalnih besedil, kjer je pričakovana čistost, slovnična pravilnost ter malo pravopisnih napak, gre pri UGV predvsem za neformalna besedila s samosvojimi zakonitostmi, ki analizo sentimenta dodatno otežijo. Takšna besedila pogosto vsebujejo sarkazem, ironijo, slovnične napake, okrajšave, sleng ter emotikone, s katerimi avtorji še poudarijo svoja občutja o določeni temi. Besedila so pogosto krajša, kar je po eni strani zaradi jedrnatosti prednost, po drugi pa lahko že ena beseda identificira piščevo mnenje.

Raziskovalci se zadnje čase osredotočajo predvsem na analizo sentimenta UGV. Zelo popularna platforma za tovrstne raziskave je družbeno omrežje Twitter, kjer so objave oz. tviti omejeni na 140 znakov, kar uporabnike sili, da svoje misli posredujejo v neposredni, jedrnati obliki. Ker uporabniki običajno tvite oplemenitijo z oznakami (hashtagi), s katerimi primarno označijo temo, na katero se tvit sklicuje, je poenostavljeno tudi pridobivanje tvitov glede na željeno temo (npr. tviti z oznako #SLOprivatizacija zelo verjetno vsebujejo mnenja uporabnikov o slovenski privatizaciji). Vse to in preprost uporabniški vmesnik za dostop do zbirke tvitov so botrovali temu, da je na voljo precejšnje število prosto dostopnih korpusov za analizo sentimenta. Žal to velja le za večje svetovne jezike, kjer prevladuje angleščina. Prosto dostopnega korpusa za slovenski jezik nismo našli, zato smo se odločili, da za ovrednotenje leksikona sentimenta izdelamo svojega. Ena od zahtev je bila, da korpus vsebuje neformalna besedila. Poleg platforme Twitter so bili naravna izbira uporabniški komentarji slovenskih novičarskih portalov. Glede na to, da na platformi Twitter objavljajo tudi uradne entitete, kot so denimo podjetja, smo menili, da bo identifikacija subjektivnega besedila lažja. Hkrati pa smo se zavedali, da komentarji pogosto niso v skladu z novico (angl. offtopic) ter da je identifikacija mnenja težavnejša zaradi tega, ker se lahko v enem komentarju prepleta več različnih mnenj. Nasploh je analiza uporabniških komentarjev ena težjih nalog s tega področja.

Pri izdelavi korpusa smo uporabili spletne komentarje iz portalov 24ur, Finance, Reporter in RtvSlo. S pomočjo ključnih besed (npr. "trg nepremičnin raste", "begunska kriza"ipd.) in prilagojenih Googlovih iskalnih pogonov, s katerimi smo iskali izključno po naštetih spletnih portalih, smo avtomatizirano pridobili spletne povezave na novice ter iz njih izluščili komentarje. Popoln nabor uporabljenih iskalnih pogojev pri gradnju korpusa je objavljen v (Kadunc, 2016). Naj omenimo, da nekateri slovenski novičarski portali vodijo vse bolj restriktivno politiko komentiranja. V času izdelave korpusa tako portal Dnevnik ni več omogočal neposrednih komentarjev uporabnikov, SiolNet je po drugi strani hranil komentarje le za zadnjih 7 dni ter tako funkcionalnost komentiranja naredil povsem neuporabno za naš eksperiment. Skupaj smo pridobili 5087 uporabniških komentarjev iz 427 različnih strani omenjenih spletnih virov, v povprečju z vsake strani 11 komentarjev, od tega je bilo 4777 uporabnih (ostali so bili npr. v tujem jeziku ali pa so vsebovali le sliko). Komentarje smo uvrstili v eno od tematik: šport, politika, gospodarstvo in drugo.

Ko smo pridobili željene komentarje, jih je bilo potrebno označiti s sentimentnimi ocenami, z namenom pridobitve zadostnega števila primerov za učenje in testiranje klasifikatorjev. Za označevanje primerov je na voljo več tehnik, od povsem ročnih, do avtomatskih. V našem delu so bili vsi primeri označeni ročno, s strani človeških označevalcev. Komentarje so označevali trije označevalci (angl. annotators). Za označevanje smo poleg osnovnih sentimentnih kategorij *pozitivno*, *negativno* ter *nevtralno*, določili še *irelevantno*. S slednjo kategorijo smo želeli označiti ter tako iz končne verzije korpusa izločiti komentarje, ki za samo analizo niso relevantni, npr. vsebujejo le sliko, povezavo, so napisani v tujem jeziku ipd. Vsi označevalci so videli isti nabor komentarjev. Vsak komentar je bil označen natanko trikrat. Stopnjo strinjanja dveh označevalcev smo izmerili s statistično mero Cohen Kappa, ki upošteva tudi morebitno naključno strinjanje. Tako $\kappa = 1$ pomeni popolno strinjanje med označevalcema, $\kappa \leq 0$ pa ujemanje, ki ni večje od naključnega. Pri nas strinjanje med prvim in drugim označevalcem znaša $\kappa = 0,33$, med prvim in tretjim $\kappa = 0,31$ ter med drugim in tretjim $\kappa = 0,54$. Stopnjo strinjanja med vsemi tremi označevalci smo izračunali z mero Fleiss Kappa. Z doseženo vrednostjo $\kappa = 0,38$ smo po interpretaciji iz Landis in Koch (1977) dosegli ustrezno ujemanje, tako da bi moral biti korpus dovolj zanesljiv za uporabo. Vpliva zanesljivosti korpusa na uspešnost klasifikacije sicer nismo merili. To bi lahko naredili tako, da bi za učenje in testiranje klasifikatorjev vzeli samo tiste komentarje, pri katerih je bilo med označevalci popolno strinjanje . Omeniti je potrebno, da gre pri označevanju sentimenta za močno subjektivno nalogo, posledično je visoko stopnjo strinjanja med označevalci težko pričakovati.

Korpus je na voljo v uravnoteženi in neuravnoteženi obliki. Končna verzija korpusa brez uravnoteženja vsebuje 898 pozitivnih, 3291 negativnih ter 588 nevtralnih komentarjev, določenih z večinskim strinjanjem označevalcev. Podrobnejši razpored primerov po posameznih kategorijah novic in spletnih virih je prikazan v tabeli 1. Vidimo lahko, da so komentarji pretežno negativno nastrojeni, razen za kategorijo *šport*, kjer je razmerje med pozitivnimi in negativnimi približno uravnoteženo. Za učenje smo uporabljali tudi uravnoteženi korpus, pri katerem smo naključno izbrali po 580 komentarjev vsake vrste sentimenta. Korpus je prosto dostopen v obliki kompresirane datoteke ZIP, ki vsebuje korpus v formatu XML ter opis strukture dokumenta XML. Za XML smo se odločili zaradi velike razširjenosti uporabe tega formata, ki predstavlja de facto standard za izmenjavo podatkov preko interneta.

4 Evalvacija z metodami strojnega učenja

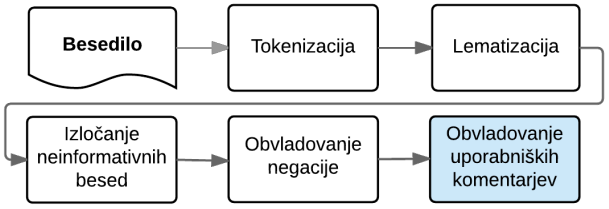
Sentimentni leksikon smo preizkusili v kontekstu strojnega učenja in predobdelave besedil. Izvedli smo vrsto ek-

	gospodarstvo	politika	šport	drugo	RtvSlo	24ur	Finance	Reporter	skupaj
pozitivno	129	26	679	64	566	255	54	23	898
nevtralno	262	33	240	53	441	48	75	24	588
negativno	1420	351	882	638	1614	584	554	539	3291
skupaj	1811	410	1801	755	2621	887	683	586	4777

Tabela 1: Razporeditev komentarjev po kategorijah novic (levo) in po spletnih virih (desno).

sperimentov, s katerimi smo merili vpliv predobdelave in izbire atributov na uspešnost klasifikacije. Zlati standard za naše eksperimentiranje je predstavljal uravnoteženi korpus z enakomerno zastopanostjo komentarjev vsake izmed treh sentimentnih kategorij. Za preverjanje uspešnosti klasifikacije smo uporabili 10-kratno prečno preverjanje. Vsi navedeni rezultati predstavljajo povprečje 10-ih iteracij. Kot klasifikatorje smo izbrali logistično regresijo (LR), metodo podpornih vektorjev (SVM), večvrednostni naivni Bayesov klasifikator (MNB) in binarni naivni Bayesov klasifikator (BNB). Navedeni klasifikatorji se pri analizi sentimenta tudi sicer največ uporabljajo. Kot mere uspešnosti smo izbrali klasifikacijsko točnost (CA) ter za vsako vrsto sentimenta posebej še mero F_1. Naj omenimo, da smo izvajali izključno trirazredno klasifikacijo, tj. klasifikacijo v eno izmed treh sentimentnih kategorij. Veliko sorodnih raziskav se namreč osredotoča le na kategoriji *pozitivno* in *negativno*. Z ozirom na uravnoteženi korpus vrednost CA pri 33,33% predstavlja spodnjo mejo še sprejemljive klasifikacije točnosti. Kot osnovo smo določili konfiguracijo, pri kateri smo za atribute izbrali posamezne besede (unigrami), skupaj jih je bilo 20.388. Z osnovno konfiguracijo smo dosegli vrednost CA 54,5%, kar je 21,2% nad vsakokratno izbiro večinskega razreda.

Najprej smo primerjali različne načine predobdelave besedil (diagram na sliki 1). S predobdelavo želimo besedilo pripraviti in ga očistiti. Pri uporabniških komentarjih lahko pričakujemo veliko šuma, zato je predobdelava še posebej pomembna. Preizkusili smo več načinov predobdelave, od bolj splošnih, ki so široko uporabljeni pri obdelavi naravnega jezika, kot so denimo tokenizacija in lematizacija, do bolj specifičnih, npr. obvladovanje negacije.



Slika 1: Koraki pri predobdelavi vhodnega besedila.

Najprej smo merili vpliv različnih načinov tokenizacije. Primerjali smo tokenizacijo s presledki, Treebank tokenizacijo in Pottsovo tokenizacijo, najboljše se je povečini izkazala slednja. Lematizacija je izboljšala, izločanje neinformativnih besed in enostavno obravnavanje negacije pa sta poslabšali delovanje vseh klasifikatorjev. Za uporabniške komentarje specifični načini predobdelave (npr. zamenjava spletnih povezav s pojavnico *URL* ali zamenjava emotikonov s sentimentno oceno) so povečini izboljšali klasifikacijo. Iz tabele 1 je razvidno, da smo s predobdelavo, v primerjavi z osnovno konfiguracijo, pridobili 8,7% pri CA. Naj omenimo, da smo poleg izboljšanja klasifikacije uspeli zmanjšati tudi število atributov, na 9.198, kar je več kot 50% manj kot pri osnovni konfiguraciji.

Klasifikator	CA	mera F_1			
		pos	*neg*	*neu*	povp.
osnova (LR)	54,5	57,6	51,5	54,3	54,5
LR	61,8	66,8	58,6	60,1	61,8
SVM	59,7	64,6	55,9	58,4	59,6
MNB	**63,2**	67,3	64,2	57,6	**63,0**
BNB	48,9	57,5	44,4	36,4	46,1

Tabela 2: Stanje konfiguracij po predobdelavi besedila. Podani so rezultati klasifikacijske točnosti ter mer F_1 za različne klasifikacijske metode. Najboljša konfiguracija je predstavljena z odebeljeno pisavo.

Nadalje smo preizkusili različne tehnike za izločanje, izbiro in uteževanje značilk. Z višjimi N-grami smo želeli preizkusiti, ali bi zajetje širšega konteksta, kot so npr. fraze, lahko pripomoglo k izboljšanju rezultatov. Če smo kot atribute dodali še bigrame, se je uspešnost nekoliko izboljšala le pri metodi SVM, kar je razvidno iz tabele 3. Dodajanje trigramov je poslabšalo rezultate vseh metod. Tudi z omejevanjem števila značilk glede na pogostost pojavitve v korpusu pri višjih N-gramih nismo dosegli večjih razlik. Da se unigrami zelo dobro obnesejo pri analizi sentimenta so pokazali že v raziskavi (Pang et al., 2002), kjer so analizirali opise filmov.

Model	LR	SVM	MNB	BNB
unigrami	**61,8**	59,7	**63,2**	**48,9**
unigrami + bigrami	61,2	**60,1**	59,0	43,3
bigrami	51,0	49,5	51,6	38,3
uni + bi + trigrami	60,6	59,4	56,4	39,4

Tabela 3: Vpliv različno visokih N-gramov na uspešnost klasifikacije. Odebeljeni so najboljši modeli za posamezno metodo.

Uteževanje značilk (pogostost, frekvenca, tf-idf) je dalo mešane rezultate (tabela 4); tako je pogostost koristila klasifikatorju MNB, uteževanje s tf-idf pa SVM, ki se je zelo približal ostalima dvema metodama. Naj navedemo, da je Smailović (2014) v sorodni raziskavi pri klasifikaciji tvitov s klasifikatorjem SVM prišla do zaključka, da se preprostejša utež tf obnese bolje od tf-idf.

Izbira podmnožice pomembnih atributov z metodo *Hi-kvadrat* je koristila klasifikatorju BNB, kjer smo dosegli izboljšanje CA za 3%. Pri ostalih večjega uspeha nismo zabeležili.

Vektorizacija značilk	LR	SVM	MNB
prisotnost (dvojiška vrednost)	**62,9**	60,2	62,1
pogostost (štetje)	61,8	59,7	**63,2**
utež TF	58,7	61,0	56,6
utež TF-IDF	61,7	**62,6**	61,3

Tabela 4: Primerjava načinov vektorizacije značilk. BNB smo izpustili zaradi definicije Bernoullijevega modela.

V klasifikacijskih modelih smo leksikalno znanje uporabili na način, da smo primerjali besede v besedilu z vnosi v leksikonu ter dodali ustrezno značilko. V kolikor so prevladovale besede s pozitivno konotacijo smo dodali značilko *oznakaSlovarja_POS*, v kolikor so prevladovale negativne besede značilko *oznakaSlovarja_NEG* in *oznakaSlovarja_NEU* za primere, ko iz leksikona ni bilo mogoče ugotoviti polaritete besedila. Rezultate združevanja različnih leksikonov in klasifikatorjev prikazuje tabela 5. Najboljše se je obnesel naš leksikon (KSS). Z njim smo dosegli zaznavno izboljšanje klasifikacije pri vseh metodah strojnega učenja, v primeru klasifikatorja MNB za 1,3%. Z leksikonom General Inquirer (GIS) smo dobili mešane rezultate. Najslabše se je odrezal leksikon avtomatsko pridobljen iz kombinacije SentiWordNeta in SloWNeta (SWN), razloge za to gre lahko iskati v tem, da SWN različne pomene iste besede točkuje z različnimi sentimentnimi ocenami. Za učinkovito izrabo leksikona SWN bi bilo potrebno vključiti sistem razdvoumljanja večpomenskih besed, kar pa bi znalo predstavljati težavo, saj v SloWNetu manjkajo glose in primeri uporabe za precejšnje število vnosov. Tudi z vključitvijo vseh leksikonov skupaj v povprečju nismo uspeli bistveno izboljšati rezultatov leksikona KSS.

Model	LR	SVM	MNB	BNB
unigrami	61,8	59,7	63,2	48,9
unigrami + KSS	**62,9**	**60,6**	64,5	49,8
unigrami + GIS	61,5	59,5	64,4	49,4
unigrami + SWN	61,0	59,8	63,4	49,2
vsi skupaj	62,2	60,5	**65,2**	**50,3**

Tabela 5: Vpliv leksikonov sentimenta na klasifikacijo. Podani so rezultati klasifikacijske točnosti za različne klasifikacijske metode pri različnih sentimentnih leksikonih. Odebeljeni so najboljši rezultati za vsako metodo.

Najboljša konfiguracija, ki vključuje vse koristne predobdelave, leksikone sentimenta in uporablja večvrednostni naivni Bayesov klasifikator, doseže na uravnoteženem besedilu 65,5% klasifikacijsko točnost, kar predstavlja 11% izboljšanje glede na osnovno konfiguracijo in 32,2% izboljšanje v primerjavi z vsakokratno izbiro večinskega razreda. Poglejmo si še primerjavo med najboljšima klasifikatorjema. Razlika v CA znaša slaba 2%, standardni odklon ali standardna deviacija (σ) za logistično regresijo znaša 3,79%, za večvrednostni naivni Bayesov klasifikator je bil izmerjen σ = 3,75%. Glede na to, da smo delali na splošnem klasifikacijskem modelu, smo z rezultati zadovoljni. Za klasifikatorja MNB in LR smo z uporabo Wilcoxonovega testa (Demšar, 2006) izračunali še statistično značilnost razlik. Pri stopnji značilnosti α = 0,05 ničelne hipoteze, da med klasifikatorjema ni razlik, ne moremo zavrniti.

Klasifikator	CA	mera F_1			
		pos	*neg*	*neu*	povp.
osnova	54,5	57,6	51,5	54,3	54,5
LR	63,6	68,1	61,3	61,6	63,7
SVM	63,2	69,0	62,1	58,6	63,2
MNB	**65,5**	68,6	66,8	60,6	**65,3**
BNB	60,1	65,0	56,7	58,4	60,0

Tabela 6: Izbira najboljše metode strojnega učenja za klasifikacijo komentarjev na uravnoteženem korpusu.

Najboljšo konfiguracijo smo preizkusili še v kontekstu neuravnoteženega korpusa. Porazdelitev komentarjev po razredih v tem primeru je: 588 nevtralnih, 898 pozitivnih ter 3291 negativnih komentarjev. Z vsakokratno izbiro večinskega razreda bi dobili 68,9% CA. Model, ki smo ga zgradili na neuravnoteženem korpusu doseže 76,2% klasifikacijsko točnost, mera F_1 na pozitivnih primerih daje vrednost 60,0%, na negativnih pa 85,4%.

4.1 Primerjava z rezultati raziskav za druge jezike

V preteklih letih je bilo narejenih veliko raziskav za večje svetovne jezike, predvsem angleščino. Rezultate raziskav, celo znotraj istega jezika, je medsebojno težko neposredno primerjati. Upoštevati je potrebno več faktorjev, ki lahko bistveno vplivajo na interpretacijo in primerjavo rezultatov posameznih raziskav. Med njimi so vrsta besedila, razrednost klasifikacije, uporabljen korpus ipd. Za angleški jezik, v obliki spletnih storitev, obstaja nekaj javno dostopnih splošnih klasifikatorjev sentimenta, kot je denimo AlchemyAPI[1]. Če bi delali na analizi sentimenta angleških besedil, bi lahko na testnih podatkih klasifikatorje preizkusili in okvirno ocenili, kako se naš klasifikator primerja z drugimi. Žal za slovenski jezik ni na voljo tovrstnih, prosto dostopnih storitev.

Vrsta besedila predstavlja pomemben dejavnik pri vrednotenju rezultatov. V času pred ekspanzijo (mikro)blogov so bili med raziskovalci priljubljeni opisi filmov in raznih produktov. Pri opisih filmov so raziskovalci dosegli rezultate, primerljive z rezultati kategorizacije novic (šport, gospodarstvo ipd.). Abbasi et al. (2008) so v raziskavi nad korpusom opisov filmov (Pang et al., 2002), ki je bil predmet številnih raziskav, pri dvorazredni klasifikaciji dosegli 91,7% CA. Takšnih rezultatov pri bolj neformalnih besedilih (tviti, uporabniški komentarji ipd.) klasifikatorji ne dosegajo. Prav tako uspešnost klasifikacije upade, če poleg kategorij pozitivno in negativno, rešujemo še problematiko nevtralnosti besedila. Smailović (2014) je primerjala nekaj prosto dostopnih klasifikatorjev na množici ročno označenih testnih besedil in pri nekaterih v primeru trirazredne klasifikacije zmogljivosti precej upadejo. Poglejmo nekaj sorodnih raziskav v angleškem jeziku, ki se

[1] http://www.alchemyapi.com/.

osredotočajo na trirazredno analizo sentimenta neformalnih besedil.

Agarwal et al. (2011) so analizo sentimenta izvajali nad 1709 ročno označenimi tviti, enakomerno razporejenimi po vseh treh kategorijah. Z najboljšo konfiguracijo so dosegli 60,83% CA, kar je nekaj slabše kot v primeru našega klasifikatorja. Kot zanimivost naj omenimo, da je tudi pri njih kot osnova služila konfiguracija z unigrami, s katero so dosegli 56,58% CA. V primerjavi z našo raziskavo jim je torej uspel manjši dvig uspešnosti najboljše konfiguracije glede na osnovo.

Prav tako so nad označenimi tviti analizo sentimenta izvajali Hamdan et al. (2013). Z najboljšo konfiguracijo so uspeli doseči 58,87% CA. Raziskava je zanimiva, ker so preizkusili sentimentni leksikon SentiWordNet in, kot mi, prišli do zaključka, da lahko leksikalni viri izboljšajo klasifikacijo.

V okviru tekmovanja SemEval-2013 so v kategoriji analize sentimenta sporočil v sistemu Twitter najboljši klasifikator izdelali Kiritchenko et al. (2014). Dosegli so povprečje mer F_1 pri 69,02% (uporabljen je bil neuravnotežen korpus tvitov, za primerjavo klasifikatorjev so uporabili povprečje mer F_1, klasifikacijske točnosti niso merili). Zmagoviti klasifikator temelji na kombinaciji metod strojnega učenja in intenzivni rabi različnih leksikonov, med drugim so vključili tudi Hujev leksikon sentimenta, ki je tudi nam služil kot osnova za izdelavo slovenskega leksikona sentimenta.

V zgoraj naštetih in mnogih drugih raziskavah se CA pri analizi sentimenta neformalnih angleških besedil giblje med 60% in 70%. Tudi naši rezultati na slovenskih besedilih so na tem intervalu, zato menimo, da smo lahko z njimi zadovoljni.

5 Zaključki

V prispevku smo predstavili rezultate analize sentimenta uporabniških komentarjev v slovenskem jeziku z uporabo leksikalnih virov v kontekstu nadzorovanega učenja. Sklenemo lahko, da leksikalni viri pozitivno vplivajo na analizo sentimenta z uporabo metod strojnega učenja. Najboljša konfiguracija, ki smo jo preizkusili, bistveno preseže klasifikacijo v večinski razred in vse osnovne konfiguracije. Leksikon, ki smo ga izdelali, je javno dostopen[2] in lahko koristno služi pri nadaljnih analizah mnenj v slovenskem jeziku. Možne so še številne izboljšave, predvsem v povezavi s kakovostnimi viri, kot so enojezični in večjezični slovarji ter avtomatsko določanje sentimentnih besed za posamezne kontekste.

Klasifikacijske modele bi lahko še dodatno izboljšali z uvedbo naprednih tehnik, kot je denimo uporaba bolj ali manj zahtevnih lingvističnih pravil za obravnavanje negacije (npr. negacija besede z negativno konotacijo sentiment spremeni v pozitiven). Običajno tovrstne raziskave vključujejo tudi oblikoslovno označevanje. V našem delu smo ga v celoti prezrli, tako da je to dobra iztočnica za nadaljne delo. Z oblikoslovnim označevanjem se odpira vrsta dodatnih možnosti pri pripravi značilk, npr. omejitev unigramov na pridevnike. Za izboljšanje leksikona SWN bi potrebovali sistem za razdvoumljanje večpomenskih besed. Prav tako so odprte možnosti pri izboljšavah korpusa uporabniških komentarjev. Korpus bi lahko razširili z novimi primeri, nekateri tuji tovrstni korpusi vsebujejo tudi več 10 tisoč označenih primerov. Kot je razvidno iz tabele 1, smo zadovoljivo pokrili predvsem kategorijo *šport*. Uporabniške komentarje bi lahko črpali iz raznovrstnejših spletnih virov, morda v korpus dodali tudi objave iz socialnih omrežij ipd.

[2]Korpus označenih uporabniških komentarjev ter slovenski leksikon sentimenta sta dostopna na naslovu http://lkm.fri.uni-lj.si/rmarko/repozitorij/opinionLexicon.

6 Literatura

Ahmed Abbasi, Hsinchun Chen in Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions On Information Systems*, 26(3).

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow in Rebecca Passonneau. 2011. Sentiment analysis of twitter data. V: *Proceedings of the Workshop on Languages in Social Media*, LSM '11, str. 30–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stefano Baccianella, Andrea Esuli in Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. V: *Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, zvezek 10, str. 2200–2204.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December.

Darja Fišer. 2008. Using multilingual resources for building SloWNet faster. V: *The Fourth Global WordNet Conference*, str. 185–193.

Hussam Hamdan, Frederic Béchet in Patrice Bellot. 2013. Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging. V: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, str. 455–459, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Minqing Hu in Bing Liu. 2004. Mining opinion features in customer reviews. V: *Proceedings of AAAI Conference on Artificial Intelligence*, zvezek 4, str. 755–760.

Matjaž Juršič. 2007. Implementacija učinkovitega sistema za gradnjo, uporabo in evaluacijo lematizatorjev tipa RDR. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Diplomsko delo.

Klemen Kadunc. 2016. Določanje sentimenta slovenskim spletnim komentarjem s pomočjo strojnega učenja. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Diplomsko delo.

Svetlana Kiritchenko, Xiaodan Zhu in Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *J. Artif. Int. Res.*, 50(1):723–762, May.

J. Richard Landis in Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Bing Liu in Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. V: *Mining text data*, str. 415–463. Springer.

Rok Martinc. 2013. Merjenje sentimenta na družabnem omrežju Twitter: izdelava orodja ter evaluacija. Univerza v Ljubljani, Fakulteta za družbene vede. Magistrsko delo.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. V: *Proceedings of the 8th European Semantic Web Conference Workshop on 'Making Sense of Microposts': Big things come in small packagess*, str. 93–98. http://arxiv.org/abs/1103.2903.

Bo Pang, Lillian Lee in Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. V: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, str. 79–86.

Christopher Potts. 2011. Sentiment symposium tutorial: Lexicons. http://sentiment.christopherpotts.net/lexicons.html.

Jasmina Smailović. 2014. *Sentiment analysis in streams of microblogging posts*. Doktorsko delo, International postgraduate school Jožef Stefan, Ljubljana, Slovenia.

Philip J Stone. 1997. Thematic text analysis: New agendas for analyzing text content. V: Carl Roberts, ur., *Text Analysis for the Social Sciences*, str. 33–54. Lawrence Erlbaum Associates Publishers, Mahwah, NJ.

Mateja Volčanšek. 2015. Leksikalna analiza razpoloženja za slovenska besedila. Univerza v Ljubljani, Fakulteta za računalništvo in informatiko. Diplomsko delo.

Konferenca
Jezikovne tehnologije in digitalna humanistika
Ljubljana, 2016

Conference on
Language Technologies & Digital Humanities
Ljubljana, 2016

Building a Gold Standard for Temporal Entity Extraction from Medieval German Texts

Natalia Korchagina

Institute of Computational Linguistics
University of Zurich
Binzmühlestrasse 14, 8050 Zürich
korchagina@cl.uzh.ch

Abstract

We present a corpus of gold standard annotation of temporal entities in Early New High German texts. This resource addresses the lack of a gold standard temporal annotation for historical German. Such a corpus is necessary for our research. The ultimate goal of our project is to develop an effective system for temporal entity extraction from historical texts. The manually annotated corpus will serve as base for quality estimation of the temporal annotation produced during the experiments.

1. Introduction

Time is a crucial dimension not only in information processing, but in humanities as well, e.g., a description of a person or a place should contain temporal terms. Not only modern texts, but also historical texts may be rich in temporal expressions. Manual extraction of this information is time-consuming, therefore some facts might still be undiscovered, and thus unknown to scientists. This research will contribute to the development of a tool for temporal entity extraction from historical texts, assisting historical text-mining.

Typical application examples exploiting temporal tagging include information extraction, i.e., the described events are summarized and chronically ordered; and information retrieval, where time is used as a query topic. Temporal annotation of historical texts would allow the digital humanities community to benefit from both scenarios, enabling a faster analysis and a time-framed search through the ever-growing amount of historical corpora available in digital form.

The project is funded by the Swiss Law Sources Foundation. As material for our research, we use historical legal texts (i.e., decrees, regulations, court transcripts) kindly provided by the Foundation. This organization has been publishing critical editions of Swiss historical legal texts in German, French, Italian, Romansh, and Latin for over a hundred years. By today 28 of 118 published volumes are available for digital processing, with a roughly estimated total of 7 million tokens of historical data. The texts' creation time ranges from the 10th to the 18th century. The biggest part of the available digital data is in German, therefore in this project we work with German texts.

There are systems for temporal information extraction from modern texts. Effective taggers such as SUTime (Chang and Manning, 2012) and HeidelTime (Strötgen and Gertz, 2013) use handcrafted rules and dictionaries for recognition and normalisation of temporal expressions. Applied to a corpus of modern narrative texts, WikiWarsDE, HeidelTime achieves f-scores of 91.3 and 85.8 for the extraction (lenient and strict, respectively) (Strötgen and Gertz, 2011). However, the application of an off-the-shelf tool developed for a modern language to a historical corpus is unlikely to lead to good results. Scheible et al. (2011) evaluated the TreeTagger (Schmid, 1994) developed for modern German on Early Modern German corpus, achieving a tagging accuracy of 69.6%, which is far from the 97% reported for modern German.

Lexical and spelling differences are some of the most evident properties of historical texts and a substantial obstacle to the application of the existing NLP tools. The example below shows some of the manually extracted expressions meaning or referring to "evening" ("Abend" in modern German).

Abend	abentt	stübgloge
abende	abentts	zenacht
abends	abent	gessen[2]
abendes	aebent	zenacht essen
ábend	aebents	znacht essen
aubent	abentz	Nacht essen
aubend	stübglogge[1]	nachtessen
aubends	stübgloggen	schlaff trunck[3]

Example 1: Expressions in medieval German with the meaning "evening".

The most common approach for dealing with the non-standard spelling is normalisation, i.e., the process of mapping historical word forms to their modern equivalents. After spelling normalisation, expressions in the range of "Abend" – "abendes" in the example above will be recognized, while those like "stübglogg" will remain unidentified because they are no longer used in temporal context or disappeared from the modern language, and thus there is no pattern to be

[1] Betzeitglocke am Abend [*en*: Bedtime bells ringing in the evening]. Schweizerisches Idiotikon – Wörterbuch der schweizerdeutschen Sprache, "S t äu b (i) g l o g g (e^n)" (II, Sp. 617), 1885.

[2] Abendbrot [*en*: supper]. Schweizerisches Idiotikon – Wörterbuch der schweizerdeutschen Sprache, "Z (e) n a c h t ë s s e^n" (I, Sp. 527), 1881.

[3] Trunk, <...>, vor dem Schlafgehen eingenommen [*en*: Nightcap, bedtime drink]. Schweizerisches Idiotikon – Wörterbuch der schweizerdeutschen Sprache, "S c h l ä f f t r u n k" (XIV, Sp. 1212), 1987.

matched in the set of rules. To overcome these limitations, at the second stage of our experiments we will use statistical methods capable to learn possible patterns of temporal expressions from a manually annotated Gold Standard corpus.

This paper describes the creation of a Gold Standard sample corpus (of about 32,000 tokens) of Early New High German containing manual annotations of temporal entities. This corpus is used in our research as base for quality estimation of temporal tagging at all stages of our experiments. Section 2 introduces the contents and design of the corpus. In Section 3, we will describe the annotation process and summarize our experience of the adaptation of the annotation guidelines for historical data. The first stage of experiments based on the corpus will be presented in Section 4.

2. Corpus Design

Two major types of documents are present in our data: legal cases and transactional documents. Legal cases describe incidents of the law violation and legal consequences that followed. Documents of this time contain, e.g., date and time when the event took place. Transactional documents represent contracts, sales agreements, and purchases. They are especially reach in temporal information, important for the legal value of the document. Schilder and McCulloh (2005) mention the following kinds of temporal information in transactional documents: the date when the transaction takes effect, the execution date, and duration clauses.

For the Gold Standard annotation, we selected manually 50 articles, corresponding to various kinds of legal documents described above. The texts were taken from 9 volumes, representing 5 Swiss cantons. This set of texts covers the period between 1450 and 1550. This particular period of time was chosen, first, because of a large number of articles created at this period (total of 4,175 articles were created between 1450 and 1550), available in digital format in the collection of the Swiss Law Sources Foundation. If our preliminary experiments will prove to be effective, larger datasets from the same period may be involved for further experiments. Figure 1 shows the distribution of the number of articles regarding the year they were issued.

Although the biggest amount of the articles belongs to the year 1425, after a closer examination of the contents of these texts we opted for a later period of time. In order to create a properly annotated Gold Standard corpus, it is important for annotators and supervisors of the task to understand well the contents of the corpus. Articles written before the second half of the 15th century were very complicated to understand even for native speakers of German, therefore the second reason for our choice of period is the language state.

The 50 chosen articles are relatively evenly distributed between 1450 and 1550. The choice of material was motivated by the idea to optimize our system for work on diachronically close texts, yet capturing a certain variety in language state due to their diverse origins. We realize that a system adapted for recognition of temporal expressions in the material from a particular period will show lower results, if applied to a text from another period of German, as spelling is period dependent. Our research should be seen as experimental ground, attempting to find a state-of-the-art method for recognition of temporal information in historical texts.

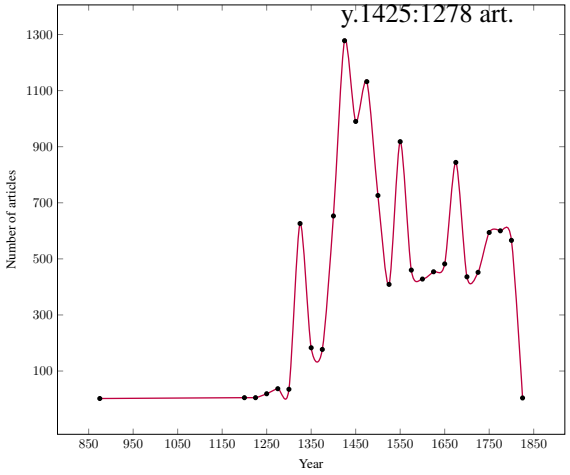


Figure 1: Distribution of the articles available in digital format with regard to the year they were issued.

Period	1450 – 1550
Language period	Early New High German
Domain	legal
Number of articles	50
Aver. length of an art., tokens	950
Total, tokens	32338

Table 1: Characteristics of the Gold Standard corpus.

3. Annotation Process and Results

3.1. Dataset Annotation

To facilitate the annotation task for the human annotators, the corpus was first processed with the rule-based temporal tagger HeidelTime adapted for the Text+Berg corpus (Rettich, 2013), (Volk et al., 2010) containing Swiss alpine texts from 1864 to 2009. Instead of the default configuration for German, the adapted version of the tagger was used because it covers some diachronic variation since the 19th century, and thus the chances of a successful extraction of temporal expressions were higher. When HeidelTime with the adapted set of the resources was applied to our corpus, 200 text segments were identified as temporal expressions.

The automatically annotated corpus was verified manually by two annotators. Their task was to correct erroneous annotations and add missing tags. For our annotation, we adopted a customized XML-like format based on TimeML language (Pustejovsky et al., 2003). It is a robust specification language for events and temporal expressions in natural language text. Several tags and their attributes are defined in TimeML adressing event markup, time stamping of events, ordering of events in time, reasoning with contextually underspecified temporal expressions (e.g., “ten days”) and reasoning about the persistence of events.

Since the work time was limited to 40 hours, it was important to provide annotators with concisely written guide-

lines, strictly relevant to their assignment. Although our instructions were based on the existing guidelines (Saurí et al., 2006) for temporal and event annotations in TimeML standard, it was not possible to reuse them entirely. First, these guidelines are very detailed: they contain descriptions of many attributes and tags which may be of use for the development of a more complex system. Second, being developed for the annotation of modern texts, the TimeML guidelines do not reflect particularities of historical corpora.

Our guidelines covered the two most important points: what to annotate and how. First, different kinds of temporal expressions (explicit/implicit/relative/markable/non-markable) were introduced. According to the TimeML annotation guidelines, only markable expressions (which can be situated on a timeline) should be annotated. Non-markable expressions are less amenable to being situated on a timeline, e.g., later, previous, sooner. It is difficult to understand old German texts even for a native speaker of German. In order not to miss a markable expression, considering it to be a non-markable one, the annotators were asked to tag any expressions with temporal semantics.

The annotators used a subset of the TimeML mark-up language, as it was implemented in HeidelTime, i.e., temporal expressions were tagged with TIMEX3 tags. In addition, the annotators were to put SIGNAL tags to mark a token signaling a relationship between two temporal expressions, and non-consuming (meta-tags, not containing any text) TIMEX3 tags detailing this relationship. These features also belong to the TimeML language. According to our guidelines, each TIMEX3 tag should contain three compulsory attributes: ID number, type and value. The TimeML standard distinguishes four different types of temporal expressions:

- DATE, for expressions describing a calendar date, e.g., December 25, 2015;
- TIME, for expressions referring to a time of the day, e.g., half past midnight;
- DURATION, for expressions describing a duration, e.g., three days;
- SET, for expressions describing a set of times, e.g., twice a week.

The value attribute specifies which temporal information is contained in the tagged span of text. In the TimeML guidelines, the value attribute should be in the form of an ISO8601 format for date and time, supposing that only markable expressions are to be annotated. Following our guidelines, the annotators were allowed to underspecify the value attribute for cases when the meaning of an expression is not entirely clear, e.g., use *"XXXX-12-25"* for "Christmas" of a year unknown from the context, or *"XXXX"* for non-markable expressions. Annotators were asked to pay special attention to the tagging of the saint feast days. They represent a large part of the temporal information in historical texts from this period and were often used instead of the calendar dates. Due to the spelling variation none of these expressions was detected during the automatic annotation using HeidelTime. Annotators were asked to assign a value attribute in ISO8601 format for fixed feasts (they normally refer to a feast of a particular saint). As for moveable feasts, i.e., relative to the Easter Sunday of a particular year, annotators could underspecify the value attribute.

old de: Verkunden uff den 8 januarii anno etc. 1545
mod. de: Verkünden auf den 8 Januar Jahr etc. 1545

en: 'Announced on the 8 January year etc. 1545'

Verkunden uff den <TIMEX3 tid="t620"
type="DATE" value="1545-01-08" >
8 januarii anno etc. 1545 </TIMEX3>.

Figure 2: Sentence in historical German (old de), its modernised spelling (mod. de), translation into English (en) and temporal annotation.

The example above shows an annotated sentence from the Gold Standard, preceded by a gloss pairing the original sentence in Early New High German with its modern equivalent and a translation into English.

3.2. Dataset Analysis

After the annotation process was finished, we calculated the inter-annotator agreement. We present values for average observed agreement and chance-corrected agreement (Cohen's Kappa) in Table 2. Relaxed matches of text spans were allowed during the calculation of the agreement on the detection of temporal expressions.

	Detection	Classification
Average observed agreement	0.75	0.89
Cohen's Kappa	0.74	0.76

Table 2: Inter-annotator agreement values.

The inter-annotator agreement values in Table 2 show that temporal entity annotation in historical texts is a highly context dependent task, and detection of a temporal expression is the most difficult part of it. Identification of a certain expression as temporal requires a thorough understanding of the context. For instance, the word *"jarzit"*, because of its similarity with "Jahreszeit" in modern German (*en* "season"), was tagged by one of the annotators as temporal expression of the type DURATION. However, in the given context *"jarzit"* should be normalised to "Jahrzeit" in modern German, referring to the event of the commemoration of a deceased person, and therefore not being a proper temporal expression.

The annotation process was finished by adjudicating the annotations, i.e., deciding which annotations should be kept in the resulting Gold Standard. According to (Pustejovsky and Stubbs, 2013), the adjudication process should be performed by those who were involved in creating the annotation guidelines, as they will have the best understanding of the annotation purpose. For this reason, the adjudication was performed by the author of the paper. The following features of each tag were adjudicated: extent of the tagged temporal expression, type and value. The tag extent was judged based

on the general rule of span economy: the tagged expressions should contain the smallest number of tokens needed to identify it as temporal expression of a particular type. For example, "1. Januar" (*en*: 1st of January) is preferred to "am 1. Januar" (*en*: on the 1st of January).

Table 3 presents the comparison between the adjudicated Gold Standard and its predecessors, i.e., annotations produced: 1) automatically by HeidelTime (HT); 2) by human annotators (A1 and A2).

	R	P	F	Type	Value
HT	0.26	0.96	0.41	96%	81%
A1	0.93	0.95	0.94	91%	84%
A2	0.88	0.90	0.89	95%	79%

Table 3: Annotation produced by a rule-based system (HT) and manual annotations (A1, A2) evaluated against the Gold Standard. Recall/precision/f-measure scores are calculated for tag extraction, whereas scores in "Type" (correctly classified) and "Value" (correctly normalised) columns are calculated based on the correctly extracted expressions.

4. Experiments Based on the Gold Standard

Several projects in the recent years applied normalisation techniques for the tasks of information extraction.

In (Pettersson et al., 2014) various methods of normalisation (i.e., rule-based, dictionary-based, Levenshtein-based, and based on statistical machine translation) are evaluated d to the task of the verb phrase extraction from Early Modern Swedish texts. The best scores for normalisation and subsequent verb phrase extraction (92.9% accuracy and 87.5 F-score respectively) were achieved by the character-based machine translation approach. Logačev et al. (2014) used a normalisation method based on weighted edit distances to improve part of speech tagging (POS) of several Early New High German texts. The tagging accuracy improved by the average of 2% for the normalised texts. We will follow the steps of these researchers and observe, to what extent the performance of a temporal tagger developed for modern texts can be improved by using normalisation as a pre-processing step.

We started our trial of spelling normalisation methods by an edit-distance based technique described in (Pettersson et al., 2013), used to improve the performance of existing NLP tools (developed for the modern language) for the task of verb extraction from historical Swedish texts, allowing to improve recall from 64.2% for unnormalised text to 86.2%. This approach benefits from context-sensitive weights (lower than 1) for commonly occurring edits and a threshold value for a dictionary entry to be considered as a normalisation candidate, both learned from a parallel corpus of manually normalised data. The only resource for historical German containing relatively large amount of manually normalised data is the GerManC corpus including texts from the period 1650–1800 (Scheible et al., 2011). The normalised subset of this corpus belongs to the period 1659–1780 and contains about 50,000 tokens. We normalised the Gold Standard corpus applying the edit-based method with context-sensitive weights and threshold value for candidates learned from the GerManC parallel data. Table 4 presents the results of the temporal annotation.

	R	P	F	Type	Value
HT	0.27	0.89	0.41	96%	85%

Table 4: Evaluation of the temporal annotation produced by HeidelTime after the Gold Standard corpus was normalised with a weighted edit-distance technique.

After normalisation, the recall value improved by only 0.01 point, while precision even dropped from 0.96 to 0.89, compared to similar values in Table 3 for the annotation produced by temporal tagger on the Gold Standard before normalisation. From this experiment we concluded that the use of manually normalised resources on texts from a slightly different period of time does not produce a positive effect on the output of the temporal tagger.

5. Conclusion

In this paper we described the process of creation of a Gold Standard corpus of Early New High German, containing manual annotation of temporal entities. Given the absence of similar corpora for historical German of this period, the creation of this annotation was a necessary step in the development of a temporal entity extraction system for historical texts. The Gold Standard corpus is used for quality estimation of the automatically produced temporal annotation. Detection of temporal expressions is a difficult task d to a high lexical and spelling variation in our data, therefore, at this point of our research, we are interested in the ability of the temporal entity extraction system to identify temporal expressions in text, reflected in the recall values. First, we evaluated the performance of the rule-based temporal tagger HeidelTime (with modern resources enhanced with the resources adapted for the Text+Berg corpus, 1964-2009) against our Gold Standard and obtained the recall of 0.26. In attempt to reduce spelling variation preventing temporal tagger developed for contemporary German from recognizing temporal expressions in our Gold standard, at the first stage of our experiments we opted for the spelling normalisation approach. We normalised the Gold Standard corpus using an edit distance metric with context-sensitive weights learned from the manually normalised subset of the GerManC corpus. The recall value after tagging the normalised text only reached 0.27. We assume, that such a little improvement is due to the fact that the subset of the GerManC corpus used for learning edit weights and the threshold for the dictionary candidates matching, belongs to a later state of German, covering the period from 1659–1780, whereas our Gold Standard contains text from 1450 to 1550.

Future work includes further evaluation of various normalisation techniques. After the best-performing normalisation approach or combination of methods will be defined, we will manually correct a portion of the output. We will then apply a modern temporal extraction system on the manually normalised subset of the Gold Standard in order to establish, to what extent spelling normalisation can improve

temporal tagging. We expect a certain portion of expressions not to be matched after spelling normalisation, because they either disappeared from the modern language, or lost their temporal semantics, e.g., *"stübglogge"* from Example 1. Machine learning techniques may be applied to deal with such expressions. For instance, a character-level classifier can be used in order to learn the shape of the word as a sequence of characters. Successful character-based systems for information extraction tasks were described in (Klein et al., 2003) and (Qi et al., 2014).

The presented Gold Standard corpus is available for research purposes. To obtain a copy of the corpus, please contact the author of the paper.

6. References

Angel X. Chang and Christopher Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning. 2003. Named entity recognition with character-level models. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 180–183, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pavel Logačev, Katrin Goldschmidt, and Ulrike Demske. 2014. Pos-tagging historical corpora: The case of early new high german. In *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories*, Tübingen, Germany, December.

Eva Pettersson, Beata Megyesi, and Joakim Nivre. 2013. Normalization of historical text using context-sensitive weighted levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference on Computational Linguistics :*.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2014. A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 32–41, Gothenburg, Sweden, April. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2013. *Natural language annotation for machine learning*. OReilly Media, Sebastopol, CA.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. Timeml: Robust specification of event and temporal expressions in text. In *in Fifth International Workshop on Computational Semantics (IWCS-5*.

Yanjun Qi, Sujatha G. Das, Ronan Collobert, and Jason Weston. 2014. Deep learning for character-based information extraction. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings*, pages 668–674.

Katrin Michaela Rettich. 2013. *Automatische Annotation von deutschen und französischen temporalen Ausdrücken im Text+Berg-Korpus Zusammenfassung*. Master thesis, University of Zurich.

Roser Saurí, Jessica Littman, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. TimeML annotation guidelines, version 1.2.1.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' pos-tagger on early modern german text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, pages 19–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

Frank Schilder and Andrew McCulloh. 2005. Temporal information extraction from legal documents. In Graham Katz, James Pustejovsky, and Frank Schilder, editors, *Annotating, Extracting and Reasoning about Time and Events*, number 05151 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Jannik Strötgen and Michael Gertz. 2011. Wikiwarsde: A german corpus of narratives annotated with temporal expressions. In *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2011)*, pages 129–134, Hamburg, Germany, September.

Martin Volk, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, and Beni Ruef. 2010. Challenges in building a multilingual alpine heritage corpus. In *LREC*, Valetta, Malta. European Language Resources Association.

Šolar 2.0: nadgradnja korpusa šolskih pisnih izdelkov

**Iztok Kosem,*† Tadeja Rozman,*† Špela Arhar Holdt,*†
Polonca Kocjančič,* Cyprian Laskowski*‡**

* Zavod za uporabno slovenistiko Trojina,
Trg republike 3, 1000 Ljubljana
iztok.kosem@trojina.si, tadeja.rozman@trojina.si, spela.arhar@trojina.si,
polonca.kocjancic@guest.arnes.si, cyp@trojina.si
† Filozofska fakulteta Univerze v Ljubljani,
Aškerčeva 2, 1000 Ljubljana
‡ Center za jezikovne vire in tehnologije, Univerza v Ljubljani,
Večna pot 113, 1000 Ljubljana

Povzetek

Korpus Šolar je nastal v letih 2009-2010, vanj pa so vključena pisna besedila učencev osnovnih in srednjih šol. Korpus vsebuje skoraj milijon besed, več kot polovica besedil pa ima označene tudi jezikovne popravke učiteljev. V prispevku opisujemo prve korake projekta nadgradnje korpusa Šolar (delovno ime Šolar 2.0), katerega glavna cilja sta povečanje korpusa, kar bo omogočilo boljše posploševanje rezultatov in izvajanje dodatnih raziskav, in pa njegovo uravnoteženje, saj so v trenutni verziji korpusa nekatere slovenske regije slabo zastopane. Predstavljamo rezultate analize obstoječega stanja in zastavljene cilje zbiranja besedil po regijah in šolah. Opišemo tudi postopek digitalizacije, ki bo igral pomembno vlogo v vzpostavitvi dolgoročnega sistematičnega nadgrajevanja korpusa Šolar. V zadnjem delu predstavimo metodo, ki jo bomo uporabili pri reviziji kategorij jezikovnih popravkov.

Šolar 2.0: Increasing the Corpus of Texts Written by Native-speaker Students

The Šolar corpus was built in 2009-2010 and comprises texts written by native-speaker students attending Slovenian elementary and secondary schools. The corpus contains nearly one million words, with over half of the texts also containing teacher corrections of student errors. This paper presents the first steps of the Šolar 2.0 project that aims to expand the corpus, which will enable better generalisations of findings and additional research, and to balance it, given that several Slovenian regions (and their schools) are poorly represented in the existing version of the corpus. We present the analysis of existing corpus structure and goals that were set for further data collection in different regions and schools. Also described is the process of text digitisation, which will play an important role in setting up regular and systematic collection of new texts for the corpus after the end of the project. Finally, a method that will be used for revising the categorization of corrections is presented.

1 Uvod

Korpus šolskih pisnih izdelkov Šolar je nastal v okviru projekta Sporazumevanje v slovenskem jeziku[1] v letih 2009–2010, vsebuje pa skoraj milijon oz. natančno 967.477 besed (Rozman et al., 2012). V korpus je vključenih 2.703 pisnih besedil srednješolcev in učencev zadnjega triletja osnovnih šol (nekaj pa je tudi besedil učencev 6. razreda), ki so jih učenci in učenke samostojno napisali pri različnih predmetih v avtentičnih šolskih situacijah. Korpus Šolar je zato velika pridobitev za slovensko korpusno jezikoslovje, saj ponuja vpogled v pisanje šolajoče se mladine, torej populacije, katere jezikovna produkcija je bila doslej s korpusnim pristopom še neraziskana.

V prispevku predstavljamo projekt nadgradnje korpusa Šolar (delovno ime Šolar 2.0[2]), ki ga sofinancira Ministrstvo za kulturo RS. Projekt poteka v letih 2015–2018, glavna cilja sta povečanje in izboljšanje uravnoteženosti korpusa, ob tem pa želimo odpraviti tudi nekatere pomanjkljivosti pri označevanju jezikovnih popravkov učiteljev, ki jih trenutno vsebuje 56 % besedil v korpusu.

2 Razvojni korpusi

Razvojni korpusi (angl. *developmental corpora*; po Leech 1997:19) so korpusi, ki vsebujejo besedila mlajših maternih govorcev, tj. tistih, ki so še v procesu usvajanja maternega jezika. V primerjavi s korpusi besedil govorcev tujega jezika oz. korpusi usvajanja jezika (angl. *learner corpora*) so razvojni korpusi precej redkejši, vendar pa tako korpusi usvajanja jezika kot razvojni korpusi že dolgo igrajo pomembno vlogo v poučevanju jezika, saj predstavljajo pristop od spodaj navzgor (Osborne, 2002). Rezultati analiz korpusov usvajanja jezika pa so bili uporabljeni tudi v slovarjih, npr. v slovarju Macmillan English Dictionary for Advanced Learners. V nadaljevanju sledi kratek pregled najbolj znanih razvojnih korpusov.

Eden najbolj znanih razvojnih korpusov je CHILDES (Child Language Data Exchange System)[3], baza več kot 130 korpusov (video)posnetkov otroškega govora v 20 različnih maternih jezikih, ki se zbirajo že vse od leta 1981. Polovico korpusov predstavljajo posnetki maternih govorcev angleščine. V bazi je tudi nekaj posnetkov govora otrok z jezikovnimi težavami (npr. disleksijo), tujih govorcev in dvojezičnih otrok.

Bazi CHILDES podobna zbirka korpusov je zbirka projekta EU SACODEYL (2005-2008)[4], ki vsebuje

[1] http://www.slovenscina.eu/
[2] http://solar.trojina.si/
[3] http://childes.psy.cmu.edu
[4] http://www.um.es/sacodeyl

transkribirane (video)posnetke najstniških govorcev angleščine, francoščine, nemščine, italijanščine, litvanščine, romunščine in španščine, starih med 13 in 18 let.

Korpus COLT (Corpus of Teenage Language), ki so ga leta 1993 izdelali na Univerzi v Bergnu, vsebuje 100 posnetkov oz. 50 ur govora 31 najstnikov iz londonskih okrožij, starih med 13 in 17 let (več o projektu gl. v Stenström, Andersen in Hasund 2002). Vseh 500.000 besed v korpusu je bilo ortografsko transkribiranih in oblikoskladenjsko označenih. Korpus je del referenčnega korpusa angleščine BNC (British National Corpus).

Govorjeni jezik otrok vsebuje tudi korpus POW (Polytechnic of Wales), ki so ga izdelali med 1978 in 1984 v južnem Walesu. Korpus vsebuje 65.000 besed, posnetih pa je bilo približno 120 otrok, starih med 6 in 12 let.

Od pisnih razvojnih korpusov sta poznana predvsem korpus LUCY in korpus LOCNESS. LUCY je bil izdelan leta 2003 in je sestavljen iz treh podkorpusov – korpusa besedil objavljenih avtorjev (za pričujoči pregled ni relevanten), korpusa besedil "mlajših odraslih" in korpusa besedil otrok. Korpus mlajših odraslih sestavljajo besedila gradiv za maturo, seminarske naloge in eseji študentov prvega letnika – skupaj 48 besedil oz. 33.000 besed. V korpusu otrok je 150 besedil oz. 30.000 besed, avtorji besedil pa so otroci, stari med 9 in 12 let.

Korpus LOCNESS je bil izdelan na Univerzi v Louvainu in vsebuje pisne izdelke (argumentativne in literarne eseje) dijakov in študentov, maternih govorcev angleščine. Korpus vsebuje 324.304 besede, od tega 60.209 besed predstavljajo eseji britanskih dijakov na maturi, 95.695 besed eseji britanskih študentov, 168.400 besed pa eseji ameriških študentov.

Od "neangleških" korpusov usvajanja maternega jezika velja omeniti korpus jezika tajvanskih otrok (TCLC). Gre za korpus govorjene tajvanščine, ki je bil izdelan v obdobju med 1997 in 2000 in vsebuje 300 ur posnetkov oz. 1,6 milijona besed.

Za slovenski prostor je od tujih korpusov usvajanja maternega jezika najrelevantnejši korpus Chyby, polmilijonski korpus češčine, ki vsebuje besedila (eseje in uvode diplomskih nalog) čeških univerzitetnih študentov (Bušta et al. 2009). Povprečna dolžina besedil je od 600 do 700 besed. Korpus Chyby je eden redkih korpusov usvajanja maternega jezika, ki ima označene napake tvorcev besedil. Te temeljijo na popravkih učiteljev, klasifikacija napak, ki je bila izdelana vnaprej, pa je zaradi predpostavke, da tuji in materni govorci delajo podobne napake, podobna tistim v korpusih usvajanja češčine kot tujega jezika.

Čeprav je razvojnih korpusov manj kot korpusov usvajanja jezika in posledično obstaja tudi precej manj na razvojnih korpusih temelječih raziskav in gradiv[5], pa raziskave, kot so Andersen (1997), Stenström, Andersen in Hasund (2002), Rowland et al. (2005), Bušta et al. (2009), v slovenskem prostoru pa na korpusu Šolar temelječe raziskave Kosem et al. (2012) ter Arhar Holdt in Rozman (2015a), nakazujejo na velik potencial razvojnih korpusov pri spoznavanju jezikovne produkcije mlajših maternih govorcev, njihovih težav pri sporazumevanju ter navsezadnje pri poučevanju jezika ter izdelavi didaktičnih gradiv in orodij (Arhar Holdt et al., in print). Posledično je nadvse smiselno še naprej razvijati razvojne korpuse slovenščine, kot je Šolar, tako z vidika njihove velikosti kot tudi raznolikosti besedil v njih.

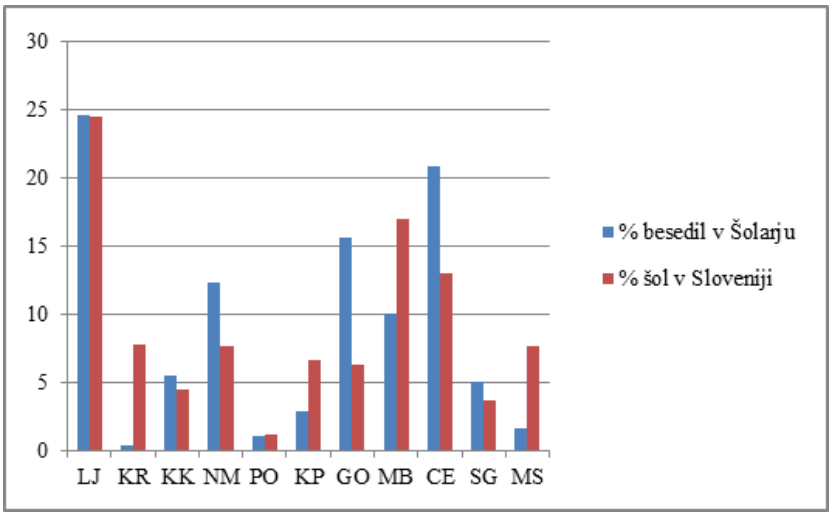
3 Velikost in uravnoteženost korpusa Šolar

Cilj projekta Šolar 2.0 je, da bi korpus Šolar povečali na približno 2 milijona besed oz. na novo dodali od 2500 do 3000 besedil. Večji korpus namreč pomeni večjo zanesljivost in uporabnost, saj po eni strani omogoča lažje posploševanje rezultatov jezikovnih analiz, omogoča raziskave, ki jih je na manjšem gradivu težko opravljati (npr. analize o rabi polnopomenskega besedišča, kolokacijah, vezljivosti, skladnji), hkrati pa nudi večji nabor avtentičnega gradiva za pripravo npr. jezikovnih priročnikov in didaktičnih gradiv.

Povečanje je tudi priložnost za uravnoteženje korpusa. Pri gradnji prvotne verzije je bil glavni cilj doseči regijsko uravnoteženost, tj. vključitev približno 60 % besedil iz regij jugozahodnega dela Slovenije in 40 % besedil iz regij severovzhodnega dela.[6] Pri nadgradnji korpusa pa želimo bolje uravnotežiti tudi razmerje med posameznimi regijami, saj je analiza vključenega gradiva pokazala, da smo – glede na razmerje deležev šol po posameznih regijah – v korpus vključili ustrezno število besedil iz ljubljanske, krške in postojnske regije, iz slovenjgraške, novomeške, goriške in celjske je besedil nekoliko preveč, premalo pa je besedil iz kranjske, koprske, mariborske in murskosoboške regije (Graf 1).

[5] Tako za korpuse usvajanja jezika obstaja zelo obsežna literatura, kot se kaže v več kot 1.100 vnosov obsegajoči bibliografiji združenja raziskovalcev, ki se ukvarjajo z korpusi usvajanja jezika (Learner Corpus Association). Bibliografija je dostopna na http://www.learnercorpusassociation.org/resources/lcb/.

[6] Regije so določene z registrskimi območji, kot jih določa 3. člen Pravilnika o registrskih tablicah motornih in priklopnih vozil (Uradni list RS, št. 83/2006, http://www.uradni-list.si/1/objava.jsp?urlid=200683&stevilka=3637). Med JZ regije sodijo LJ, KR, KK, NM, PO, KP, GO, med SV regije pa MB, CE, SG in MS.



Graf 1: Delež besedil v Šolarju v primerjavi z deležem šol po regijah.

Poleg tega bo pri nadgradnji potrebno vključiti večje število osnovnošolskih besedil, saj je delež besedil iz osnovnih šol bistveno premajhen: v korpusu Šolar je trenutno le 18,6 % osnovnošolskih besedil, čeprav je osnovnih šol v Sloveniji približno 75 %. Zavedamo sicer se, da so šole različno velike in da deleža besedil v korpusu nima smisla pretirano uravnavati glede na število šol, vendar pa je razmerje med deležem osnovnih in deležem srednjih šol v regijah dobra orientacija pri načrtovanju nadgradnje. Idealno bi bilo, če bi po posameznih regijah lahko dosegli razmerja, v katerih bi bilo osnovnošolskih besedil od 20 % do 30 %, čeprav bo, sodeč po trenutni sestavi korpusa, to razmeroma težko doseči (Tabela 1 in Graf 2).[7]

	OŠ		**SŠ**		**skupaj**	
	Šolar	Šolar 2.0	Šolar	Šolar 2.0	Šolar	Šolar 2.0
LJ	52	954	615	369	**667**	**1323**
KR	12	306	0	117	**12**	**423**
KK	83	198	65	45	**148**	**243**
NM	40	324	295	90	**335**	**414**
PO	0	45	28	18	**28**	**63**
KP	0	270	77	90	**77**	**360**
GO	137	252	286	90	**423**	**342**
MB	0	693	271	225	**271**	**918**
CE	0	522	563	180	**563**	**702**
SG	136	153	0	45	**136**	**198**
MS	43	342	0	72	**43**	**414**
skupaj	**503**	**4059**	**2200**	**1341**	2703	5400

Tabela 1: Število besedil v Šolarju po stopnjah in regijah v primerjavi z idealnim številom besedil v načrtovanem Šolarju 2.0 glede na razmerja šol po stopnjah in regijah.

4 Dodajanje novih besedil

V korpus bomo skušali vključiti čim več besedil, ki smo jih prejeli pri gradnji korpusa Šolar v šolskem letu 2009/2010 in v korpus niso bila vključena. Teh besedil je 5891, kar bi bilo dovolj za načrtovano povečanje korpusa, vendar pa s temi besedili nikakor ne moremo doseči želene uravnoteženosti po regijah in stopnjah šolanja. Prednost pri vključitvi bodo zato imela besedila, ki bodo pripomogla k boljši uravnoteženosti korpusa, s šolskim letom 2016/2017 pa začenjamo tudi nov krog zbiranja besedil po šolah, ki bo bomo, tudi zaradi časa in truda, ki smo ga in ga še bomo skupaj z učitelji vložili v zbiranje besedil, skušali v čim večji meri izogniti.

[7] Dejansko bi idealna razmerja ob korpusu s 5400 besedili imela za posledico to, da bi morali odstraniti nekatera besedila iz prve verzije (zlasti iz ljubljanske in mariborske regije), čemur pa se

trajal dve leti. K sodelovanju bomo povabili osnovne šole iz vseh regij, hkrati pa bo potrebno dobiti tudi srednješolska besedila iz kranjske, slovenjgraške in murskosoboške regije, ki jih zaenkrat v Šolarju ni (Graf 2), saj jih pri gradnji korpusa Šolar sploh nismo prejeli.

Graf 2: Deleži osnovnošolskih in srednješolskih besedil po regijah.

Zbiranje besedil bo potekalo podobno kot v šolskem letu 2009/2010 (Rozman et al., 2012): potrebno bo dobiti soglasja učencev oz. njihovih staršev za vključitev besedil v korpus, učitelji in učiteljice bodo vsa besedila označili z metapodatki (gl. poglavje 3.1), nato sledi skeniranje besedil, kar je novost, saj smo pri gradnji korpusa Šolar besedila fotokopirali (gl. poglavje 4). Nova je tudi razširitev zbiranja na besedila 6. razreda, ki jih je v Šolarju le za vzorec; za to smo se odločili, da bi dobili več osnovnošolskih besedil, saj učitelji, ki učijo v zadnji triadi, učijo tudi učence 6. razredov, po drugi strani pa z nižanjem starostne meje avtorjev besedil dobimo boljši vpogled v razvoj pisne jezikovne zmožnosti. V prihodnosti bi zato bilo smiselno korpus razširiti tudi z besedili učencev na razredni stopnji, kar v okviru tega projekta sicer ni predvideno.

Novost, ki do sedaj ni bila vključena zaradi predvidenih možnih težav pri zbiranju gradiva, je tudi informacija, ali ima učenec odločbo o specifičnih učnih primanjkljajih na področju branja in pisanja. V primeru, da bomo ta podatek od učiteljev oz. učencev lahko pridobili, ga bomo vključili v različico korpusa za raziskovalce, javno pa ne bo dostopen. Čeprav na tak način zbranih besedil morda ne bo nujno dovolj za izdelavo uravnoteženega podkorpusa, pa bo zbiranje omogočilo preizkus postopka za gradnjo tovrstnega vira in prve korake pri uporabi gradiva za statistične analize jezikovnih težav in zmožnosti učencev s primanjkljaji v primerjavi s preostalo populacijo. Gradivo bo omogočilo vpogled v posredovane povratne informacije učencem in dijakom s primanjkljaji, in s tem dragoceno avtentično gradivo za izobraževanja bodočih učiteljev (Arhar Holdt in Rozman, 2015b).

4.1 Metapodatki o besedilih

Vsako besedilo mora biti označeno z naslednjimi metapodatki, sicer ni primerno za vključitev v korpus:

- šola, naslov šole in učitelj, ki je besedilo prispeval (ta podatek je shranjen v elektronskem arhivu in v korpusu ni viden);
- program: OŠ, gimnazija, SSI, SPI, PTI, NPI (trenutno so v Šolarju srednješolski programi razdeljeni na gimnazije, strokovne šole, kamor so uvrščeni programi SSI, ter poklicne šole, kamor so uvrščeni programi SPI, PTI in NPI, vendar razmišljamo, da bi zaradi večje jasnosti kategorije preimenovali, kategorijo s poklicnimi programi pa razdružili);
- razred oz. letnik (tukaj sta v obrazcu, ki ga morajo učitelji izpolniti, predvidena tudi poklicni in maturitetni tečaj);
- predmet;
- šolska situacija/besedilna vrsta: ker poimenovanja besedilnih vrst v Šolarju niso najboljša, saj povzročajo težave pri interpretaciji,[8] smo za namene zbiranja določili naslednje kategorije:[9]
    - esej/spis: eseji, spisi in ostala daljša besedila (npr. domišljijski dnevniški zapisi, lastne basni, pravljice ipd.), ki so nastala v okviru šolske naloge (testna situacija),
    - praktično besedilo: vabila, prošnje, opravičila ipd., ki so nastala v testni situaciji (so torej napisana v šoli pri pouku slovenščine za oceno),

[8] V korpusu so besedilne vrste razdeljene na: esej/spis, pisni izdelek (učna ura), test (daljše besedilo), test (odgovori na vprašanja). Kategorije so opredeljene v Rozman et al. (2012).

[9] Verjetno bomo kategorije spremenili tudi v Šolarju.

  - test: test z odgovori na (esejska) vprašanja; test mora vsebovati vsaj dve esejski vprašanji oz. vprašanji, ki zahtevata nekoliko daljši odgovor (kot test se označijo tudi testi pri slovenščini, ki vsebujejo tako vprašanja kot tvorjenje praktičnega besedila),
  - delo v razredu: vsa besedila, ki so nastala pri pouku v netestni situaciji (torej niso za oceno), učitelji dopišejo besedilno vrsto;
- regija;
- šolsko leto.

Učitelji morajo tudi s podpisom potrditi, da dovolijo vnos učiteljskih popravkov v korpus. Zbiramo sicer tako besedila s popravki kot besedila brez popravkov, kljub temu da vnos popravkov v okviru projekta ni predviden.

5 Digitalizacija

Na začetku projekta smo največjo pozornost posvetili optimizaciji procesa zbiranja in pretvorbi besedil v obliko, primerno za vključitev v korpus, saj smo pri izdelavi prve verzije korpusa opazili precej možnosti za izboljšavo in pohitritev. Med ključnimi odločitvami je bila tudi ta, da bomo vsa na roko napisana besedila digitalizirali, tako že zbrana kot tista, ki jih bomo pridobili v prihodnje. Digitalna oblika besedila, tj. sken (pa tudi že na računalniku natipkana besedila), je za nas izhodiščna, saj v primerjavi s fotokopijami omogoča enostavnejše arhiviranje, boljšo in hitrejšo organizacijo transkripcije, hitrejšo izsledljivost izvirnika in navsezadnje tudi boljšo čitljivost besedil oz. njihovih delov (pri izdelavi prve verzije so nam npr. učitelji pošiljali črno-bele kopije, na katerih so bili sicer barvni učiteljski popravki včasih težko razločljivi od učenčevega besedila).

V začetnih mesecih smo tako pripravili in preizkusili postopek digitalizacije besedil, ki so bila pridobljena v šolskem letu 2009/2010 in v korpus še niso bila vključena. V testnem obdobju smo z digitalizacijo pridobili 1651 dokumentov v formatu .pdf, ki so že primerna za vključitev v interni repozitorij in nadaljnjo obdelavo, tj. transkripcijo. Opravila, ki jih predvideva digitalizacija, so naslednja: pregled in ureditev posameznega snopiča z besedili (priprava za skeniranje, ki je odvisna tudi od lastnosti skenerja), določitev identifikacijske številke besedila, evidentiranje metapodatkov, označitev vseh strani besedila z identifikacijsko številko, priprava metapodatkov za skeniranje (t. i. »nosilna stran« ali zbirnik, ki spremlja digitalizirano skupino besedil), skeniranje in lokalno shranjevanje datotek, kompresija datotek, združevanje večstranskih izdelkov v eno datoteko ter pretvorba .tiff v .pdf. Tako pripravljene dokumente nadalje prenesemo v interni repozitorij.

Pri digitalizaciji upoštevamo tudi posebnosti besedil in določene dodatne kriterije, saj so vhodna gradiva dokaj raznolika. S posameznih šol so prišla organizirana v snopiče, ki so lahko tudi notranje razdeljeni na več podsnopičev - pri digitalizaciji to informacijo ohranjamo. Gradiva so lahko enostranska ali dvostranska, formata A4 ali A3. V veliki večini gre za fotokopije, le majhen odstotek gradiv so originali, ki so pogosto neenotnega formata. Večstranska gradiva so pogosto speta. Pri digitalizaciji je poleg naštetega eden najpomembnejših kriterijev ta, da en izdelek enega avtorja postane en dokument oziroma datoteka. To je pomembno upoštevati zlasti takrat, kadar je na eni strani več vsebinsko nepovezanih izdelkov (lahko so različni tudi avtorji) ali pa se konča en izdelek ter začne naslednji. Če je vhodno gradivo fotokopija ali besedilo brez učiteljskih popravkov, digitaliziramo sivinsko. Če gre za original z učiteljskimi popravki, pa skeniramo barvno, saj bodo informacije za transkriptorje tako popolnejše oziroma lažje dostopne. Priprava korpusa predvideva tudi anonimizacijo. Nekateri učitelji so jo izvedli že sami, preostale izdelke pa bomo anonimizirali ob transkripciji besedila. Anonimizacija zajema zakrivanje podatkov v besedilu (metapodatkov o besedilu, kot so ime in priimek avtorja in ime šole, sploh ne beležimo), ki bi lahko razkrili avtorja besedila, npr. imen in/ali priimkov avtorja, njegovih sorodnikov ali prijateljev, imena šole, kraja ipd.

Na podlagi testne digitalizacije smo opravili tudi časovne izračune in pripravili dokument s podrobnim opisom postopka ter predstavitvijo posebnih primerov. Dolgoročni načrt je, da v korpus vključimo vsa zbrana besedila, v času pisanja konferenčnega prispevka pa že pripravljamo spletni repozitorij.

6 Pripis kategorij jezikovnih popravkov

Pomemben del korpusa Šolar so označeni jezikovni popravki učiteljev, ki omogočajo različne raziskave (npr. Arhar Holdt in Rozman, 2015a; Kosem et al., 2012), prav tako pa so pomembni za izdelavo jezikovnih tehnologij, jezikovnih priročnikov in didaktičnih gradiv (npr. Pedagoški slovnični portal[10]). Pomanjkljivost, ki so jo zainteresirani uporabniki večkrat izpostavili, je odsotnost podrobnejše kategorizacije napak šolarjev. Obstoječe kategorije so trenutno namreč zelo splošne (besedišče, oblika, zapis, skladnja – pri čemer imata zapis in skladnja še podkategorije), kar otežuje direktno uporabo korpusa v razredu in urjenje orodij in programov na korpusu Šolar za namene, kot je npr. avtomatska prepoznava napak.

V korpusu Šolar trenutno najdemo 35.035 učiteljskih jezikovnih popravkov, od tega največ na ravni zapisa (61,1 %), sledijo popravki skladnje (17,7 %), besedišča (10,9 %) in oblike (10,3 %). Podrobnejša analiza popravkov je bila opravljena pri izdelavi Pedagoškega slovničnega portala (Kosem et al., 2012), v sklopu katere so bili popravki tudi ročno razvrščeni v približno 700 kategorij jezikovnih težav. Obstoječa kategorizacija se izkazuje za problematično s treh vidikov: (I) Kategorizacijo je izvajalo več označevalcev, vsak na posamezni jezikovni ravnini in po principu od spodaj navzgor. Rezultat so medravninsko deloma različni sistemi kategoriziranja, ki so do sedaj ostajali razdruženi. (II) Ker je bila kategorizacija v veliki meri ciljno usmerjena v pripravo portala, so določene vrste popravkov, npr. popravki ločil, ostali le delno obravnavani. (III) Pripisane kategorije popravkov še niso bile umeščene v XML-strukturo korpusa, torej ostajajo nedosegljive za sintetične jezikoslovne analize. Našteti problemi so botrovali odločitvi o natančnejši reviziji in nadgradnji sistema kategorij ter vpisu le-teh med oznake korpusa Šolar. Pri slednji nalogi se izkazuje za poseben izziv zagotavljanje možnosti označevanja določenega popravka z več različnimi kategorijami (npr. popravek oblike *uplivat* v *vplivati,* ki sodi tako v napake črkovanja –

[10] http://slovnica.slovenscina.eu/

popravek zapisa ụ v besednem vzglasju – kot tudi na raven morfologije – popravek kratkega nedoločnika).

Za učinkovito revizijo obstoječih kategorij je ključna uporaba sistematičnega in fleksibilnega postopka, ki omogoča pregledovanje večjih količin označenih konkordanc in sprotne prekategorizacije označenih napak v njih. V ta namen smo preizkusili orodja, izdelana bodisi predvsem za namene korpusov z jezikovnimi napakami bodisi za namene pripisovanja različnih oznak (npr. TEITOK, WebAnno). Na koncu smo se odločili za uporabo orodja Sketch Engine (Kilgarriff et al., 2004), ki podpira pregleden prikaz jezikovnih popravkov in ponuja možnost pripisovanja ter shranjevanja uporabniško določenih oznak za prikazane konkordance.

Korpus s pripisanimi revidiranimi kategorijami jezikovnih popravkov bomo v drugi polovici projekta uporabili za izdelavo učnega korpusa, namenjenega razvoju sistema za avtomatsko kategorizacijo učiteljskih popravkov in na drugi strani šolskemu pisanju prilagojenih jezikovnotehnoloških orodij, kot so denimo črkovalniki in slovnični pregledovalniki.

7 Sklep

V okviru projekta Šolar 2.0 bomo obstoječi korpus izboljšali in nadgradili, kar bo dobrodošlo tako za raziskovalce in jezikovne tehnologe kot učitelje in učence. Večji korpus, boljša uravnoteženost po regijah in stopnjah šolanja ter dodane podkategorije bodo omogočili nove vpoglede v pisno produkcijo učencev in s tem tudi izdelavo problemsko naravnanih jezikovnih priročnikov in učnih gradiv za osnovne in srednje šole, osnovanih na analizah dejanske jezikovne rabe, ki jih v slovenskem prostoru močno primanjkuje. Poleg tega želimo v okviru projekta vzpostaviti postopek zbiranja in digitalizacije besedil, ki bo omogočal dolgoročno sistematično nadgrajevanje korpusa Šolar.

Korpusi, ki bodo nastali v okviru projekta, bodo na voljo konec leta 2018, in sicer bosta korpus Šolar 2.0 in učni korpus z jezikovnimi popravki na voljo pod licenco CC BY-NC-SA 2.5 SI[11] (gre za licenco, pod katero je na voljo tudi trenutna verzija korpusa Šolar), novo zbrana besedila v korpusu Šolar pa najbrž tudi kot ločen korpus pod licenco CC BY 4.0[12].

8 Literatura

Gisle Andersen. 1997: Pragmatic markers in teenage and adult conversation. *18. konferenca ICAME*. Neobjavljeni prispevek.

Špela Arhar Holdt, Iztok Kosem in Polona Gantar. In print. Corpus-based resources for L1 teaching: The Case of Slovene. V: *Handbook on Digital Learning for K-12 Schools*. Springer.

Špela Arhar Holdt in Tadeja Rozman. 2015a. Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov. V: M. Smolej, ur., *Obdobja 34: Slovnica in slovar - aktualni jezikovni opis*, 1. del, str. 67–74. Znanstvena založba Filozofske fakultete UL. http://centerslo.si/wp-content/uploads/2015/11/34_1-Arhar-Hol-Roz.pdf.

Špela Arhar Holdt in Tadeja Rozman. 2015b. Korpus Šolar: gradivni vir za raziskave pisne produkcije slovenskih učencev. *Bilten društva Bravo*, XI/23: 17-23.

Jana Bušta, Dana Hlaváčková, Miloš Jakubíček, Karel Pala. 2009. Classification of Errors in Text. V: P. Sojka, A. Horák, ur., *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009*. str. 109–119. Brno: Masaryk University. http://nlp.fi.muni.cz/raslan/2009/papers/6.pdf.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz in David Tugwell. 2004. The Sketch Engine. V: G. Williams in S. Vessier, ur., *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004,* str. 105–116. Lorient: Universite de Bretagne - sud.

Iztok Kosem, Mojca Stritar Kučuk, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012. *Analiza jezikovnih težav učencev: korpusni pristop*. Trojina, zavod za uporabno slovenistiko.

Geoffrey Leech. 1997: Teaching and language corpora: A convergence. V: A. Wichmann, S. Fliegelstone, T. McEnery in G. Knowles, ur., *Teaching and language corpora,* str. 1–23. London: Longmann.

John Osborne. 2002. Top-down and bottom-up approaches to corpora in language teaching. Connor, Ulla and Thomas A. Upton (ur.): *Applied Corpus Linguistics: A Multidimensional Perspective,* str. 251–265. Amsterdam: Rodopi.

Caroline F. Rowland, Julian M. Pine, Elena V. Lieven, Anna L. Theakston. 2005. The incidence of error in young children's Wh-questions. *Journal of speech, language and hearing research* 48/2:384–404.

Tadeja Rozman, Irena Krapš Vodopivec, Mojca Stritar Kučuk in Iztok Kosem. 2012. *Empirični pogled na pouk slovenskega jezika*. Trojina, zavod za uporabno slovenistiko.

Anna-Brita Stenström, Gisle Andersen, Ingrid Kristine Hasund. 2002. *Trends in Teenage Talk: Corpus compilation, analysis and findings.* Studies in corpus Linguistics 8. Amsterdam: John Benjamins.

[11] https://creativecommons.org/licenses/by-nc-sa/2.5/si/

[12] https://creativecommons.org/licenses/by/4.0/

Baza kolokacijskega slovarja slovenskega jezika

Simon Krek,[+]* Polona Gantar,[†]* Iztok Kosem,[‡†] Vojko Gorjanc,[†] Cyprian Laskowski[‡]*

[+] Laboratorij za umetno inteligenco, Institut »Jožef Stefan«, Jamova 29, 1000 Ljubljana
* Center za jezikovne vire in tehnologije, Univerza v Ljubljani, Večna pot 113, 1000 Ljubljana
simon.krek@guest.arnes.si
† Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani,
Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@ff.uni-ljs.si, vojko.gorjanc@ff.uni-lj.si
‡ Center za uporabno jezikoslovje, Trojina
Partizanska cesta 5, 4220 Škofja Loka
iztok.kosem@trojina.si, cyp@trojina.si

Povzetek

V prispevku so opisani postopki izdelave Baze kolokacijskega slovarja slovenskega jezika, ki predstavlja samostojno fazo v procesu izdelave Kolokacijskega slovarja slovenskega jezika. Opis zajema prilagoditev že obstoječih kolokacijskih podatkov v predhodno izdelani Leksikalni bazi za slovenščino in se osredotoča na dopolnitev z avtomatsko izluščenimi podatki iz korpusa. V osrednjem delu prispevka je opisan nabor avtomatsko izluščenih podatkov, ki jih trenutno vsebuje kolokacijska baza, ter njihov prikaz v spletnem vmesniku. Prispevek zaključuje analiza evalvacije uporabniškega vmesnika in načrti za nadaljnje delo.

Slovene Collocations Dictionary Database

The paper describes the compilation of Slovene Collocations Dictionary Database which represents a separate stage in the process of compiling the Collocations Dictionary of Slovene. The described process includes the adaptation of the existing collocations data in the Slovene Lexical Database and focuses on the upgrade with the automatically extracted data from the corpus. The central part of the paper describes the new data collection currently included in the collocations database, and its visualisation in the web interface. We conclude with the presentation of evaluation results, both of the data and the user interface, and with the plans for future work.

1 Uvod

V prispevku opišemo izdelavo Baze kolokacijskega slovarja slovenskega jezika (BKSSJ), ki predstavlja del procesa izdelave Kolokacijskega slovarja slovenskega jezika (KSSJ).[1] Ta bo ob zaključku vseboval 5.000 gesel in bo prosto dostopen na spletu. Ideja je povezana z dejstvom, da kolokacijski slovar za slovenščino ne obstaja in da so kolokacijski podatki, ki so strukturirani v obstoječih slovarjih v t. i. iztržkih in slovarskih zgledih, nedostopni kot računalniško procesljiva baza, hkrati pa ne odražajo stanja v sodobnem slovenskem jeziku. Potreba po kolokacijskem slovarju za slovenščino je bila v jezikovni skupnosti že večkrat izražena (Čibej et al., 2015). BKSSJ trenutno vsebuje 2.500 gesel, v načrtu je razširitev na približno 50.000 iztočnic.

Kolokacijski slovarji (Rundell et al., 2010; Cleveland-Marwick et al., 2013) se selijo na splet (prim. Roth, 2013), hkrati pa zaradi vse bolj naprednih orodij ter statistik za luščenje kolokabilno izstopajočih sopojavitev nastajajo avtomatsko generirani kolokacijski slovarji (Baisa in Suchomel, 2014; Kallas et al., 2015). Ti so potem lahko kot taki predstavljeni uporabnikom ali pa služijo kot osnova za izdelavo kolokacijskih in podobnih slovarjev (gl. Kallas et al., 2015).

2 Od leksikalne baze do kolokacijskega slovarja

Osnovo KSSJ sestavljajo kolokacijski podatki, vključeni v Leksikalno bazo za slovenščino (LBS; Gantar, 2015), in avtomatsko izluščeni kolokacijsko relevantni podatki iz korpusa Gigafida (BKSSJ). Končni slovar bo vključeval večnivojski opis kolokacij, ki ga bo mogoče vključiti v obsežnejšo slovarsko bazo, namenjeno izdelavi slovarja sodobnega slovenskega jezika (SSSJ; Gorjanc et al., 2015), hkrati pa bo uporabnikom na voljo kot samostojen prosto dostopen kolokacijski slovar. Glede na to da proces izdelave SSSJ predvideva več med seboj soodvisnih leksikografskih faz (Gantar et al., 2015), pri izdelavi KSSJ preizkušamo in nadgrajujemo tako proces avtomatskega luščenja podatkov iz korpusa kot tudi možnosti vključevanja množičenja pri čiščenju in urejanju avtomatsko izluščenih podatkov.

Prvi del izdelave KSSJ je bila priprava in ureditev podatkov v LBS za izdelavo druge verzije, tj. LBS2. V ta namen smo podatke iz LBS obogatili z (a) metapodatki, kot so identifikator leme, podatek o frekvenci leme v korpusu, podatek o izvoru, tj. ali je bila iztočnica v LBS izdelana ročno ali že na podlagi avtomatsko izluščenih podatkov, (b) dodali XML-oznake posameznim elementom sheme, npr. identifikator kolokatorja, pridobljenega iz Sloleksa, in oznake za besedno vrsto ter (c) odstranili napake, ki so nastale pri ročni izdelavi gesel v LBS.

[1] Projekt se izvaja v okviru raziskovalnega programa Slovenski jezik – bazične, kontrastivne in aplikativne raziskave ter infrastrukturnih programov Centra za jezikovne vire in tehnologije Univerze v Ljubljani in Centra za uporabno jezikoslovje na zavodu Trojina: http://www.cjvt.si/kssj/.

Drugi del izdelave KSSJ je zajemal avtomatsko luščenje kolokacijskih podatkov. Ta proces je bil na manjšem obsegu lem preizkušen že pri izdelavi LBS (Kosem et al., 2012). Avtomatsko izluščeni podatki in podatki iz LBS2 so bili združeni v osnovo za kolokacijski slovar. V tem postopku smo najprej avtomatsko izluščili kolokacijske podatke, tj. lemo in kolokatorje za omejen nabor kolokacijsko relevantnih skladenjskih struktur in pripadajoče korpusne zglede, za 2.500 v LBS že obstoječih gesel ter jih združili z na novo izluščenimi podatki za dodatnih 2.500 lem. Postopek je zahteval več prilagoditev, ki smo jih opravili v postopku postprocesiranja, kot npr. poenotenje poimenovanj skladenjskih struktur, prepoznavo struktur pri glagolih, pripravo podatkov po novi shemi XML, vzpostavljanje povezav med kolokacijami v LBS in njihovimi zgledi, pripisovanje novih avtomatskih podatkov LBS geslom na podlagi obstoječih pomenskih členitev ipd. Za luščenje smo uporabili orodje Sketch Engine (Kilgarriff et al., 2004) in za slovenščino prilagojeno aplikacijo GDEX za izbiro dobrih korpusnih zgledov (Kosem et al., 2011).

Združeni podatki iz 2.500 gesel LBS2 in 2.500 novih avtomatsko izluščenih gesel torej predstavljajo osnovo, pripravljeno za nadaljnjo obdelavo, v okviru katere predvidevamo uporabo množičenja (Fišer et al., 2015), zlasti pri razporejanju kolokacij pod ustrezne pomene ter pri čiščenju podatkov, ki ga ni bilo mogoče izvesti v fazi luščenja in postprocesiranja.

3 Baza kolokacijskega slovarja slovenskega jezika (BKKSJ)

V tem delu prispevka opišemo BKSSJ, avtomatsko izluščenih 2.500 gesel iz korpusa Gigafida z izbranim naborom skladenjskih relacij, skupaj s korpusnimi zgledi. Avtomatsko izluščeni podatki so bili v postopku postprocesiranja dodatno prilagojeni, predvsem glede ujemanja po spolu, sklonu in številu, kot samostojna baza pa so dostopni tudi v spletnem vmesniku.[2]

BKSSJ ima večdelno vlogo, saj predstavlja podatkovno osnovo za gesla v KSSJ, je samostojna zbirka tako za analizo in izboljšavo in nadgradnjo avtomatskih metod luščenja podatkov kot za preizkušanje množičenja, in je navsezadnje tudi spletno dostopen vir za različne uporabnike.

3.1 BKSSJ kot zbirka

Baza z 2.500 iztočnicami (1.000 samostalnikov, 750 glagolov, 625 pridevnikov in 125 prislovov) vsebuje 2.310.100 kolokacij v 72.117 skladenjskih strukturah, za vsako kolokacijo je tipično izločenih tudi pet zgledov rabe iz korpusa. Izhajajoč iz metodologije, uporabljene pri izdelavi LBS, je bilo izluščenih 528 različnih skladenjskih struktur, od tega pri samostalnikih 192, pri glagolih 147, pri pridevnikih 107 in pri prislovih 82. V Tabeli 1 navajamo prvih pet najpogostejših.

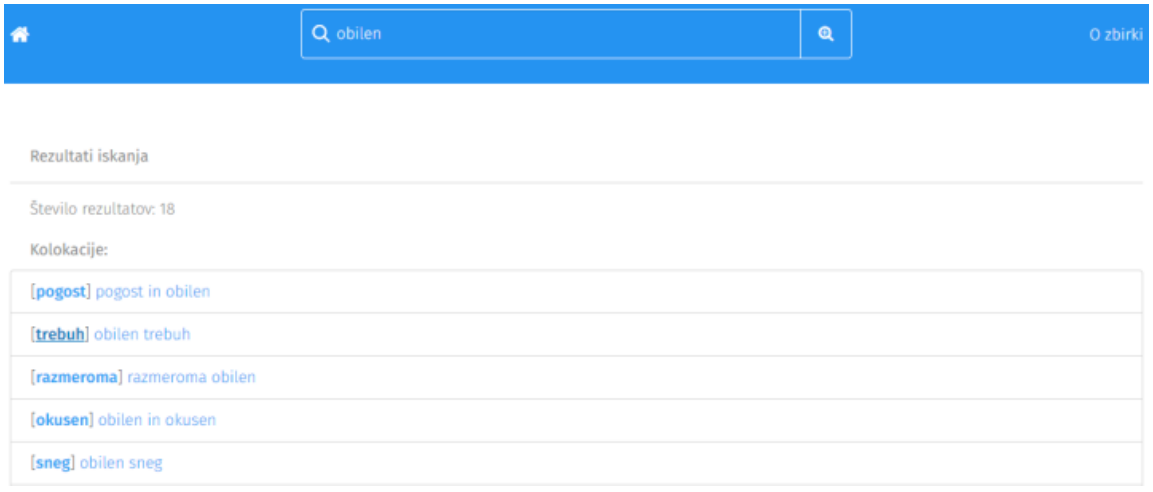
Ker gre za strojno luščenje, BKSSJ vsebuje tudi napake, ki izhajajo deloma iz jezikoslovnega označevanja korpusa Gigafida, deloma iz luščenja v orodju Sketch Engine. Primer prvega je kolokacija *kraj kolesa* namesto *kraja kolesa* v strukturi *SBZ0 + sbz2,* primer drugega je *bolezen očija* v enaki strukturi. Pri kolokatorju *oči* sta možni dve lemi: *oči* (eden od staršev) ali *oči* (množinski samostalnik – *oko*). V postprocesiranju je bila izbrana rodilniška oblika prve leme (*očija*) namesto rodilniške oblike druge leme (*oči* = bolezen oči).

Odpravo napak lahko izvajamo na dveh ravneh. Kot prvo, z izboljševanjem postopka avtomatskega luščenja, tj. z izboljšavo slovnice besednih skic, orodja GDEX za luščenje dobrih zgledov in navsezadnje tudi oblikoskladenjskega označevalnika za slovenščino. Tovrstne izboljšave potrebujejo bolj sistematičen pristop in obsežnejši pregled napak, zaradi česar smo se tudi lotili evalvacije podatkov v BKSSJ z vidika potencialnih uporabnikov (glej 3.3). Po drugi strani se napake lahko odpravljajo z izboljševanjem postopka postprocesiranja. Na primer, za zgornji primer *bolezen očija* in podobne primere lahko pravo lemo identificiramo z analizo oblik, ki se pojavljajo v izluščenih zgledih.

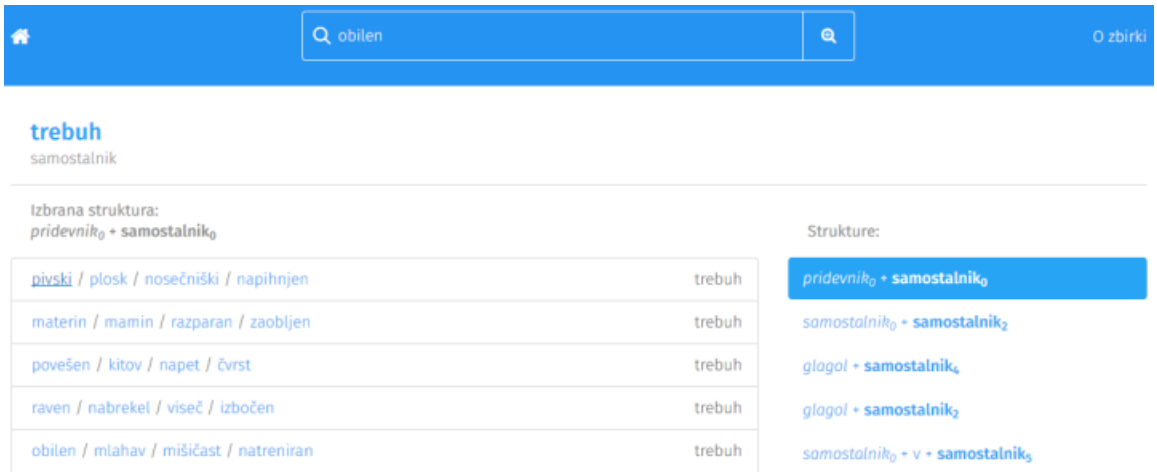
št.	**struktura**	**opis**	**primer kolokacije**	**število struktur**
1	sbz0 SBZ2	samostalnik v poljubnem sklonu + samostalnik v rodilniku	[štafeta, eliksir, vrelec, kult] mladosti	1.783
2	GBZ sbz4	glagol + samostalnik v tožilniku	priznati [premoč, krivdo, zmoto, neodvisnost]	1.672
3	PBZ0 sbz0	pridevnik v poljubnem sklonu + samostalnik v poljubnem sklonu	mlada [generacija, ženska, družina, igralka]	1.609
4	GBZ sbz2	glagol + samostalnik v rodilniku	priznati [imunitete, očetovstva, krivde, neodvisnosti]	1.598
5	GBZ z sbz6	glagol + z + samostalnik v orodniku	priznati z [nasmeškom, grenkobo, obžalovanjem, nasmehom]	1.193

Tabela 1: Najpogostejše skladenjske strukture v BKSSJ.

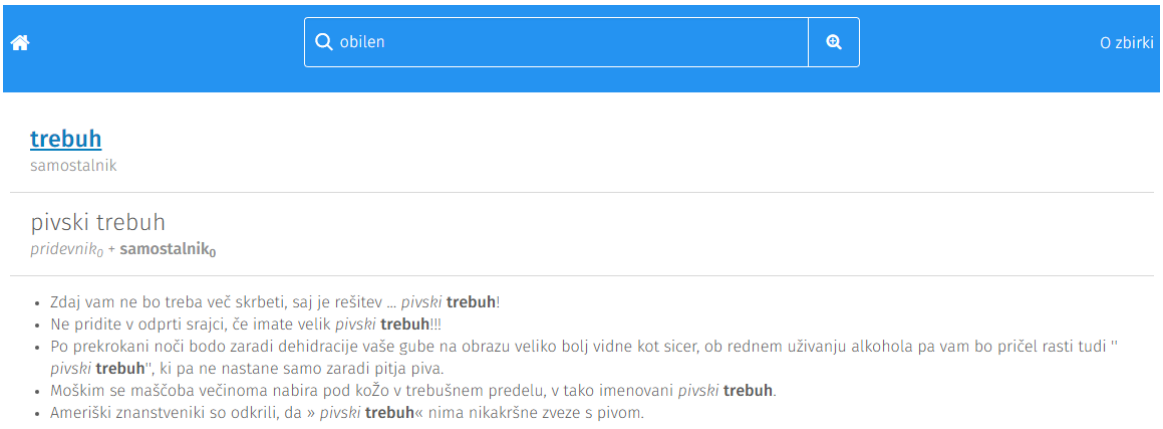
[2] Pri izdelavi vmesnika so sodelovali: Rok Rejc, Gašper Uršič, Simon Krek, Iztok Kosem, Polona Gantar, Vojko Gorjanc. Spletna stran: http://bkssj.cjvt.si/.



Slika 1: Prikaz kolokacij, ki vključujejo iskano besedo.



Slika 2: Prikaz kolokacij znotraj izbranega kolokatorja, ki je v bazi obdelan kot iztočnica.



Slika 3: Prikaz korpusnih zgledov za izbrano kolokacijo.

3.2 BKSSJ v spletnem vmesniku

Spletni vmesnik, ki omogoča iskanje po celotni bazi, je bil zasnovan v okviru diplomske naloge na Naravoslovnotehniški fakulteti (Uršič, 2015) ter študentskega dela na Fakulteti za računalništvo in informatiko UL. Zadetki pri iskanju so ločeni na prikaz iztočnic, ki vsebujejo kolokacije za iskano lemo in na prikaz seznama kolokacij (zadnje prikazuje Slika 1 zgoraj), pri čemer je pri seznamu kolokacij vedno poudarjeno izpisan kolokator, ki je v bazi predstavljen kot iztočnica.

Pri prikazu slovarskega gesla je mogoče kolokacije filtrirati po skladenjskih strukturah, ki so prikazane na desni strani vmesnika (Slika 2), ob kliku na posamezno kolokacijo pa so na novi strani prikazani pripadajoči izluščeni zgledi iz korpusa (Slika 3).

3.3 Evalvacija BKSSJ

Z namenom izboljšati uporabniško izkušnjo in optimizirati prikaz izluščenih podatkov je bila že v tej fazi izvedena kratka anketa med potencialnimi uporabniki BKSSJ oz. bodočega KSSJ.[3] Poleg osnovnih podatkov o anketirancih (smer študija; starost, spol) in o izbiri naprave za dostop do BKSSJ (namizni ali prenosni računalnik, tablica, telefon) je bil osnovni namen ankete zbrati predvsem opažanja, predloge in želje, ki se nanašajo na organizacijo in postavitev podatkov na strani, navigacijo med razdelki in vizualizacijo podatkov. Odgovore anketirancev smo razdelili v 5 kategorij, ki jih na kratko povzemamo v nadaljevanju.

(1) Razvrščanje kolokacij. Večina anketirancev je opozorila na preobsežnost seznamov kolokacij, ki se izpišejo za iskano besedo v primeru, ko ta v bazi ni prikazana kot iztočnica (prim. Sliko 1). Predlagane so predvsem naslednje možnosti razvrščanja: po pogostnosti, po obliki oz. strukturi kolokacije, po abecedi iztočnice, po pomenskih sklopih in po besedni vrsti iztočnice. Predlagane so tudi kombinacije razvrščanja, npr. najprej po strukturi, znotraj posamezne strukture pa po pogostnosti. Taka razvrstitev se zdi smiselna zlasti z vidika relevantnosti in intuitivne prepoznavnosti tipičnih sopojavitev, hkrati pa bi bilo mogoče na ta način potisniti napačno izluščene sopojavitve, ki se jim v postopku avtomatizacije ni mogoče povsem izogniti, na konec obsežnih seznamov.

(2) Prikaz podatkov na strani. Največ nejasnosti in neintuitivnosti pri branju podatkov povzroča razmerje med prikazom iskane besede kot kolokatorja (kadar ni iztočnice; Slika 1) in prikazom iskane besede kot iztočnice (Slika 2). Ker je trenutno v bazi le 2.500 iztočnic, je večina iskanih besed v bazi prikazana le, če se pojavljajo kot kolokatorji pri v bazi obstoječih iztočnicah, pri čemer te kolokacije za iskano besedo po pričakovanju niso najbolj tipične. Ta problem bo delno odpravljen z razširitvijo baze na 50.000 iztočnic in s katerim od zgoraj predlaganih načinov razvrščanja. Anketirance je pri prikazu podatkov na strani motilo tudi preveč praznega prostora, postavitev menija za filtriranje struktur na desno stran zaslona, preveč zapleten (metajezikoslovni) zapis struktur, prikaz kolokatorjev v nizih za poševnicami namesto v stolpcih ipd. Načeloma so bili anketiranci zadovoljni z barvnimi kombinacijami, velikostjo fonta in številom izpisanih zgledov.

(3) Navigacija po strani/geslu. Kot dobro rešitev so anketiranci izpostavili spustni meni s predlogi tipičnih sopojavitev, kritizirali pa so premikanje po straneh med obsežnimi seznami kolokacij na prvem nivoju in preobsežne sezname struktur na drugem nivoju. V ta namen so bile predlagane rešitve združevanja struktur pod opisne razlage, npr. kakšen je (+ iskani samostalnik); kaj počnemo z (+ iskani samostalnik) oz. združevanje struktur s posameznimi konkretnimi predlogi pod skupno oznako "predlog", npr. glagol + predlog + samostalnik$_5$, ki bi vsebovala strukture z vsemi predlogi (v, na, po, pri itd.), ki se vežejo s samostalnikom v mestniku.

(4) (Ne)ustreznost kolokacij. Po pričakovanju je anketirance zmotila neustreznost prikazanih podatkov, vezana pretežno na napake pri avtomatskem luščenju kolokacij iz korpusa. Izpostavljeni so bili predvsem primeri, kjer je prikazana kolokacija oz. kolokator del širše (ustaljene) zveze, npr. *mali sneg ← nekaj malega snega; hud sneg ← v hudem snegu; zdrav duh ← zdrav duh v zdravem telesu.* Prav tako prihaja pri nekaterih prikazanih sopojavitvah za naključne kombinacije, ki sicer formalno ustrezajo skladenjski strukturi, vendar v besedilu pripadajo dvema različnima skladenjskima strukturama, npr. *sanjati v ligi* (gbz v SBZ5: sanjati o čem + zmaga v ligi). Poleg omenjenega, so anketiranci opozorili še na napačno lematizacijo, neustrezno obliko kolokatorja, npr. osnovnik namesto presežnik: *lep med lepotico ← najlepša med lepoticami*; razlikovanje med glagoli s in brez *se*/*si* ter med zanikanimi in nezanikanimi glagoli, npr. *spomniti zabavo ← spomniti se zabave*; priznati [premoč, krivdo] in ne priznati [imunitete, očetovstva, krivde], na napake v zapisu struktur, npr. s + prislovom namesto "s prislovom"; na vrstni red znotraj kolokacije, npr. zamenjava med osebkom in predmetom v imenovalniku: *poskrbeti vreme ← vreme poskrbi za kaj*; neločevanje med stalnimi zvezami in kolokacijami, npr. *reševanje/jabolko/zgladitev spora* (SZ: jabolko spora), vključevanje lastnoimenskih kolokatorjev, npr. *trgovina v Citypark/v Globokem/v Žireh* itd., ki so s kolokacijskega vidika manj relevantni.

(5) (Ne)uporabnost. Med razlogi, ki delajo bazo v tej fazi manj uporabno, so anketiranci izpostavili predvsem premajhno strukturiranost in izčiščenost podatkov, preveliko količino podatkov in nerazločevalnost med relevantnimi (pogostimi) in nerelevantnimi oz. manj relevatnimi kolokacijami. Med manjkajočimi podatki so našteli predvsem: podatke o (ne)standardnosti oz.

[3] Za namene evalvacije BKSSJ sta bili izdelani dve anketi: prva je bila namenjena študentom (prevajalstvo, novinarstvo, računalništvo idr.), druga pa sodelavcem pri projektu priprave novega slovarja sodobne slovenščine. V času priprave prispevka sta bili anketi še aktivni, zato so v prispevku upoštevani vmesni rezultati, in sicer zgolj ustrezno rešene ankete (od skupno 243 le 92 anket) oz. ankete, ki vsebujejo komentarje (35 anket). Anketi sta dostopni na https://www.1ka.si/a/94327 in https://www.1ka.si/a/92168.

zaznamovanosti kolokacije; pogostnosti kolokacije; pomenu iztočnice; med manjkajočimi funkcionalnostmi pa možnost določiteve števila besed v okolici iskane besede in dostop do baze za tekstovno rudarjenje.

4 Zaključek in nadaljnje delo

Nadaljnje delo bo potekalo v dveh smereh: pri izdelavi KSSJ bo poudarek na izdelavi pomenske členitve za leme, ki ne izhajajo iz LBS, vzpostavitvi procesa množičenja ter evalvaciji rezultatov, pridobljenih v tem procesu. Pri izdelavi BKSSJ predvidevamo množični izvoz podatkov iz korpusa (približno 50.000 iztočnic) in izboljšanje njihove vizualizacije na spletu

Poleg tega preučujemo tudi izrabo dodatnih funkcionalnosti v orodju Sketch Engine, kot sta gručenje (ang. clustering) in najpogostejši besedni niz (ang. longest comment match), za namene izboljšave postopka avtomatskega luščenja. Prvi namreč kaže dober potencial za grupiranje (semantično) podobnih kolokacij, kar je koristno tako za vizualizacijo BKSSJ (ena od preferenc uporabnikov pri evalvaciji) kot za samo pomensko členitev pri izdelavi končnih gesel v KSSJ, medtem ko najpogostejši besedni niz lahko služi kot dodatna informacija pri postprocesiranju izluščenih podatkov.

5 Literatura

Vit Baisa in Vit Suchomel. 2014. SkELL: Web Interface for English Language Learning. *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*, str. 63–70. Brno: Tribun EU.

Karen Cleveland-Marwick et al. (ur.). 2013. Longman collocations dictionary and thesaurus. Harlow, Essex: Pearson Education.

Jaka Čibej, Vojko Gorjanc in Damjan Popič. 2015. Vloga jezikovnih vprašanj prevajalcev pri načrtovanju novega enojezičnega slovarja. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 168–181. Ljubljana: Znanstvena založba Filozofske fakultete.

Darja Fišer, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Iztok Kosem, Špela Arhar Holdt, Damjan Popič in Tomaž Erjavec. 2015. Množičenje za slovar sodobnega slovenskega jezika. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 566–586. Ljubljana: Znanstvena založba Filozofske fakultete.

Polona Gantar. 2015. *Leksikografski opis slovenščine v digitalnem okolju*. Ljubljana: Znanstvena založba Filozofske fakultete UL. http://www.ff.uni-lj.si/Portals/0/Dokumenti/ZnanstvenaZalozba/e-knjige/Leksikografski.pdf.

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. 280–297. Ljubljana: Znanstvena založba Filozofske fakultete.

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek, ur., 2015. *Slovar sodobne slovenščine: problemi in rešitve*. Ljubljana: Znanstvena založba Filozofske fakultete.

Jelena Kallas, Adam Kilgarriff, Kristina Koppel, Elgar Kudritski, Margit Langemets, Jan Michelfeit, Maria Tuulik in Ülle Viks. 2015. Automatic generation of the Estonian Collocations Dictionary database. V: I. Kosem, M. Jakubiček, J. Kallas, S. Krek, ur., *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015*. Herstmonceux Castle, United Kingdom, str. 1–20. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, David Tugwell. 2004. The Sketch Engine. V: G. Williams, S. Vessier, ur., *Proceedings of the Eleventh EURALEX International Congress.* EURALEX 2004 Lorient, France July 6-10, 2004, str. 105–116. Lorient: Universite de Bretagne - sud.

Iztok Kosem, Polona Gantar in Simon Krek. 2012. Avtomatsko luščenje leksikalnih podatkov iz korpusa. V: T. Erjavec, J. Žganec Gros, ur., *Zbornik 15. mednarodne multikonference Informacijska družba - IS 2012, zvezek C,* str. 117–122. Ljubljana: Institut Jožef Stefan.

Iztok Kosem, Miloš Husák in Diana Mccarthy. 2011. GDEX for Slovene. V: I. Kosem, K. Kosem, ur., *Electronic Lexicography in the 21st Century: New applications for new users. Proceedings of eLex 2011.* Bled, 10-12 November 2011, str. 151–159. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Tobias Roth. 2013. Going Online with a German Collocations Dictionary. V: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik, ur., *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17–19 October 2013,* str. 152–163. Tallinn, Estonia. http://eki.ee/elex2013/proceedings/eLex2013_11_Roth.pdf.

Michael Rundell et al. (ur.). 2010. *Macmillan collocations dictionary*. Oxford: Macmillan Education.

Gašper Uršič. 2015. *Oblikovanje uporabniškega vmesnika za spletni kolokacijski slovar*. Diplomsko delo. Ljubljana: Univerza v Ljubljani, Naravoslovnotehniška fakulteta.

Označevanje udeleženskih vlog v učnem korpusu za slovenščino

Simon Krek,*+ Polona Gantar,♦ Kaja Dobrovoljc,† Iza Škrjanec‡

* Laboratorij za umetno inteligenco, Institut »Jožef Stefan«, Jamova cesta 39, 1000 Ljubljana
+ Center za jezikovne vire in tehnologije Univerze v Ljubljani, Večna pot 113, 1000 Ljubljana
simon.krek@ijs.si
♦ Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@ff-uni.lj.si
† Zavod za uporabno slovenistiko, Trojina, Trg republike 3, 1000 Ljubljana
kaja.dobrovoljc@trojina.si
‡ Ljubljana
skrjanec.iza@gmail.com

Povzetek

V prispevku predstavimo postopek, nabor oznak, merila ter orodje za semantično označevanje učnega korpusa za slovenščino. V prvem delu prispevka predstavimo teoretična izhodišča raziskave in uporabljeno metodologijo, nato pa podrobno opišemo nabor oznak za semantično označevanje učnega korpusa za slovenščino in merila za njihovo določanje. Posebej izpostavimo konkurenčne udeleženske vloge in potencialne nove udeleženske vloge za razreševanje mejnih primerov. Prispevek zaključimo s kratkim povzetkom sprejetih odločitev in predvidenim nadaljnjim delom v okviru bilateralnega projekta Označevanje semantičnih vlog v slovenščini in hrvaščini.

Semantic Role Labeling in the Training Corpus for Slovene

The paper describes the procedure, tagset, criteria and tools for semantic role labeling in the training corpus for Slovene. In the first part we present the theoretical foundations of our research and the methodology. The following part includes a detailed description of the tagset used for semantic role labeling of Slovene, together with annotation criteria. Ambiguous cases are highlighted and potential now semantic roles are suggested for solving borderline cases. The paper finishes with a short summary of the decisions that were taken in the process, and future work in the context of the bilateral Slovene-Croatian project Semantic Role Labeling in Slovene and Croatian.

1 Uvod

Označevanje semantičnih vlog (ang. Semantic Role Labeling – SRL) je postopek, ki je z jezikoslovnega vidika namenjen (avtomatskemu) prepoznavanju udeleženskih vlog, z jezikovnotehnološkega pa razvoju sistemov za luščenje informacij, sistemov za odgovarjanje na vprašanja (ang. question answering system), izboljšavi delovanja skladenjskih razčlenjevalnikov ter strojnih prevajalnikov ipd. (Shen in Lapata, 2007; Christensen et al., 2011). Ker pomanjkanje konsenza glede različnih kategorij in meril za njihovo določanje, ki so danes sicer že na voljo za številne jezike, povzroča težave pri čezjezikovnem modelu semantičnega označevanja, mora po našem mnenju uspešen sistem meril in oznak za označevanje udeleženskih vlog ali natančneje predikatno-argumentnih razmerij (a) zagotavljati nabor kategorij, ki je kar najbolj optimalen, tj. pokriti vse (v našem primeru za slovenščino) ključne udeleženske vloge in hkrati (b) ne vsebovati kategorij, ki so prepodrobne ali medsebojno prekrivne, (c) temeljiti primarno na semantičnih in ne na morfoloških, leksikalnih ali skladenjskih lastnostih, (d) omogočati formalni opis oz. uporabnost v jezikovnotehnoloških aplikacijah ter (e) biti čim bolj kompatibilen s kategorijami in merili, ki veljajo za druge jezike (prim. Petukhova in Bunt, 2008: 39). V ta namen je bil v okviru projekta izdelave učnega korpusa za označevanje semantičnih vlog za slovenščino izdelan sistem meril za prepoznavanje in označevanje udeleženskih vlog za slovenščino. Naš cilj je bil ročno označiti polovico skladenjsko označenega dela učnega korpusa ssj500k,[1] na njegovi podlagi pa naj bi bilo v prihodnje mogoče avtomatsko označiti tudi obsežnejše korpuse.

V nadaljevanju prispevka predstavimo izhodišča za določitev semantičnih kategorij ter nabor oznak za slovenščino, postopek označevanja in orodje za semantično označevanje učnega korpusa za slovenščino.

2 Teoretično in metodološko ozadje

Pri izbiri metode semantičnega označevanja in določanju semantičnih kategorij za slovenščino smo najprej analizirali posamezne pristope, ki so bili razviti in uporabljeni za druge jezike, npr. PropBank (Palmer et al., 2005), Verbnet (Kipper et al., 2006) in FrameNet (Backer et al., 1998) za angleščino, AnCora (Taulé et al., 2011) za španščino, SoNaR (Schuurman et al., 2010) za nizozemščino. Poleg tega pa še nabor oznak za hrvaščino (Filko et al., 2012) in češki valenčni leksikon Vallex.[2] Osredotočili smo se na primerjavo formalnih opisov (tj. naborov semantičnih oznak) za posamezne udeleženske vloge ter meril za njihovo določanje. Z vidika optimizacije nabora oznak, ki bi zagotavljal dovolj robusten sistem in hkrati v čim večji meri upošteval specifike slovenščine, smo upoštevali še stopnjo semantične razdrobljenosti, ki jo predvideva posamezni sistem, in dejstvo, da za slovenščino nimamo na voljo strojno berljivega leksikona glagolske vezljivosti. Poleg tega smo merila za semantično označevanje želeli določiti tako, da bodo

[1] Opis in prenos korpusa: http://www.slovenscina.eu/tehnologije/ucni-korpus.

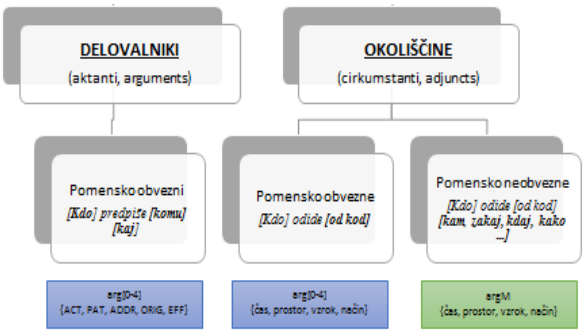
[2] Vallex: http://ufal.mff.cuni.cz/vallex.

omogočala zanesljivo in čim bolj konsistentno označenost učnega korpusa.

Pri naboru udeleženskih vlog in njihovih formalnih opisov smo izhajali iz funkcijskega generativnega pristopa Praške odvisnostne drevesnice (ang. Prague Dependency Treebank; PDT; Mikulová et al., 2006), ki znotraj obsega prototipičnega glagolskega vzorca (propozicije) določa razmerja med udeleženci, ki imajo lahko udeleženske vlogo delovalnika ali okoliščine. Udeležence in njihove različne udeleženske vloge določa torej prototipična prepozicija za dani pomen glagola, ki se uresničuje v njegovi vezljivostni shemi. Konkretno bi lahko rekli, da predvideva glagol *narediti* v enem od svojih pomenov tako delovalniške kot okoliščinske udeležence, ki se na oblikoskladenjski ravni realizirajo v obliki argumentov, ki jih je mogoče zapisati kot: kdo *naredi* komu kaj (kdaj, kje, kako, zakaj), kar predstavlja vezljivostno shemo konkretnega glagolskega pomena. Ob tem, da je število delovalnikov za dani primer predvidljivo, čeprav ne nujno realizirano v vsakem vzorcu omenjenega glagola oz. pomena, je mogoče za okoliščine reči le, da jih glagolski pomen predvideva, odprto pa ostaja vprašanje, ali so dejansko potrebne za ustrezno evokacijo glagolskega pomena. Na drugi strani je jasno, da tudi (ne)realizacija posameznih delovalnikov sama na sebi ne vpliva nujno na pomen glagola, saj govorec lahko udeležence predvideva, tudi če v vzorcu niso izraženi (Hanks, 2010; Žele, 2010). Tako velja, da obstaja v mentalnem leksikonu govorca za vsak glagolski pomen prototipična propozicija, ki se v realnem besedilu udejanja na različne načine, pri čemer izrablja tako oblikoskladenjski inventar jezika, vključno z elipsami in sobesedilnimi referencami, kot tudi zunajezikovne in pragmatične okoliščine izrekanja. Pri določanju razmerja med pomensko obveznimi in pomensko neobveznimi udeleženci smo zato v izhodišču izhajali iz sistema PropBank (Palmer et al., 2005), (Slika 1). Ta model določa pomensko obveznost le na ravni delovalnikov (določila), ki so vedno pomensko obvezni (arg[0-4]), medtem ko so okoliščine (prislovna dopolnila; argM) določene kot pomensko neobvezne, pri čemer, kot rečeno, pomenska obveznost ne implicira tudi strukturne obveznosti.

Slika 1: Obligatornost udeležencev v sistemih PropBank in VerbNet.

Na drugi strani sistemi, kot so PDT ter FrameNet (Backer et al., 1998), predvidevajo ugotavljanje obligatornosti tudi na ravni okoliščin (Slika 2). V tem primeru se tako pomensko obvezni delovalniki (ACT, PAT, ADDR, ORIG, EFF) kot pomensko obvezne okoliščine (čas, prostor, vzrok, način) označujejo z oznakami arg[0-4], pomensko neobvezne okoliščine pa z oznakami argM (gl. Tabelo 1).



Slika 2: Obligatornost udeležencev v sistemu PDT.

Ob združitvi obeh sistemov z vidika obligatornosti udeležencev, bi zgornji primer za glagol *narediti* izgledal takole:

Kdo	*naredi*	kaj	komu	kdaj	kje	kako	zakaj
arg0		arg1	arg2	{arg3-4}	{arg3-4}	{arg3-4}	
				ArgM			

Tabela 1: Vezljivostna shema glagola *narediti* s pripisom udeleženskih vlog in njihove obligatornosti v sistemu PDT.

Čeprav se zdi glede na pomenske lastnosti glagola ustreznejši sistem PDT, ki ob združitvi pomensko-skladenjskih meril prepoznava tudi obligatornost okoliščinskih udeležencev (prim. tudi Žele, 2010), obligatornosti za slovenščino ni bilo mogoče dosledno izpeljati brez leksikonskih podatkov za posamezni glagol. V trenutni fazi semantičnega označevanja smo zato ohranjali obligatornost le pri delovalnikih, pri okoliščinah pa te razmejitve nismo upoštevali, je pa to eden od izzivov, ki se jih nameravamo lotiti v nadaljnjih fazah označevanja.

3 Semantično označevanje za slovenščino: nabor oznak in merila za njihovo določitev

Osnovo za nabor udeleženskih vlog in njihovih oznak nam je, kot rečeno, predstavljal nabor oznak praške odvisnostne drevesnice. Z vidika optimizacije pomenske razdrobljenosti, upoštevanja slovenskih specifik in hkrati prekrivnosti oznak med posameznimi sistemi, smo nabor ustrezno zreducirali. Tabela 2 prikazuje združitev nabora delovalniških vlog glede na PDT ter nabor oznak za slovenščino.

Oznaka	PDT		SLO	
ACT	ACT	actor		**vršilec, aktant**
PAT	PAT	patient		**prizadeto**
REC	ADDR	addressee	prejemnik	**prejemnik**
	BEN	benefactor	koristnik	
ORIG	ORIG	origo	izvor	**izvor**
	HER	inheritence	podedovanost	
RESLT	EFF	effect		**učinek**

Tabela 2: Nabor oznak za delovalniške udeleženske vloge za slovenščino glede na PDT.

Praška odvisnostna drevesnica predvideva med okoliščinskimi udeleženci naslednje kategorije: čas, prostor, vzročnost in način, ki smo jih upoštevali tudi v

slovenskem naboru (Tabela 2). Pri notranji razčlenjenosti smo težili k združevanju pomenskih kategorij pod eno oznako. Tako smo npr. za različne časovne kategorije (*when, parallel, from when, to when; how long, since when* ipd. – skupaj 9), ki imajo v PDT ločene oznake, v slovenskem naboru združili v le tri: TIME (čas), DUR (trajanje) in FREQ (pogostnost). Hkrati smo določili, da oznaka TIME zajema semantične povezave, ki ustrezajo opredelitvam kot: *kdaj, sočasnost, z/od kdaj, do kdaj*, oznaka DUR določa povezave, ki opredeljujejo trajanje stanja ali dejanja (*kako dolgo, koliko časa*), oznaka FREQ pa pogostnost (*kako pogosto, kolikokrat*).

Tabela 3 prikazuje razmerje med PDT in slovenskimi oznakami za druge okoliščinske kategorije: prostor, vzrok in način.

	Oznaka	PDT		SLO	
PROSTOR	**LOC**	LOC	locative		**kraj**
		DIR2	which way	smer	
	SOURCE	DIR1	from		**začetna lokacija**
	GOAL	DIR3	where to		**končna lokacija**
VZROČNOST	**AIM**	AIM	aim	namen	**namen**
		INTT	intent	namera	
	CAUSE	CAUS	cause		**vzrok**
	CONTR	CNCS	concession	dopustnost	**protivnost**
		CONTRD	contradiction	protivnost	
	COND	COND	condition		**pogojnost**
OZIR	**REG**	REG	regard	ozir	**ozir**
		CRIT	criterion	merilo	
		CPR	comparison	primerjava	
	ACMP	ACMP	accompaniment		**spremstvo**
NAČIN	**RESTR**	RESTR	restriction	omejitev	**omejitev**
	MANN	MANN	manner	način	**način**
		RESL	result	rezultat	
	MEANS	MEANS	means		**sredstvo**
KOLIČ	**QUANT**	DIFF	difference	razlika	**količina**
		EXT	extent	količina	

Tabela 3: Nabor oznak za okoliščinske udeleženske vloge za slovenščino glede na PDT.

Pri označevanju večbesednih enot smo od nabora PDT ohranili le oznako DPHR (ang. dependant part of phraseme), ki smo jo preimenovali v PHRAS. Z njo označujemo frazeološke zveze tipa: *iti na živce*$_{PHRAS}$*, zaviti v molk*$_{PHRAS}$ ipd. Na novo smo v slovenskem naboru dodali oznako za zložene povedke MWPRED (ang. multi-word predicate), ki smo jo uporabili za označevanje zvez nedoločnika in faznega glagola, npr. *začeti vpiti*$_{MWPRED}$, ter za zveze nedoločnika in modalnega glagola, npr. *bo uspelo prepričati*$_{MWPRED}$*, zmore brati*$_{MWPRED}$*, niso želeli prikrajšati*$_{MWPRED}$. Zvez glagola in povedkovega prilastka nismo označevali s posebno oznako (PDT za te zveze uporablja oznako COMPL – ang. predicative complement), pač pa z delovalniško oznako (navadno RESLT): *zdela se mi je nekoliko bolj vsakdanja*$_{RESLT}$. Za označevanje modalnih glagolskih zvez smo uvedli oznako MODAL, npr. *je treba*$_{MODAL}$, *bi bilo mogoče*$_{MODAL}$. Zvez glagola z oslabljenim pomenom in samostalnika oz. samostalniške zveze (V PDT oznaka CPHR – ang. nominal part of the complex predicate) v tej fazi ne ločujemo od polnopomenskih vlog istih glagolov. To pomeni, da ne vzpostavljamo razlike med *dati ime* (PAT), *ne imeti namena* (PAT) – glagoli z oslabljenim pomenom – in *dati gol* (PAT), *imeti prijatelja* (PAT) – kjer gre za zvezo glagola in predmetnega določila. Razlikovanje med pomensko oslabljenimi glagoli kot sestavnimi deli glagolske zveze, prim. še *imeti na voljo, imeti v spominu,* in glagoli kot podeljevalci udeleženske vloge, *imeti denar*, bo prišlo do izraza pri prepoznavanju večbesednih enot, za katerega je predviden samostojni nivo označevanja učnega korpusa. To označevanje trenutno poteka v okviru COST akcije PARSEME kot del skupne naloge za identifikacijo večbesednih enot v različnih jezikih. Projekt predvideva v prvi fazi določitev formata in meril za označevanje večbesednih enot, v nadaljevanju pa ročno označenost približno 11.400 enot učnega korpusa.

Skupaj z oznakami smo na podlagi PDT določali tudi merila za njihovo prepoznavanje. V Tabeli 4 so poleg nabora oznak ter slovenskih imen zanje navedeni še kratki opisi udeleženskih vlog.

Udeleženska vloga		Opis
DELOVALNIKI	ACT	delujoči udeleženci, povzročitelji ali nosilci dejanja
	PAT	prizadeti predmet dejanja
	REC	prejemnik, posredni udeleženec dejanja; nedelovalniški udeleženec, ki mu je dejanje v škodo ali v prid
	ORIG	izhodišče, izvor/vir/povod dejanja
	RESLT	učinek, rezultat, cilj dejanja
OKOLIŠČINE	TIME	konkretni trenutek ali interval dejanja; kdaj
	DUR	trajanje stanja, dejanja koliko časa
	FREQ	frekvenca dejanja
	LOC	konkretna lokacija, kraj, mesto dejanja; kje
	SOURCE	začetna točka v prostoru; od kod
	GOAL	končna točka v prostoru; kam
	AIM	namen dejanja; čemu, s kakšnim namenom
	CAUSE	vzrok dejanja; zakaj
	CONTR	nepričakovana posledičnost dejanja; kljub čemu
	COND	pogoj za obstoj dejanja ali dogodka
	REG	glede na, primerjava
	ACMP	predmet, oseba ali dogodek, ki spremlja dejanje ali druge udeležence
	RESTR	izjema, omejitev
	MANN	načinovna lastnost dejanja, rezultat ob koncu dejanja
	MEANS	sredstvo ali orodje za izvedbo dejanja
	QUANT	količina, razlika
GLAGOLSKE ZVEZE	MWPRED	zveze z nedoločniki
	MODAL	zveze biti + modalnega prislova/pridevnika
	PHRAS	pomensko neprozorne zveze

Tabela 4: Merila za določanje udeleženskih vlog v učnem korpusu za slovenščino.

V slovenskem naboru smo od skupno 34 oznak v PDT ohranili 22 oznak (+ 2 za glagolske zveze), hkrati pa je analiza pokazala, da v nekaterih mejnih primerih potrebujemo podrobnejše pomenske opredelitve. V nadaljevanju opišemo nekatera ključna pomenska razmerja, ki zahtevajo natančnejšo opredelitev semantičnih oznak, in sicer tako z vidika pomenskih kot formalnih (skladenjskih in morfoloških) meril, ter potencialne dodatne/nove udeleženske vloge.

3.1 Konkurenčne udeleženske vloge/pomenska razmerja

Posamezna razmerja med udeleženci si z vidika določitve ustrezne udeleženske vloge pogosto konkurirajo. Do konkurenčnih povezav prihaja tako znotraj delovalnikov, npr. med vršilcem in prizadetim: *Dogodek v Ankaranu (ACT) je bila dramatična nesreča (PAT)*, kot tudi znotraj okoliščinskih razmerij, npr. med prostorskimi (LOC) in vzročnimi (CAUSE) ali časovnimi (TIME) udeleženskimi vlogami: *se dušijo v številkah (LOC→CAUSE), ministrica je na enem od sestankov (TIME→LOC) dejala*, ter med delovalniškimi in okoliščinskimi udeleženci, npr. pri prostorsko izraženih delovalnikih: *v bolnišnici (LOC→ACT] bodo uvedli,* o čemer več v nadaljevanju.

Razmerje vršilec – prizadeto pride do izraza predvsem pri upoštevanju skladenjskih razmerij, npr. pri razlikovanju med pasivnimi in aktivnimi zgradbami, kjer ločujemo med dvema oblikoskladenjskima možnostima izražanja trpnosti, tj. s prostim glagolskim morfemom *-se/-si* in z deležnikom na *-n/-t,* ter med trpnimi in aktivnimi povratnosvojilnimi skladenjskimi zgradbami. Na podlagi tega obravnavamo primere kot npr. *pozitivna diskriminacija* (PAT) *se označuje kot privilegij* kot pasivne, primere kot *dogodki* (ACT) *so se odvijali* pa kot aktivne povratnosvojilne skladenjske zgradbe. Pri trpnih zgradbah smo pozorni na to, da ostajajo udeleženske vloge v aktivnih in pasivnih skladenjskih zgradbah prekrivne, npr. *stvar* (PAT) *je malce bolj zapletena – kdo* (ACT) *zaplete stvar* (PAT); *pozitivna diskriminacija* (PAT) *se označuje kot privilegij – kdo* (ACT) *označuje pozitivno diskriminacijo* (PAT) *kot privilegij.*

V pomenskih razmerjih med delovalniškimi in okoliščinskimi udeleženci si konkurirajo zlasti predložne delovalniško-prostorske pomenske povezave kot npr. LOC→ACT/REC/RESLT: *v bolnišnici* (LOC→ACT) *so uvedli; povzročiti nevšečnosti na televiziji* (LOC→REC); *zaposliti v podjetju* (LOC→RESLT), in GOAL→ RESLT/REC: *to ga je spravilo v dobro voljo* (GOAL→RESLT); *Nokia je v paket* (REC→GOAL) *priložila polnilnik*. O prostorsko izraženih aktantih govorimo takrat, ko glagolski pomen ne predvideva prostorske komponente, kot jo predvidevajo npr. glagoli premikanja *priti, oditi, iti* itd. KAM, ampak dejavnost, npr. *v osnovni šoli* (ACT) *pripravljajo, pri Fujifilmu* (ACT) *so objavili, na centru* (ACT) *so se lotili*. Aktanti so v teh primerih dejansko metonimično izraženi delovalniki, hkrati pa vezljivostni vzorec glagola predvideva tudi potencialno okoliščinsko udeleženske mesto, ki se v nekaterih primerih tudi dejansko izraža, npr. *v osnovni šoli v Bistrici* (ACT) *pripravljajo.*

3.2 »Manjkajoče« udeleženske vloge/pomenska razmerja

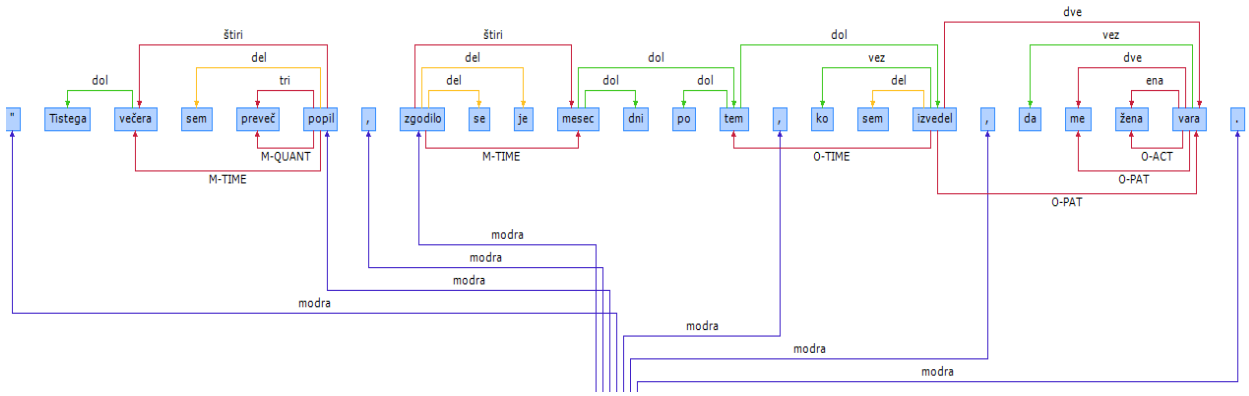
Pri glagolskih pomenih, ki predvidevajo obsežnejši vezljivostni vzorec, kot so npr. glagoli sporočanja, zaznavanja, mišljenja ipd., npr. kdo-ACT *reče, pove, izjavi* ... komu-REC kaj-RESLT o čem-PAT (kdaj, kje, kako, zakaj), se glede na možnost pripisa iste udeleženske vloge samo enemu udeležencu pojavlja potreba po pomensko podrobnejši razčlenitvi udeleženskih vlog.

Sem sodi razlikovanje med konkurenčnimi okoliščinskimi udeleženskimi vlogami, ki smo jih omenjali že v prejšnjem poglavju. V stavkih kot: *v drevišnjih tekmah bodo igrali*; *sploh ni padel v vojni*; *na enem od sestankov je dejala*, se podčrtanim udeležencem pripisuje bodisi časovna (TIME) bodisi prostorska (LOC) udeleženska vloga. Ker ob tem vezljivostni vzorec lahko predvideva več časovnih ali prostorskih udeležencev, kot npr. v stavku *Na veleslalomu za mladince na SP na Pohorju*, se odpira vsaj še prazno mesto »dogodka« (tekma, vojna, sestanek, SP), kot ga denimo predvideva sistem FrameNet (EVENT), ki združuje tako prostorsko kot časovno pomensko komponento. Podobno se za konkurenčni udeleženski vlogi načina (MANN) in prostora (LOC) v primerih kot: *informacijo po elektrodnem kablu pripeljejo v napravo; mimo se je pripeljala deklica,* ponuja možnost bolj specializirane udeleženske vloge, npr. »pot« (PATH), ki jo prav tako pozna sistem FrameNet.

4 Orodje in format označevanja učnega korpusa za slovenščino

Za semantično označevanje korpusa smo uporabili orodje SentenceMarkup, ki je bilo primarno razvito za skladenjsko označevanje slovenščine (Dobrovoljc et al., 2012). Orodje smo prilagodili za namene semantičnega označevanja tako, da smo mu dodali neodvisen in hkrati medsebojno povezljiv semantični nivo (Slika 3).

Ker želimo program v prihodnje nadgraditi za različne tipe označevanja (npr., za označevanje večbesednih enot), je pomembno, da zagotavlja čim večjo avtonomnost pri spreminjanju nabora oznak na več ravneh in možnosti tako ločenega kot kombiniranega iskanja po tipih povezav na skladenjski, pomenski ter drugih ravneh označevanja. Program omogoča izvoz podatkov v tabelarični obliki in XML formatu, ki poleg podatkov o tipu povezave na posameznem nivoju označevanja vsebuje tudi podatke o lemi, MSD-oznake ter omogoča izpis celotnega stavka.



Slika 3: Skladenjski in semantični nivo oznak v orodju SentenceMarkup.

5 Zaključek in prihodnje delo

V trenutni fazi semantičnega označevanja učnega korpusa je bil naš cilj določiti dovolj robusten in hkrati optimalen nabor udeleženskih vlog za slovenščino. Nabor oznak in merila za njihovo označevane smo določili na podlagi obstoječih označevalskih modelov, kjer smo izhajali zlasti iz PDT, v posameznih odločitvah pa smo upoštevali tudi rešitve v sistemu FameNet in drugih.

V postopku ročnega označevanja učnega korpusa smo, izhajajoč iz dejstva, da ne razpolagamo z leksikonom glagolske vezljivosti za slovenščino, težili k izboru udeleženskih vlog, ki omogočajo konsistentnost označevanja z upoštevanjem tako skladenjskega kot pomenskega nivoja. Pri konkurenčnih pomenskih oznakah smo zato skušali uvesti čim bolj jasna razločevalna merila, hkrati pa smo predlagali dodatne udeležene oznake, ki razrešujejo mejne primere. V prihodnje je naš namen na podlagi analize semantičnih povezav določiti tudi stopnjo obligatornosti tako pri delovalniških kot tudi pri okoliščinskih udeleženskih vlogah.

V naslednji fazi nameravamo v okviru bilateralnega projekta med Slovenijo in Hrvaško oblikovati sistem za označevanje semantičnih vlog, ki bodo pripisane obstoječim skladenjskim odvisnostnim povezavam v učnih korpusih, ki so uporabljeni za algoritme strojnega učenja za oba jezika. Vzorčni del slovenskega in hrvaškega učnega korpusa bo označen s kompatibilnimi oznakami, na njih pa bodo izpeljani tudi prvi eksperimenti avtomatskega označevanja z nadzorovanim strojnim učenjem. V okviru projekta bodo tako izdelana skupna navodila za označevanje semantičnih vlog v slovenščini in hrvaščini, orodje za označevanje semantičnih vlog za označevalce na obeh straneh, vzorčna učna korpusa za slovenščino in hrvaščino in eksperimentalno orodje, ki uporablja strojno učenje, za avtomatizacijo označevanja semantičnih vlog.

Del korpusa ssj500k je bil novembra 2015 vključen v zbirko skladenjskih drevesnic Universal Dependencies (UD) (Nivre et al., 2016). To omogoča, da se sistem označevanja semantičnih vlog, razvit v obstoječem sistemu JOS, prenese in preveri tudi v sistemu UD, kar je ena od prihodnjih nalog. Prenos je smiseln tudi s stališča kompatibilnosti med slovenskim in hrvaškim sistemom označevanja v okviru bilateralnega projekta, saj hrvaška drevesnica uporablja oznake UD.

6 Literatura

Collin F. Backer, Charles J. Fillmore in John B. Lowe. 1998. The Berkeley FrameNet project. *Proceedings of the COLING-ACL*. Montreal, Canada. 86–90.

Janara Christensen, Stephen Soderland Mausam in Oren Etzioni. 2011. An Analysis of Open Information Extraction based on Semantic Role Labeling. *International Conference on Knowledge Capture* (KCAP). Banff, Alberta, Canada. June 2011. 113–120.

Kaja Dobrovoljc, Simon Krek in Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. V T. Erjavec, J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 42–47.

Matea Filko, Daša Farkaš in 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.

Patrick Hanks. 2010. Elliptical Arguments: a Problem in relating Meaning to Use. S. Granger, M. Paquot (ur.): eLexicography in the 21st century: New challenges, new applications. *Proceedings of ELEX2009*. Cahiers du CENTAL. Louvain-la-Neuve: Presses universitaires de Louvain.

Karin Kipper, Anna Korhonen, Neville Ryant in Martha Palmer. 2006. Extensive Classifications of English verbs. *Proceedings of the 12th EURALEX International Congress*. Turin, Italy. September. 1–15.

Marie Mikulová et al. 2006. *Annotation on the tectogrammatical level in the Prague Dependency Treebank*. Aannotation manual. Technical Report 30. 5–11.

Joakim Nivre et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. V: *Proceedings of LREC'16*. 1659–1666.

Martha Palmer, Daniel Gildea in Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics* 31(1). 71–106.

Volga Petukhova in Henry Bunt. 2008. LIRICS semantic role annotation: Design and evaluation of a set of data categories. V *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008). Marrakech, Morocco. Paris: ELRA. 39–45.

Ineke Schuurman, Véronique Hoste in Paola Monachesi. 2010. Interacting semantic layers of annotation in sonar, a reference corpus of contemporary written dutch. *Proceedings of LREC'10*, Valletta, Malta. ELRA. 2471–2477.

Dan Shen in Mirella Lapata. 2007. Using Semantic Roles to Improve Question Answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning*. Prague. 12–21.

Mariona Taulé, Antònia M. Martí in Oriol Borrega. 2011. AnCora 2.0: Argument Structure Guidelines for Catalan and Spanish. *Working paper 4: TEXT-MESS 2.0 (Text-Knowledge 2.0)*. Universitat de Barcelona. Barcelona.

Andreja Žele. 2010. Elipsa med glagolsko intenco in besedilno koherenco (Izpust med glagolsko usmerjenostjo in besedilno soveznostjo). *Slavistična revija*, 58(1). 117–131.

Priprema usporedivih korpusa za usporedbu

Ivana Lalli Paćelat

Odjel za interdisciplinarne, talijanske i kulturološke studije, Sveučilište Jurja Dobrile u Puli,
I. M. Ronjgova 1, 52100 Pula
ilalli@unipu.hr

Sažetak

Osim formalne usklađenosti šest korpusa prema rasponu, opsegu i strukturi, zbog prirode planirane kvantitativne analize neizostavna je i njihova usklađenost na razini POS i MSD označavanja. Budući da je ta usklađenost samo djelomično prisutna prikazuju se u ovome radu potrebni postupci svođenja postojećih oznaka pod zajedničku oznaku kako bi rezultati bili usporedivi. No, i u slučaju potpune usklađenosti skupa oznaka sa smjernicama koje propisuju standarde, neizbjeţno je promišljanje i usklađivanje oznaka s obzirom na razlike u poimanju i postojanju gramatičkih kategorija u pojedinim jezicima, u slučaju ovoga rada hrvatskoga i talijanskoga jezika. Nakon što su se utvrdile i prikazale razlike među korpusima, koje su proizašle iz kontrastivne analize dvaju jezika, i u skladu s time odabrane moguće zajedničke oznake za vrste riječi i druge gramatičke kategorije, pribjeglo se postupku normalizacije korpusa. Radi postizanja bolje usporedivosti rezultata na međujezičnoj razini promatrala se tako distribucija unutar zajedničkih dijelova korpusa na način da cjelinu čine samo one oznake koje su zajedničke i relevantne za ciljno istraţivanje što je doprinijelo ujedno i većoj pouzdanosti rezultata.

Ovime se radom s jedne strane potvrdila vaţnost sustavnoga planiranja izrade skupa oznaka za vrste riječi i gramatičke kategorije za pojedini jezik u skladu s međunarodnim smjernicama koje propisuju standarde i stvaraju preduvjete usporedivosti među korpusima kako na unutarjezičnoj razini tako i na međujezičnoj razini, a s druge strane pokazalo kako se usporedba i usklađivanje MSD ili POS oznaka mogu smatrati dobrim temeljem i zanimljivim pristupom u kontrastivnoj analizi dvaju jezika.

Preparing comparable corpora for comparison

Although the six corpora included in the research were comparable with respect to size, purpose and structure, it was indispensable, due to the nature of the planned quantitative analysis, to make them comparable at the POS and MSD tagging level. Since the tagsets used to annotate the corpora were only partially compatible, several procedures were needed to convert the existing tags to a common tagset in order to have comparable results. However, also in case of full compatibility with international standards, it is inevitable to think about and to compare the tagsets because of the differences in the perception and in the existence of grammatical categories in different languages, i.e. Croatian and Italian. After the differences among the tagsets of the six corpora were identified, followed by a detailed contrastive analysis of the two languages and after the only possible common POS and MSD tagset was found, the normalization of the corpora was performed. In order to achieve better comparability of results at inter-lingual level only the distribution within the common, comparable and relevant tags were taken into account which contributed to greater reliability and accuracy of results.

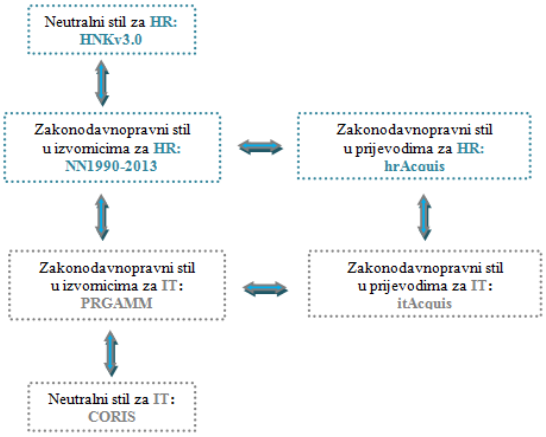
On one hand this paper confirmed the importance of systematic planning of linguistic annotation scheme for each language in accordance with guidelines which prescribe international standards and create conditions for the comparability across corpora at both inter-lingual and intra-lingual levels. On the other hand, the paper showed that comparing and analysing MSD or POS tagsets can be considered a good basis and an interesting approach for the contrastive analysis.

1 Uvod

Cilj je ovoga rada istaknuti vaţnost sustavnog planiranja izrade skupa oznaka za pojedini jezik u skladu sa smjernicama koje propisuju standarde i stvaraju preduvjete za usporedbu korpusa kako na unutarjezičnoj razini tako i na međujezičnoj razini. U slučaju kada takvi skupovi oznaka ne postoje, a priroda analize koja se planira provesti to zahtjeva, potrebno je uskladiti skupove oznaka korpusa koji se ţele uspoređivati.

Predstavit će se u ovome radu postupci pripreme korpusa za usporedbu, koji su bili nuţni kako bi se omogućila analiza registra i analiza univerzalnih obiljeţja prijevoda na hrvatsko-talijanskome jezičnom paru, a koje su detaljno prikazane u Lalli Paćelat (2014). Temeljni preduvjeti za analizu registra prema Biberu (1995) jesu: komparativni pristup, kvantitativna analiza i reprezentativni uzorak. Kako bi se zadovoljili navedeni uvjeti istraţivanje je provedeno na šest različitih računalnih korpusa odnosno četiri vrste korpusa za oba jezika: referentni općejezični jednojezični korpusi (1) Hrvatski nacionalni korpus (HNK v 3.0) i (2) Corpus di italiano scritto (CORIS); (3) specijalizirani dvojezični usporedivi korpus (potkorpusi referentnih korpusa: NN1990-2013/PRGAMM); (4,5) jednojezični usporedivi korpusi izvornika i prijevoda na istome jeziku istoga stila (NN1990-2013/hrAcquis i PRGAMM/itAcquis) (6) i usporedni korpus hrvatskih i talijanskih prijevoda (hrAcquis/itAcquis). Više o korpusima u Tadić (2009) i Rossini Favretti i dr. (2002).

Komparativni pristup pretpostavlja usporedbu ciljnoga funkcionalnog stila s drugim stilovima. U planiranome istraţivanju usporedit će se tako zakonodavnopravni stil s neutralnim stilom referentnoga korpusa koji čini skup različitih funkcionalnih stilova jednoga jezika, s istim stilom drugoga jezika te s prevedenim tekstovima istoga stila na istome jeziku. Komparativni pristup i analiza u navedenome istraţivanju izgledat će kao što je prikazano na Slici 1.



Slika 1: Komparativni pristup.

U sličnim istraţivanjima (Neumann, 2013; Teich, 2003; Xiao 2010 i dr.) istraţivači su koristili korpuse u čijem su obiljeţavanju odnosno označavanju posredno ili neposredno sudjelovali, a svi su korpusi bili označeni na isti način. Suprotno njima, u navedenome istraţivanju koristit će se korpusi koji su obiljeţeni i označeni neovisno jedni o drugima zbog čega su samoj analizi prethodile provjere usporedivosti korpusa i usklađivanja skupova oznaka. To je usklađivanje bilo potrebno kako bi se omogućilo jasnije prikazivanje rezultata, ali i kako bi se osigurala visoka razina usporedivosti među korpusima odnosno među jezicima na temelju vrsta riječi i drugih gramatičkih kategorija.

2 Obilježavanje hrvatskih i talijanskih korpusa

Od ključne je vaţnosti prilikom sastavljanja korpusa njegova usklađenost sa svjetskim standardima za obiljeţavanje. Korpusi koji će se koristiti u navedenome istraţivanju usporedivi su na najvećoj mogućoj razini: prema rasponu, opsegu, i što je jako vaţno za ovaj rad i prema strukturi. Tipologija tekstova, njihov omjer i druge vaţne smjernice za izradu korpusa, kako za HNK-a, tako i za CORIS, usklađene su koliko je u danim uvjetima bilo moguće s preporukama EAGLES-a (1996). O strukturi HNK-a u Tadić (1996), (1998), (2003), a CORIS-a u Rossini Favretti (2000).

Međutim, osim temeljnih pretpostavki usklađenosti potrebnih radi uspoređivanja korpusa, zbog prirode analize koja se planira provesti, neizostavna je i usklađenost na razini POS i MSD obiljeţavanja odnosno označavanja. Ta je usklađenost samo djelomično prisutna, kao što će se u nastavku rada i predstaviti.

POS i MSD označavanje hrvatskih tekstova, kako onih u HNKv3.0, naravno onda i potkorpusu NN1990–2013, tako i onih u hrvatskome prijevodnom korpusu hrAcquis, obavljeno je prema MULTEXT-East (v4.0) specifikaciji (Erjavec, 2004).

Za talijanski je jezik razrađeno niz preporuka za obiljeţavanje korpusa u okviru EAGLES projekta (Monachini, 1995). Isto tako, nekoliko je istraţivačkih skupina radilo na POS označavanju te je svaka skupina razradila svoj način označavanja i svoj skup oznaka. Više o usporedbi skupa oznaka za vrste riječi u talijanskome jeziku u Bernardi i suradnici (2005), (2006), Tamburini i suradnici (2008), Venturi (2009) i dr. Tamburini i suradnici (2008: 97) ističu kako postoji načelna suglasnost oko oznaka za osnovne vrste riječi, dok značajne razlike postoje u kriterijima za dodjeljivanje oznaka za podvrste riječi. Neusuglašenost u korištenju oznaka za POS i MSD označavanje talijanskih korpusa znatno oteţava njihovu međusobnu usporedbu, a onda i usporedbu s korpusima na drugim jezicima. S obzirom na to da je MULTEXT preporuka standardiziranoga sustava za označavanje gramatičkih kategorija bila razrađena u suradnji s EAGLES inicijativom iz 1996., a da je MULTEXT-East bio samo „srednjo- i istočnoeuropski odvjetak“ toga projekta (Tadić, 2003: 107) te da su i sastavljači korpusa navodili prihvaćanje smjernica EAGLES-a očekivalo se da neće biti razlike barem u oznakama temeljnih gramatičkih kategorija za talijanski i za hrvatski jezik. Iako se skup oznaka izrađenih za označavanje talijanskoga referentnoga korpusa naziva „EAGLES-like tagset for Italian“ (Tamburini, 2000, Monachini, 1995) oznake su nešto drugačije od onih standardiziranih. Tamburini (2000) ističe da je skup oznaka primijenjen u CORIS-u u skladu sa standardima EAGLES-a razrađenih za talijanski jezik u Monachini (1995). S obzirom na već spomenutu neusuglašenost i neslaganje oko označavanja vrsta riječi za talijanski jezik autor je uzeo u obzir preporuke i primjere navedene u on-line inačici rječnika De Maura iz 2007. Skup oznaka korištenih za označavanje itAcquisa sastavljen je po uzoru na skup oznaka za španjolski jezik (Prokopidis et al., 2012) i sličniji je skupu oznaka za hrvatski jezik odnosno bliţi j e smjernicama EAGLES-a.

Najveća razlika između talijanskih i hrvatskih korpusa koji se rabe u planiranome istraţivanju odnosi se na razinu označavanja. Dok su talijanski korpusi označeni samo na razini vrste riječi (POS), hrvatski su korpusi označeni na način da je svakoj gramatičkoj kategoriji dodijeljena i vrijednost. Stoga svaka pojavnica u hrvatskim korpusima ima morfosintaktički opis (MSD), dok takav opis nedostaje kod talijanskoga referentnog i specijaliziranog korpusa, a samo djelomično postoji kod prijevodnoga korpusa.

Ako se uzme u obzir do sada spomenuto, odnosno da svaka pojavnica u talijanskim korpusima ima oznaku o vrsti riječi, a da pojavnice u hrvatskim korpusima imaju oznake koje, osim podataka o vrsti riječi, donose i vrijednosti relevantnih morfosintaktičkih kategorija te da su same oznake, nazivi za iste gramatičke kategorije, različiti, jasno je da je u uspoređivanju hrvatskih i talijanskih prvi korak bio stvaranje preduvjeta za morfosintaktičku usporedbu hrvatskih i talijanskih korpusa.

3 Usklađivanje korpusa

3.1 Svođenje na zajedničke oznake

Prilikom uspoređivanja svih šest korpusa koristit će se samo one oznake koje su zajedničke svim korpusima, dok će se detaljnije analize prikazati samo onda kada se budu analizirali korpusi samo na jednome jeziku. Kako bi i na razini vrsta riječi hrvatski i talijanski korpusi bili usporedivi, bilo je potrebno uskladiti oznake i svesti ih pod zajednički nazivnik prvenstveno radi jasnijega prikazivanja rezultata. U odabiru zajedničkih oznaka odlučilo se prikloniti standardiziranome skupu oznaka prema MULTEXT-East (v3.0) (Tadić iz 1998. u Erjavec, 2004) preporuci.

Za svaku će se vrstu riječi prikazati u nastavku svođenje na zajedničku oznaku. Valja podsjetiti da sva tri hrvatska korpusa koriste isti skup oznaka, pa će se u nastavku rada i prikazima navoditi samo oznake u HNK-u, a vrijedit će i za korpus NN1990-2013 i hrAcquis. Što se tiče talijanskih korpusa, uspoređivat će se dva skupa oznaka, jedan koji se koristi u CORIS-u, koji vrijedi i za potkorpus PRGAMM te drugi koji se koristi za itAcquis.

3.1.1 Imenice i prilozi

Svođenje oznaka imenica pod isti nazivnik nije predstavljalo problem. U svim korpusima razlikuju se opće i vlastite imenice, kao što je to prikazano u Tablici 1.

itAcquis			CORIS	HNK	Odabrana oznaka i objašnjenje	
S			N	N	N	imenica
Ss	Sp	Sn	NN	Nc	Nc	opća imenica
SP			NN_P	Np	Np	vlastita imenica
SWs	SWp	SWn				
B			ADV	R	R	prilog
				Qr		
BN				Qz		

Tablica 1: Prijedlog zajedničkih oznaka za imenice i priloge.

Problematični nisu bili ni prilozi za koje prema skupovima oznaka nije predviđena daljnja podjela na vrste priloga. Valja napomenuti da u hrvatskim korpusima postoji daljnja podjela na vrste priloga, koja međutim nije dokumentirana u specifikaciji za hrvatski jezik MULTEXT-East-a. Pod oznaku priloga (R) u hrvatskim korpusima uključit će se i dio pojavnica s oznakom čestica, kao što prikazuje Tablica 1. S obzirom na to da u talijanskim korpusima ne postoji posebna oznaka za čestice budući da u tradicionalnim talijanskim gramatikama čestice (tal. *particelle*) nemaju status vrste riječi, smatralo se to dobrim rješenjem za dio hrvatskih čestica koje se u talijanskome jeziku smatraju prilozima. O spornome statusu talijanskih čestica mnogo se pisalo, a čak i o njihovome odnosu prema česticama u hrvatskome jeziku, primjerice u Tekavčić (1989), i Jernej (1990). Pojavnice koje će priključiti prilozima u hrvatskim korpusima imaju oznake Qr i Qz. Riječ je o potvrdnim odnosno niječnim česticama koje se u talijanskome jeziku tradicionalno svrstavaju među priloge (tal. *avverbi di affermazione/negazione*)[1].

3.1.2 Glagoli

Suprotno imenicama i prilozima, svođenje oznaka glagola na zajedničku nije bilo jednako zahvalno, kao što je prikazano u Tablici 2.

itAcquis	CORIS	HNK	Odabrana oznaka i objašnjenje	
V	V	V	V	glagol
VA	V_ESSERE	Va	Va	pomoćni glagol
	V_AVERE	Vc		
VM	V_MOD	Vm	Vm	glagolski oblik

Tablica 2: Prijedlog zajedničkih oznaka za glagole.

[1] Valja napomenuti da se stavovi talijanskih lingvista razilaze se po pitanju opravdanosti svrstavanja jesno-potvrdnih čestica pod priloge.

Zbog različitih podjela unutar svakoga korpusa bilo je moguće svesti na samo dvije zajedničke oznake. S obzirom na to da se u oba jezika podjela na pomoćne i kopulativne glagole uglavnom podudara, odlučilo se za zajedničku oznaku (Va). Treba međutim imati na umu nekoliko mogućih razlika. Pomoćni su glagoli oni koji služe za tvorbu složenih glagolskih oblika, a u hrvatskome su jeziku to *biti* i *htjeti*, dok su u talijanskome to *essere* i *avere* (Silić i Pranjković, 2005: 185; Dardano i Trifone, 2003: 200). Valja spomenuti da se u talijanskome jeziku i glagoli *venire, andare, dare* i *stare* pojavljuju u ulozi pomoćnih glagola u tvorbi određenih glagolskih oblika. Pod kopulativnim glagolima podrazumijeva se glagol *biti* odnosno *essere* (usp.Težak i Babić, 1994: 198 i dr.), no postoje i drugi glagoli, koji vrše istu funkciju, no nekad se nazivaju polukopulativni (Silić i Pranjković, 2005: 269) ili kopulativni glagoli (Dardano i Trifone, 2003: 192), međutim u korpusima koji se koriste u navedenome istraživanju nisu označeni kao kopulativni.

Isto tako u hrvatskome jeziku ne postoje glagolski oblici združeni s nenaglašenim zamjenicama (poput *dillo, prendertelo, andarci, affittasi* i sl.), a glagolski pridjevi i prilozi ne čine posebnu kategoriju već su svrstani pod oznaku glavnih glagola (Vm). Potrebno je napomenuti da u talijanskome jeziku postoje *participio presente* i *participio passato*. S vremenom je particip prezenta (npr. *amante, vincente, studente*) izgubio glagolska svojstva te se danas smatra uglavnom pridjevom ili imenicom (Dardano i Trifone, 2003: 245). Svoju glagolsku vrijednost zadržao je samo u administrativnome funkcionalnom stilu, ističu Dardano i Trifone (2003: 246). Glagolski pridjev radni i trpni ostvaruju se u talijanskome jeziku istim oblikom, participom prošlim, što znači da ih nije moguće odvajati samo prema obliku. To onemogućava analizu distribucije pasivnih glagolskih oblika u talijanskim korpusima budući da nemaju posebno označen podatak o radnom ili trpnom obliku, dok ga hrvatski korpusi imaju. S druge strane talijanski *gerundio* ima dva oblika koji odgovaraju hrvatskim oblicima glagolskih priloga. Glagolski prilog radni odgovara tako talijanskome *gerundio presente*, dok glagolski prilog prošli odgovara talijanskome *gerundio passato* (Dardano i Trifone, 2003: 246). Dok bi se iz hrvatskih korpusa i itAcquisa glagolski pridjevi i glagolski prilozi mogli izvući i pretragama s MSD odnosno POS oznakama, takvu pretragu nije moguće obaviti za CORIS gdje su označeni samo glagolski pridjevi. U CORIS-u je moguća samo djelomična pretraga pomoću regularnih izraza.

Jedino što je moguće, jest, kao što to predočuje Tablica 2, razlikovanje pomoćnih (Va) i glavnih glagola (Vm). Valja pojasniti da se pod oznaku glavnih glagola (Vm) ne misli samo na samoznačne glagole, jer uključuju i modalne, fazne i perifrazne glagole koji su suznačni. Iako bi potpuno odvajanje suznačnih od samoznačnih glagola bilo zanimljivo, takva pretraga s oznakama nije ostvariva za svih šest korpusa. Moguće je, međutim, da već i odvajanje glavnih i pomoćnih glagola može biti pokazateljem neke tendencije, koju je potrebno detaljnije obraditi.

3.1.3 Zamjenice i pridjevi

Kategorije zamjenica i pridjeva vrlo se različito poimaju u hrvatskome i talijanskome jeziku zbog čega je svođenje spomenutih kategorija pod istu oznaku bilo vrlo zahtjevno, kao što to pokazuje Tablica 3.

itACQUIS			CORIS	HNK	Odabrana oznaka i objašnjenje	
A			ADJ	A	A	pridjev
As	Ap	An	ADJ	Af	Af	opisni pridjev
			~~ADJ_NUM~~	~~Ao~~	~~Ao~~	~~brojevni pridje~~
APs	APp	APn	ADJ_POS	Ps	Psx	posvojna zamj./pridjev (it)/povratno-posvojna zamj. (hr)
			PRON_POS	Px		
P			PRON	P	P	zamjenica
PE			PRON_PER	Pp	Ppx	lična zamj./ povratna zamj. (hr)
PC				Px		
PD			PRON_DIM	Pd	Pd	pokazna zamj. /pridjev (it)
DD			ADJ_DIM			
PI			PRON_IND	Pi	Piqr	neodređena zamj./pridjev (it)/upitna zamj./pridjev (it)/odnosna zamj./pridjev (it)
DI			ADJ_IND			
PQ			PRON_IES	Pq		
DQ			ADJ_IES			
PR		DR	PRON_REL	Pr		

Tablica 3: Prijedlog zajedničkih oznaka za pridjeve i zamjenice.

U Tablici 3 mogu se zamijetiti i značajne razlike u oznakama kod talijanskih korpusa.

Dok se skup oznaka u CORIS-u drţi podjela tradicionalnih talijanskih gramatika, skup oznaka itAcquisa ima nešto drugačiju podjelu. Kod skupa oznaka itAcquisa kod promijenjivih riječi posebnu oznaku imaju oblici u jednini (s), mnoţi ni (p) i oni koji su neutralni što se tiče broja (n), zbog čega su oznake u tablicama usporedno prikazane. Zanimljivo je zamijetiti da je ono što se u CORIS-u označava kao pokazni, neodređeni i upitni pridjev (ADJ_DIM, ADJ_IND, ADJ_IES) u itAcquisu označeno kao pokazni, neodređeni, upitni, ali i odnosni „determinante“ (DD, DI, DQ) prema uzoru na španjolski skup oznaka. Riječ je o posebnoj oznaci (D) za vrstu riječi koja uključuje i članove, što svakako ukazuje na njezin sporan status.

Iz ovoga kratkog prikaza samo nekoliko uočenih razlika u označavanju talijanskih korpusa, jasno je da odabir oznaka ovisi o teoriji na koju se oslanja te da nije u potpunosti teorijski neutralno. Ovaj je pokušaj usklađivanja oznaka dobar primjer kako nije dobro imati različite skupove POS i MSD oznaka za isti jezik jer se na taj način onemogućava unutarjezična analiza korpusa označenih različitim označivačima koji se nisu vodili istim smjernicama, a što u konačnici ograničava njihovu primjenu.

Ako se usporede samo oznake, koje donosi Tablica 3, vidljivo je da u oba jezika postoje posvojni pridjevi. Međutim, ono što se u talijanskome jeziku smatra posvojnim pridjevom u hrvatskome se jeziku smatra posvojnom zamjenicom. Definicije se, kako posvojnih zamjenica u hrvatskome, tako i talijanskih posvojnih pridjeva, podudaraju: obje kategorije iskazuju pripadnost te se određuju prema licima (usp. Silić i Pranjković, 2005: 123; Dardano i Trifone, 2003: 138). U hrvatskome jeziku smatraju se zamjenicama jer zamjenjuju posvojne pridjeve (Teţak i Babić, 2005: 127).

U Primjeru 1 prenosi se rečenica koju Silić i Pranjković (2005: 123) navode kao jedan od primjera uporabe posvojnih zamjenica. Ista će se rečenica prevesti na talijanski jezik. Posvojni će pridjev odnosno posvojna zamjenica, kao istovrijednice, biti posebno istaknuti.

Tvoja profesorica dobro pjeva
La tua insegnante canta bene.

Primjer 1: Posvojne zamjenice u hrvatskome jeziku i posvojni pridjevi u talijanskome jeziku.

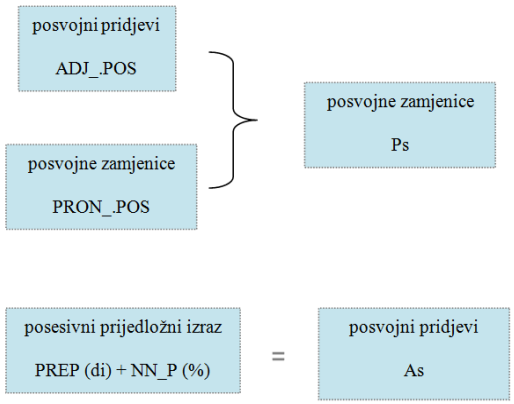
Osim oblika posvojnih zamjenica pod posvojne pridjeve u talijanskome spadaju još i pridjevi *altrui* i *proprio* (Dardano i Trifone, 2003: 139). U talijanskome jeziku posvojne zamjenice i posvojni pridjevi jednaki su po obliku, podsjeća Jernej (2005: 126). Ulogu onoga što se u hrvatskim gramatikama naziva posvojnim pridjevima, u talijanskome jeziku preuzeli su prijedloţn i izrazi *complemento di specificazione* točnije *complemento di specificazione di appartenenza o di specificazione possessiva* (usp. Dardano i Trifone, 2003: 139; Jernej, 2005: 290). Primjer 2 prema uzoru na primjere ponuđene u Silić i Pranjković (2005: 123) to zorno prikazuje. Njihovi će se primjeri prevesti na talijanski jezik, a posvojna zamjenica odnosno pridjev u hrvatskome i prijedloţn i izraz u talijanskome jeziku biti će posebno istaknuti.

Njegova sestra studira engleski.
Sua sorella studia inglese.

Petrova sestra studira engleski.
La sorella di Pietro studia inglese.

Primjer 2: Posvojna zamjenica odnosno pridjev u hrvatskome i prijedloţn i izraz u talijanskome jeziku.

Moguće rješenje u usporedbi talijanskih i hrvatskih korpusa saţeto je prikazano na Slici 2. Lijevi dio prikaza odnosi se na talijanski jezik, dok se desni dio prikaza odnosi na hrvatski jezik.

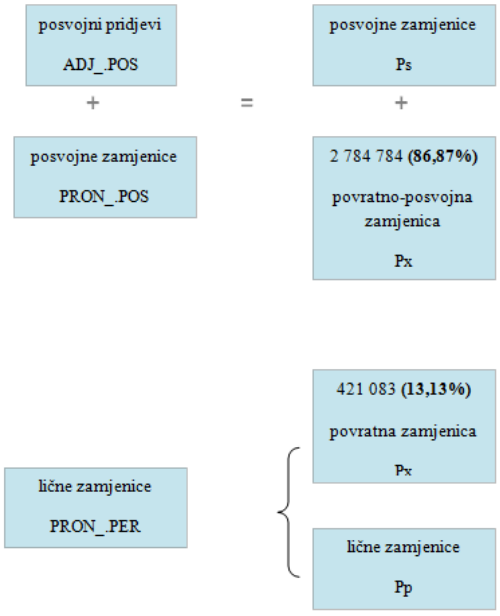


Slika 2: Mogućnosti kontrastivne analize posvojnih zamjenica odnosno pridjeva u postojećim korpusima.

Kada se uzme u obzir sve do sada rečeno, proizlazi da je vrlo teško odijeliti samo posvojne pridjeve jer treba uzeti u obzir da se oni u talijanskome jeziku drugačije ostvaruju, a budući da su talijanski tekstovi označeni samo na razini vrste riječi, potrebno bi bilo uzeti u obzir određeni broj prijedloga i određeni broj vlastitih imenica. Međutim, ono što se svakako može uspoređivati na temelju postojećih oznaka i korpusa jesu posvojne zamjenice u hrvatskome koje odgovaraju talijanskim posvojnim zamjenicama i posvojnim pridjevima. Kada je riječ o zamjenicama, problematična je i pripadnost povratnih zamjenica. Pod oznaku Px u hrvatskim korpusima spadaju povratna (*se, sebe*) i povratno-posvojna zamjenica (*svoj*), dok se u talijanskome jeziku povratne zamjenice svrstavaju pod lične zamjenice i broje više oblika (*mi, ti, si, ci vi si*), a povratno-posvojna zamjenica spada pod posvojne zamjenice odnosno posvojne pridjeve (*suo*). Stoga se broj pojavnica označenih s Px morao podijeliti u dvije skupine. Za podjelu se koristila sljedeća pretraga s regularnim izrazima:

1) [msd="Px.*pn.*"]
Pojašnjenje: p→personal n→nominal
2) [msd="Px.*nsa.*"]
Pojašnjenje: s→possessive a→adjectival

Kao što je vidljivo iz pretrage i same oznake upućuju na njihovu pripadnost, pa će se povratne zamjenice radi usklađivanja s talijanskim oznakama uključiti u lične zamjenice, dok će se povratno-posvojna zamjenica priključiti skupini hrvatskih posvojnih zamjenica odnosno talijanskih posvojnih zamjenica i posvojnih pridjeva u omjeru prikazanome na Slici 3.



Slika 3: Prijedlog usklađivanja oznaka povratne i povratno-posvojne zamjenice.

3.1.4 Prijedlozi, veznici i brojevi

Nakon usklađivanja oznaka zamjenica i pridjeva, prelazi se na ponešto jednostavnije usklađivanje oznaka prijedloga, veznika i brojeva, kao što prikazuje Tablica 4.

itAcquis	CORIS	HNK	Odabrana oznaka i objašnjenje	
E	PREP	Sp	Sp	prijedlog
EA	PREP_A			
Cc	CONJ_C	Ccs	Cc	veznik nezavisnosloženih rečenica
		Ccc		
Cs	CONJ_S	Css	Cs	veznik zavisnosloženih rečenica
		Csc		
NOs NOn	ADJ_NUM	Ao	M	brojevi i brojevni pridjevi
NUM	C_NUM	M		

Tablica 4: Usklađivanje oznaka prijedloga, veznika i brojeva.

Dok se u talijanskome razlikuju jednostavni prijedlozi i prijedlozi združeni s članom, u hrvatskome postoji i podjela na jednostavne i složene prijedloge, koja ovom prilikom nije prikazana jer nije usporediva s talijanskom podjelom. Budući da talijanska podjela prijedloga ne postoji u hrvatskome jeziku, svi prijedlozi obaju jezika svedeni su pod jednu zajedničku oznaku (Sp).

Što se tiče veznika, u oba jezika razlikuju se veznici nezavisnosložene rečenice i zavisnosložene rečenice. Hrvatski korpusi spomenute veznike dijele još i na jednostavne i složene. Radi pojednostavljivanja smatra se dovoljnim podjela na veznike nezavisnosložene rečenice ili konjunktore i na veznike zavisnosložene rečenice ili subjunktore, kao što je prikazano u Tablici 4.

Usporedbom tradicionalnih podjela na vrste riječi u hrvatskim i talijanskim gramatikama uočava se razlika što se tiče statusa brojeva. Brojevi (na tal. *numerali*) se uglavnom u talijanskome jeziku svrstavaju među pridjeve (*aggettivi cardinali, ordinali e multiplicativi*; npr. Dardano i Trifone, 2003: 138). Naravno, postoje i u obliku brojevnih imenica i prijedložnih izraza, no u manjem broju, a odnose se na posebne brojeve (*numerali frazionari, distributivi i collettivi*; npr. Dardano i Trifone, 2003: 150; Faloppa, 2011 i dr.). Brojevi se u talijanskim gramatikama ne smatraju posebnom vrstom riječi (Dardano i Trifone, 2003; Lo Duca, 2011; Faloppa, 2011 i dr.), dok oni u hrvatskome uživaju status zasebne vrste riječi (Barić et al., 1997; Težak i Babić, 2005; Silić i Pranjković, 2005 i dr.). Problematiku statusa brojeva kao vrstu riječi i općenito morfoloških obilježja brojevnih riječi opširno je prikazala za hrvatski jezik Tafra (1989), (2000), (2005), a za talijanski jezik Faloppa (2011) i dr.

S obzirom na to da pod oznaku za talijanske brojevne pridjeve (ADJ_NUM) iz CORIS-a spada i dio pojavnica koje su u hrvatskim korpusima označene kao brojevi (M.*l), a ostatak pojavnica s oznakom brojeva (M.*d i M.*r) odgovara oznaci brojeva (C_NUM) u CORIS-u te da slično vrijedi i za itAcquis, u kojemu se za razliku od CORIS-a ne dodijeljuje oznaka pridjeva, bilo je važno svesti usporedive dijelove pod jednu zajedničku oznaku. Rješenje spomenutih nepodudarnosti u označavanju korpusa, ali i u poimanju brojeva u hrvatskim i talijanskim tradicionalnim gramatikama, pronađeno je, za spomenuto istraživanje, u odvajanju brojevnih pridjeva od ostalih pridjeva te u njihovom uključivanju pod zajedničku oznaku brojeva (M), kao što je detaljno prikazano u Tablici 4.

3.2 Problem člana

Prvi rezultati usporedbe omjera vrsta riječi u hrvatskome i talijanskome jeziku, pokazali su znatne razlike u distribuciji riječi u referentnim korpusima.

Prilikom promatranja rezultata bilo je vaţno uzeti u obzir i činjenicu da je „talijanski sustav djelomično analitički, dok je hrvatski u potpunosti sintetički, s vrlo sloţenim sklonidbenim sustavom i slobodnim redom riječi“ (Sočanac, 2004: 151). Ta činjenica ima kao posljedicu veliku razliku u broju prijedloga, koji su u talijanskome jeziku preuzeli i funkciju padeţnih nastavaka (Sočanac, 2004: 151). Teško je, međutim, odrediti točan omjer prijedloga s takvom funkcijom u talijanskome jeziku. Još jedna bitna razlika među jezicima jest nepostojanje člana u hrvatskome jeziku. Suprotno razlici u prijedlozima, ova će se razlika, koja još u većoj mjeri utječe na usporedbu distribucije vrsta riječi u hrvatskim i talijanskim korpusima, pokušati neutralizirati. Način na koji će se to provesti i njegova opravdanost prikazat će se u nastavku rada.

Podsjetit će se na početku da vrijednost člana većinom ostaje implicitna u hrvatskome, dok je ona u talijanskome eksplicitna (Ljubičić, 2000: 226). Talijanski određeni član nastao je od pokaznoga pridjeva, a u nekim je slučajevima do danas sačuvao pokazno značenje, koje se onda i očituje u prijevodu uporabom pokazne zamjenice u hrvatskome jeziku (Ljubičić, 2000: 179). S druge strane, Ljubičić (2000: 193–194) upozorava kako je „teško izvršiti razgraničenje između neodređenoga člana i broja, ali još je teţe odijeliti broj i član od neodređenoga pridjeva“ (u hrvatskome zamjenice) te ističe kako se neodređeni član ponekad prevodi neodređenim pridjevom (u hrvatskome zamjenicom) kada član nema samo gramatičko značenje, nego i ono leksičko.

Međutim i Ljubičić (2000: 228) i Karlić (2014) slaţu se da samo u slučajevima kada članovi nose i leksičko značenje te kada nose komunikacijski relevantnu informaciju koja se ne moţe zaključiti iz konteksta, onda i jezici, koji nemaju član, poput hrvatskoga, mogu eksplicitno izraziti vrijednost člana drugim sredstvima.

Izraţava li se zaista i koliko često vrijednost člana u hrvatskome jeziku provjerilo se na odabranome uzorku korpusa itAcquisa i hrAcquisa[2]. Valja napomenuti da se u oba slučaja radi o prijevodima s istoga jezika, dakle oba prijevoda imaju isti status te jednaki utjecaj izvornika s obzirom na to da među njima ne postoji izravan prijevodni proces. Test je pokazao da u talijanskim prijevodima članovi čine od 7,7% do čak 9,2% ukupnoga broja pojavnica. Isto tako, test je pokazao da od ukupnoga broja članova u talijanskim usporednim tekstovima, maksimalno se za njih 4%, a nekad i mnogo manje, vrijednost eksplicitno iskazuje drugim sredstvima. Da se vrijednost članova u većini slučajeva ne iskazuju u hrvatskome jeziku pokazuje Primjer 3.

1) „...u vezi s kvalitativnim Ø značajkama riţe ...“(jrc31999R0691)
„...per quanto riguarda le caratteristiche qualitative del riso...“(jrc31999R0691)

2) „...za potvrđivanje usklađenosti Ø određenog proizvoda ili Ø obitelji proizvoda...“(jrc31999R0691)
„...se per un dato prodotto o un gruppo di prodotti determinati...“(jrc31999D0089)

3) „...je li Ø postojanje nadzornog Ø sustava tvorničke proizvodnje za koji je odgovoran proizvođač potreban i dovoljan Ø uvjet...“(jrc31999D0089)
„...l’esistenza nella fabbrica di un sistema di controllo della produzione, effettuato dal fabbricante, sia una condizione necessaria e sufficiente...“(jrc31999D0089)

4) „...Ø Uredba (EZ-a) br. 708/98 ovime se izmjenjuje i dopunjuje kako slijedi ...“(jrc31999R0691)
„...Il regolamento (CE) n. 708/98 è modificato come segue...“(jrc31999R0691)

5) „...Ø EUROPSKI PARLAMENT I Ø VIJEĆE EUROPSKE UNIJE...“(jrc32006L0012)
„...IL PARLAMENTO EUROPEO E IL CONSIGLIO DELL’UNIONE EUROPEA...“(jrc32006L0012)

Primjer 3: Primjeri neiskazivanja vrijednosti člana u hrvatskome jeziku.

Eksplicitno iskazivanje vrijednosti člana posvojnom, neodređenom i pokaznom zamjenicom prikazano je u Primjeru 4.

1) „...posljednji puta izmijenjena i dopunjena Uredbom (EZ-a) br. 2072/98, a posebno njezin članak 8. točku b...“(jrc31999R0691)
„...modificato da ultimo dal regolamento (CE) n. 2072/98 (2), in particolare l’articolo 8, lettera b...“(jrc31999R0691)

2) „...budući da je zato poţeljno odrediti onaj koncept proizvoda ili obitelji...“(jrc31999D0089)
„...è opportuno definire il concetto di prodotto o di gruppo di prodotti...“(jrc31999D0089)

3) „...«proizvođač» je svaka osoba čijom aktivnošću nastaje otpad...“(jrc32006L0012)
„...«produttore»: la persona la cui attività ha prodotto rifiuti...“(jrc32006L0012)

Primjer 4: Primjeri iskazivanja vrijednosti člana u hrvatskome jeziku.

S obzirom na to da su članovi činili sveukupno 7,97% pojavnica u referentnome talijanskom korpusu odnosno 7,12% svih pojavnica u specijaliziranome korpusu i 5,26% u prijevodnome korpusu, a da članovi ne postoje u hrvatskome, odlučilo se radi usporedivosti omjera isključiti broj pojavnica s oznakom člana. Na taj način uspoređivali bi se omjeri samo postojećih vrsta riječi u oba jezika. Takva metodološka odluka, osim što nalazi opravdanje u prethodnim radovima usmjerenim na prevođenje hrvatsko-talijanskoga jezičnog para (Ljubičić, 2000; Katušić, 1981; 1982a; 1982b; 1982c), ali i onim usmjerenim na proučavanje određenosti i neodređenosti u hrvatskome jeziku (Karlić, 2014), potvrđena je i testom na usporednim tekstovima.

3.3 Normalizirana veličina korpusa

Nakon što su utvrđene i prikazane razlike među korpusima, koje su proizašle iz kontrastivne analize dvaju jezika na razini vrsta riječi i pokušaja usklađivanja oznaka korištenih u morfosintaktičkome označavanju hrvatskih i

[2] Riječ je o posebnoj vrsti usporednoga korpusa, sastavljenoj od prijevoda na dva jezika bez izvornika.

talijanskih korpusa, provjerio se njihov utjecaj na pouzdanost rezultata distribucije vrsta riječi svih šest korpusa.

Iz provjere proizlazi kako te razlike ne mijenjaju u velikoj mjeri sliku distribucije vrsta riječi pojedinoga korpusa. Međutim, radi postizanja veće usporedivosti odlučeno je promatrati distribuciju unutar zajedničkih dijelova korpusa na način da cjelinu čine samo one oznake koje su zajedničke i relevantne za ciljno istraţivanje. Relativne će se frekvencije stoga izračunavati u odnosu na normaliziranu veličinu korpusa, a podrazumijevat će veličinu korpusa umanjenu za apsolutnu frekvenciju oznaka koje nisu zajedničke ili koje nisu relevantne za ovo istraţivanje. Radi se o vrstama riječi koje ne postoje u oba jezika, poput člana i čestica ili o oznakama koje nisu relevantne, a nisu u jednakoj mjeri označene u korpusima poput kratica, pokrata, simbola i interpunkcijskih znakova. Postoji i vrsta riječi koja postoji i označena je u svim korpusima, no neće se uzimati u obzir u normaliziranim korpusima. Radi se o uzvicima koje se zbog više razloga odlučilo isključiti iz normaliziranih korpusa. Osim što čine jako mali dio korpusa, uzvici nisu uobičajeni u zakonodavnopravnome stilu, što je vidljivo iz njihove još manje uključenosti u specijaliziranim korpusima. Dodatni je razlog niski postotak točnosti u označavanju uzvika u specijaliziranim korpusima, neovisno o jeziku[3].

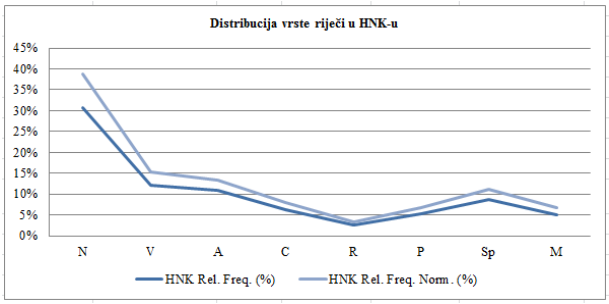
Normalizirani korpus uključivat će dakle samo apsolutnu frekvenciju sljedećih vrsta riječi odnosno oznaka: imenica (N), glagola (V), pridjeva (A), veznika (C), priloga (R), zamjenica (P), prijedloga (Sp) i brojeva (M). Tablica 5 prikazuje apsolutni broj pojavnica u izvornim oblicima korpusa, relativni broj pojavnica u normaliziranim korpusima i njihov omjer.

Korpus	Apsolutni broj pojavnica u izvornome korpusu	Apsolutni broj pojavnica u normaliz. korpusu	Omjer broja pojavnica u izvornome i normaliz. korpusu
HNK	216 812 148	177 216 713	0,8173
NN_{1990_2013}	92 363 788	70 852 385	0,7671
hrAcquis	11 209 795	8 784 240	0,7836
CORIS	130 294 347	98 300 670	0,7544
PRGAMM	9 575 784	7 325 921	0,7650
itAcquis	16 955 483	12 464 404	0,7351

Tablica 5: Apsolutni brojevi pojavnica u izvornim i u normaliziranim korpusima.

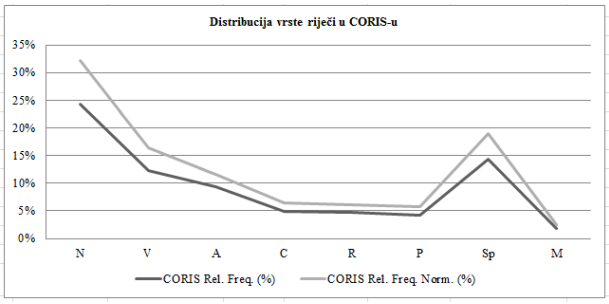
Iz Tablice 5 vidljivo je da su se korpusi smanjili za približno isti postotak. Da postotak nije u potpunosti isti i da se ne odraţava jednako na svaku vrstu riječi, prikazat će u nastavku Grafikon 1 i Grafikon 2.

Grafikon 1 prikazuje distribuciju relativne frekvencije vrste riječi u referentnome korpusu hrvatskoga jezika (HNKv3.0) u njegovu izvornome obliku i distribuciju vrsta riječi u normaliziranome korpusu.



Grafikon 1: Usporedba distribucije vrste riječi u HNK-u.

U Grafikonu 2 prikazana je distribucija relativne frekvencije vrste riječi u referentnome korpusu talijanskoga jezika (CORIS) u njegovu izvornome obliku i distribuciju vrsta riječi u normaliziranome korpusu.



Grafikon 2: Usporedba distribucije vrste riječi u CORIS-u.

Kao primjer razlika u distribuciji vrsta riječi između izvornih oblika korpusa i normaliziranih korpusa prikazana je distribucija frekvencija vrsta riječi referentnih korpusa. Iako je ta razlika kod referentnih korpusa najmanja, vidljivo je, kao što to prikazuju i Grafikon 1 i Grafikon 2, da se svođenje korpusa na normaliziranu veličinu ne odraţava jednako na svaku vrstu riječi te da je odluka o korištenju normaliziranih korpusa opravdana i da će doprinijeti većoj pouzdanosti rezultata.

4 Zaključak

Iz svega prikazanog u ovome radu očita je vaţnost i sustavno planiranje izrade skupa oznaka za vrstu riječi i ostale gramatičke kategorije za svaki pojedini jezik u skladu sa zajedničkim međunarodnim smjernicama koje propisuju standarde i stvaraju preduvjete usporedivosti među korpusima kako na unutarjezičnoj tako i na međujezičnoj razini.

Dok takav skup oznaka za hrvatski jezik postoji i dosljedno se koristi, puno je problematičnije stanje sa skupovima oznaka za talijanski jezik budući da postoje različiti skupovi oznaka. Ovaj rad jasno pokazuje kako neusklađenost oznaka, tj. postojanje različitih skupova POS i MSD oznaka za isti jezik, onemogućuje unutarjezičnu analizu korpusa označenih različitim skupovima oznaka, što u konačnici ograničava njihovu primjenu. Istovjetna primjedba vrijedi i na međujezičnoj razini i u toliko je bitno da se korpusni lingvisti pridrţavaju zajedničkih međunarodnih smjernica jer se time omogućuje jednostavnija usporedivost rezultata pretrage korpusa na različitim jezicima.

[3] Pojavnice koje su označene kao uzvici u specijaliziranim, kako hrvatskim tako i talijanskim korpusima, uglavnom nisu uzvici, već se radi o kraticama, pokratama ili vrlo često stranim riječima, što ukazuje na problematičnost označavanja uzvika neovisno o jeziku i o korištenome automatskom označivaču.

No, i kada bi postojala potpuna usklađenost sa smjernicama, neizbjeţno je promišljanje i usklađivanje oznaka s obzirom na razlike u poimanju i postojanju gramatičkih kategorija u pojedinim jezicima. Moglo bi se zaključiti dakle da se usporedba i usklađivanje MSD ili POS oznaka mogu smatrati dobrim temeljem i zanimljivim pristupom u kontrastivnoj analizi dvaju jezika.

S druge strane, treba spomenuti da već neko vrijeme postoje pokušaji sastavljanja univerzalnih skupova oznaka za vrste riječi i druge gramatičke kategorije kao primjerice univerzalni skup oznaka za vrste riječi prikazan u Petrov i suradnici (2012: 2089), koji uključuje 12 vrsta riječi određenih na temelju analize skupova oznaka 22 jezika, među kojima se nalazi i talijanski, ali ne i hrvatski jezik ili poput onoga sastavljenog u sklopu projekta *The Universal Dependencies*[4], a koji uključuje i hrvatski jezik (Agić et al., 2015). Moţe se stoga i zaključiti da ako se budući sastavljači korpusa budu pridrţavali međunarodnih smjernica i univerzalnih skupova oznaka za vrste riječi i druge gramatičke kategorije, usklađivanja poput ovoga prikazanog u ovome radu u budućnosti moţda više neće biti potrebna.

5 Bibliografija

Ţeljko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richard Farkas, Jennifer Foster, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Jan Hajič, Anders Trærup Johannsen, Jenna Kanerva, Juha Kuokkala, Veronika Laippala, Alessandro Lenci, Krister Linden, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Hector Alonso Martınez, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Joakim Nivre, Hanna Nurmi, Petya Osenova, Slav Petrov, Jussi Piitulainen, Barbara Plank, Prokopis Prokopidis, Sampo Pyysalo, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Kiril Simov, Aaron Smith, Reut Tsarfaty, Veronika Vincze i Daniel Zeman. 2015. *Universal dependencies 1.1*.

Eugenija Barić, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević i Marija Znika. 1997. *Hrvatska gramatika*. Školska knjiga, Zagreb

Raffaella Bernardi, Andrea Bolognesi, Corrado Seidenari i Fabio Tamburini. 2006. POS tagset design for Italian. U *Proceedings 5th International Conference on Language Resources and Evaluation – (LREC 2006)*, str. 1396–1401. European Language Resources Association (ELRA), Genova.

Raffaella Bernardi, Andrea Bolognesi, Corrado Seidenari i Fabio Tamburini. 2005. Automatic induction of a POS tagset for Italian. U *Proceedings Australasian Language Technology Workshop -ALTW 2005*, Sydney.

Douglas Biber. 1995. *Dimensions of register variation: a cross-linguistic comparison.* Cambridge University Press, Cambridge.

Maurizio Dardano i Pietro Trifone. 2003. *La lingua italiana: morfologia, sintassi, fonologia, formazione delle parole, lessico, nozioni di linguistica e sociolinguistica* (7. izd.). Zanichelli, Bologna.

Tomaţ Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. U *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, LREC'04, str. 1535–1538. European Language Resources Association (ELRA), Paris.

Federico Faloppa. 2011. Numerali. U R. Simone, ur., Enciclopedia dell'italiano, str. 972–974. Istituto dell'Enciclopedia Italiana, Roma.

Josip Jernej. 1990. Riflessioni sulle unità linguistiche chiamate „particelle". *Italica Belgradensia*, 3: 1–4.

Josip Jernej. 2005. *Konverzacijska talijanska gramatika za početnike i napredn*e (11. izd.). Školska knjiga, Zagreb.

Virna Karlić. 2014. *Određenost i neodređenost u srpskom i hrvatskom jeziku*. Neobjavljena doktorska disertacija, Filozofski fakultet Sveučilišta u Zagrebu.

Maslina Katušić. 1981. Note preliminari sulla traduzione dell'articolo italiano. *Studia Romanica et Anglica Zagrabiensia*, XXVI (1-2): 149-158.

Maslina Katušić. 1982a. Neka razmišljanja o mogućnosti prevođenja □određenog□ i neodređenog□članа. *Strani jezici*, XI (1-2): 17–26.

Maslina Katušić. 1982b. L'articolo italiano: un problema di traduzione (I). *Studia Romanica et Anglica Zagrabiensia*, XXVII (1-2): 145–196.

Maslina Katušić. 1982c. L'articolo italiano: un problema di traduzione (II). *Studia Romanica et Anglica Zagrabiensia*, XXVIII (1-2): 111–166.

Ivana Lalli Paćelat. 2014. *Analiza zakonodavnopravnoga stila hrvatskoga i talijanskoga jezika: unutarjezična, međujezična i prijevodna perspektiva*. Neobjavljena doktorska disertacija, Filozofski fakultet Sveučilišta u Zagrebu.

Maslina Ljubičić. 2000. *Studije o prevođenju*. Hval, Zagreb.

Maria Giuseppa Lo Duca. 2011. Parti del discorso. U R. Simone, ur., *Enciclopedia dell'italiano. Istituto dell'Enciclopedia Italiana*, Roma. Dostupno na: http://www.treccani.it/enciclopedia/parti-del-discorso_(Enciclopedia_dell'Italiano)/

Monica Monachini. 1995. *ELM-IT: An Italian Incarnation of the EAGLES-TS. Definition of Lexicon Specification and Classification Guidelines.* Technical report, Pisa.

Stella Neumann. 2013. *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German.* De Gruyter Mouton, Boston/Berlin.

Slav Petrov, Dipanjan Das i Ryan McDonald. 2012. A universal part-of-speech tagset. U Proceedings of the 8th International Conference on Language Resources and Evaluation – (LREC 2012), str. 2089–2096. http://arxiv.org/pdf/1104.2086v1.pdf

Prokopis Prokopidis, Vassilis Papavassiliou, Antonio Toral, Marc Poch, Francesca Frontini, Francesco Rubino i Gregor Thurmair. 2012. WP-4.5: Report on the revised Corpus Acquisition & Annotation subsystem and its components. Panacea Project. http://hdl.handle.net/10230/22514

Rema Rossini Favretti. 2000. Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS. U R. Rossini Favretti, ur., *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*, str. 39–56. Bulzoni, Roma.

[4]Više o projektu na http://universaldependencies.org/hr/pos/index.html [posjećeno 5. rujna 2016.].

Rema Rossini Favretti, Fabio Tamburini i Cristiana De Santis. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. U A. Wilson, P. Rayson, i T. McEnery, ur., *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, str. 27–38. Lincom-Europa, Munich.

Josip Silić i Ivo Pranjković. 2005. *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*. Školska knjiga, Zagreb.

Lelija Sočanac. 2004. *rva tsko-talijanski jezični dodiri: s rječnikom talijanizama u standardnome hrvatskom jeziku i du rovačkoj dramskoj knji evnosti.* Nakladni zavod Globus, Zagreb.

Marko Tadić. 1996. Računalna obrada hrvatskoga i nacionalni korpus. *Suvremena lingvistika*, 22(41–42): 603–612.

Marko Tadić. 1998. Raspon, opseg i sastav korpusa suvremenoga hrvatskoga jezika. *Filologija*, 30–31: 337–347.

Marko Tadić. 2003. *Jezične tehnologije za hrvatski jezik.* Exlibris, Zagreb.

Marko Tadić. 2009. New version of the Croatian National Corpus. U D. Hlaváčková, A. Horák, K. Osolsobě i P. Rychlý, ur., *After Half a Century of Slavonic Natural Language Processing*, str. 199–205. Masaryk University, Brno.

Branka Tafra. 1989. Što su brojevi? (gramatički i leksikografski problem). *Rasprave Zavoda za jezik IFF,* 15: 219–237.

Branka Tafra. 2000. Morfološka obilje ja brojevnih riječi. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 26: 261–275.

Branka Tafra. 2005. *Od riječi do rječnika*. Školska knjiga, Zagreb.

Fabio Tamburini, Corrado Seidenari, Andrea Bolognesi i Raffaella Bernardi. 2008. Italian Lexical-Classes Definition Using Automatic Methods. U R. Rossini Favretti, ur., *Frames, Corpora and Knowledge Representation*, str. 95–120. Bononia University Press, Bologna.

Fabio Tamburini. 2000. Annotazione grammaticale e lemmatizzazione di corpora in italiano. U R. Rossini Favretti, ur., *Linguistica e informatica: multimedialità, corpora e percorsi di apprendimento*, str. 57–73. Bulzoni, Roma.

Fabio Tamburini, Corrado Seidenari, Andrea Bolognesi i Raffaella Bernardi. 2008. Italian Lexical-Classes Definition Using Automatic Methods. U Rema Rossini Favretti, ur., *Frames, Corpora and Knowledge Representation*, str. 95–120. Bononia University Press, Bologna.

Elke Teich. 2003. *Cross-linguistic variation in system and text.* Mouton de Gruyter, Berlin/New York.

Pavao Tekavčić. 1989. Prema kontrastivnoj gramatici tzv. „čestica“ u hrvatskom ili srpskom jeziku i talijanskom jeziku. *Rad JAZU*, 427: 127–194.

Stjepko Te ak i Stjepan Babić. 2005. *ra matika hrvatskoga jezika: priručnik za osnovno jezično obrazovanje* (15. izd.). Školska knjiga, Zagreb.

Giulia Venturi. 2009. Rassegna comparativa degli schemi di annotazione morfosintattica per la lingua italiana, *Technical report TRIPLE - RTT/1*.

Richard Xiao. 2010. How different is translated Chinese from native Chinese. *International Journal of Corpus Linguistics*, 15(1): 5–35.

Easily Accessible Language Technologies for Slovene, Croatian and Serbian

Nikola Ljubešić,[†‡] Tomaž Erjavec,[†] Darja Fišer,[*†] Tanja Samardžić,[♣] Maja Miličević,[♠] Filip Klubička,[‡] Filip Petkovski[♢]

[†]Department of Knowledge Technologies, Jožef Stefan Institute
tomaz.erjavec@ijs.si

[‡]Faculty of Humanities and Social Sciences
nljubesi@ffzg.hr, fklubicka@ffzg.hr

[*]Faculty of Arts, University of Ljubljana
darja.fiser@ff.uni-lj.si

[♣]CorpusLab, University of Zürich
tanja.samardzic@uzh.ch

[♠]Faculty of Philology, University of Belgrade
m.milicevic@fil.bg.ac.rs

[♢]Freelance developer, Macedonia
filip.petkovsky@gmail.com

Abstract

In this paper we present the pipeline of recently developed language technology tools for Slovene, Croatian and Serbian. They currently cover text segmentation, text normalisation, part-of-speech tagging, lemmatisation and inflectional lexicon lookup. Most rely on machine learning approaches, such as statistical machine translation and conditional random fields, capable of producing high-quality models for the phenomenon covered. Special emphasis is put on easy accessibility of these tools by offering them and the trained models for all three languages as (1) open source via public git repositories and (2) online in the form of web applications and web services.

1. Introduction

With the increasing availability of language technologies for various languages, different scientific areas, including those of social sciences and humanities (SSH), have started to perceive the usefulness of such technologies for their own research. Given the lower level of technical competence of most researchers in SSH in comparison to the areas language technologies are developed in, a significant technological gap has to be filled, which would enable SSH scholars to include the developed technologies in their own research.

This paper presents a joint effort to make language technology for three western South Slavic languages – Slovene, Croatian and Serbian – more widely accessible. For Slovene there are already tools available for tagging and lemmatisation in form of web applications, such as ToTaLe[1] (Erjavec et al., 2005) and Obeliks[2] (Grčar et al., 2012), but of lower quality than the one presented in this paper. For Croatian there was a web application available hosting tools trained on the SETimes.HR corpus (Agić and Ljubešić, 2014), but given the superior quality of the tools presented in this paper, this web application is currently forwarding requests to the new solution. For Serbian there were no technologies available up to this point.

While many toolchains already exist, e.g. Gate (Cunningham et al., 2011), FreeLing (Padró and Stanilovsky, 2012), OpenNLP (Apache Software Foundation, 2014), there are two main reasons why they do not suit our needs. First, the choice of technology in existing toolchains is mostly oriented toward the major world languages. Subsequently, for part-of-speech tagging HMMs are used which, in case of more complex inflectional languages such as the Slavic ones, do not yield the best results. The other reason is that most of the toolchains only cover basic tasks like part-of-speech tagging, parsing and named entity recognition while our toolchain has already touched on more specific tasks like non-standard language normalisation.

Furthermore, we put special emphasis on bridging the aforementioned technology gap by offering three modes of using the developed technologies: (1) as open source programs available from the public GitHub repository, (2) as RESTful web services and (3) through a web application. The first is intended for technically experienced people who are capable of installing the tools and their dependencies and want to process large amounts of data, as well as control input and output formats. The latter two are better suited for those who are either processing smaller datasets or do not have the knowledge or hardware capabilities to install and run the tools locally. The web services can be used from code either directly as JSON-based RESTful services or through an available Python library. The developed web application is primarily intended for teaching purposes, trying out the technologies, debugging or processing only a handful of documents.

The tools are, for the most part, based on the machine

[1] http://nl.ijs.si/tei/convert/

[2] http://eng.slovenscina.eu/tehnologije/oznacevalnik

learning paradigm, and comprise the learning and execution components as well as models for Slovene, Croatian and Serbian developed by training the tool on the best resources available for the task.

The paper is structured as follows: the following section gives a short overview of the developed technologies, Section 3 describes the available modes of using them, while the last section gives a short-term plan of future developments.

2. Language Technology Tools

2.1. Inflectional Lexicons

Slavic languages in general have a complex inflectional system and lexicons covering this layer of language are important for almost any task of automatic language processing. All three languages of interest now have available large inflectional lexicons, in particular:

- the Slovene Sloleks lexicon (Dobrovoljc et al., 2015), 100,805 lexemes in size;
- the Croatian hrLex lexicon (Ljubešić, 2016a), 99,680 lexemes in size;
- the Serbian srLex lexicon (Ljubešić, 2016b), 105,358 lexemes in size.

The entry in each lexicon consists of the lemma and its complete inflectional paradigm comprising the word forms, their morphosyntactic descriptions and their corpus frequencies.

Through our web services and application we give a unified interface to all three resources.

2.2. Diacritic Restoration Tool

In computer-mediated communication, such as emails, instant messages, tweets etc. users of Latin-based scripts often replace characters with diacritics with their ASCII equivalents for ergonomic reasons, especially when typing on tablets and smartphones. Such text is typically easily understandable to humans but very difficult for computational processing because many words without the diacritics become ambiguous or unknown. At the same time, computer-mediated communication has become a hot topic of research and application, which is why high-quality processing of such language is in high demand.

We have developed a diacritic restoration tool called REDI (Ljubešić et al., 2016a)[3] with models covering all three languages of interest. The tool is trained on large corpora and consists of two components: the translation model (the probability of a standard word given its dediacritised version) and the language model (the probability of the standard word given its context). For estimating the token translation probability we use the maximum likelihood estimate of a diacritised form given the dediacritised one, while for estimating the context probability we use KenLM (Heafield, 2011) with the default parameters. These two components are combined with a log-linear model.

The token-level accuracy of the tool is around 99.5% on standard text and around 99.2% on non-standard text (Ljubešić et al., 2016a). The tool significantly outperforms charlifter,[4] so far the only open source tool available for this task on the target languages, which achieves around 97% accuracy on standard and around 94% on non-standard text.

2.3. Non-Standard Text Normalisation

Computer-mediated communication is often written in non-standard language, where users are either not acquainted with the language norm or, more often, intentionally use phonetic and dialectal spelling. Similarly, historical texts are also written in language which is significantly different from the contemporary standard. However, annotation tools, such as PoS taggers and lemmatisers, are typically trained on standard language and perform poorly on non-standard texts. As developing new text tools for each language variety is very time consuming and expensive, a typical approach is to first standardise the spelling of words and only then apply further processing on them.

For normalising words in user-generated content we use character-level statistical machine translation (CSMT), the Slovene variant of which, applied both to computer-mediated communication and historical texts, is presented in Ljubešić et al. (2016). The technology is based on the well-known SMT system Moses (Koehn et al., 2007), which is trained on a manually normalised collection of tweets split into sequences of characters. For all three languages the training set comprises 80,000 tokens.

The last experiments on Slovene show that for less standard tweets the error reduction obtained when applying CSMT is $\sim$70% while for more standard tweets it is $\sim$50%. When comparing the CSMT systems to a baseline which applies the most probable token transformation as estimated on the same training data, the error reduction on less standard tweets is $\sim$35% and on more standard tweets $\sim$45% (Ljubešić et al., 2016).

2.4. Morphosyntactic Annotation and Lemmatisation

For Slavic languages morphosyntactic tagging is probably the most important step in text annotation, and is still an interesting topic of research. Such languages with their large tagsets of morphosyntactic descriptions (MSDs) and often limited training data still offer significant room for improvement in tagging accuracy. Similar points hold for lemmatisation, the process of assigning the base form to a word form in running text. On one hand, the rules for predicting the lemma of a word form are complex and have many exceptions, while, on the other, the word forms are often ambiguous and their MSD tag is needed to correctly determine the lemma.

We recently developed a new tagger combined with a lemmatiser, explicitly developed for high-quality processing of the languages of interest (Ljubešić and Erjavec, 2016; Ljubešić et al., 2016b).[5] The tagger follows the approach by Grčar et al. (2012) but replacing their instance classifier (SVM) with a sequential one (CRF) and re-engineering the optimal features given the different nature

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

[4] https://sourceforge.net/projects/lingala/files/charlifter/

[5] https://github.com/clarinsi/reldi-tagger

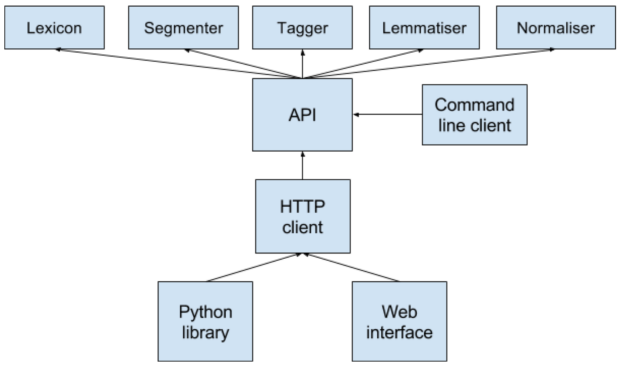


Figure 1: Architecture of the system exposing the developed language technologies.

of the classifier. With this we significantly improve their results, with an error reduction of ∼25% on both known and unknown words.

The Slovene model for our tagger is trained on the ssj500k corpus (Krek et al., 2015) and the Sloleks lexicon with the reported tagging accuracy at 94.27%. The Croatian and Serbian models are trained on the Croatian hr500k corpus (Ljubešić et al., 2016b), with the Croatian model using the hrLex lexicon (Ljubešić, 2016a) and the Serbian one the srLex lexicon (Ljubešić, 2016b). The reported tagging accuracy for Croatian is 92.53% and for Serbian 92.33%.

The currently used lemmatiser applies the mentioned lexicons in case the surface form and guessed MSD can be found in them. Otherwise, it uses supervised machine learning to predict the transformation of the surface form to the lemma. The predicted transformation is formalised as a 4-tuple (prefix length, prefix substitute, suffix length, suffix substitute); for example, in the transformation from *najvećih* to *velik* the 4-tuple is `(3,"",3,"lik")`. The features used for prediction are the suffixes of different length and the guessed MSD.

3. Accessibility

3.1. Open Source

Most of the tools described above are already available as open source distributions inside the CLARIN.SI GitHub organisation.[6] Git has become the most popular platform for (distributed) code development, and GitHub additionally offers a free platform for sharing the code, reporting bugs and requests for improvements, monitoring the activity of a project etc. It, of course, also offers the possibility for third party developers to post improvements to the code with a well-defined procedure for incorporating them into the master branch. For all technologies, we plan in the near future to deposit stable versions of the code to the repository of the Slovene research infrastructure CLARIN.SI,[7] as this frees us from the dependence on a U.S. based repository, and, more importantly, gives additional visibility and citability to the code. Namely, CLARIN.SI has an OAI-PMH endpoint, which enables it to expose the repository metadata to harvesting services, with the repository already being harvested by the European CLARIN Virtual Language Observatory. Furthermore, CLARIN.SI uses the Handle system for persistent identifiers and recommends the correct way of citing its items in publications, thus giving a better chance of acquiring citations for the tools in scientific publications.

3.2. Web application and services

The envisaged architecture of our system that joins the developed language technologies in one ecosystem is presented in Figure 1. There are three approaches to access our technologies: via a command line client, a web interface and a Python library. The latter two approaches access the technologies through the HTTP protocol and have no local requirements besides a browser and a Python interpreter. The command line client is planned for researchers who want to install all the technologies locally as a single package and this component of our system will be finished once all the intended technologies are added to the system. In the remainder of this subsection we describe the two HTTP-based access methods.

Access to both the web interface and to the API via a Python library requires authentication, in order to ensure the stability of the service. To obtain a user name and password one has to register at `http://nl.ijs.si/services/`. This URL is also the entry point to the web application.

3.2.1. Web Application

The technologies currently available through the web application are the lexicon, segmenter, tagger and lemmatiser. A screenshot of the interface to the tagger and lemmatiser is given in Figure 2. The interface enables either writing / pasting text into the form or uploading a text file, choosing the language, defining the input format (either plain text or the text corpus format TCF[8] and choosing the function one wants to run on the input data. Currently the available functions are "Tag", "Lemmatise" and "Tag + Lemmatise". Each of the functions also contains the preprocessing step of segmenting the input on sentences and tokens. For future versions of the technology a higher level of control is planned by allowing building custom pipelines for tasks like tagging already tokenised text, both normalising and tagging text etc.

The result of applying the function on the input data is presented on the right side of the screen in three different modes: as a table, as raw response from the web service and for download. The downloaded file contains either vertical text with tab-separated annotations if the input format was text, or a TCF file if such input format was given.

The main purposes of the presented web application are the following: (1) a first insight in the quality of the output of the language technologies, (2) an insight in the raw response given from the API and (3) a way to process smaller amounts of data, mostly present in form of text files. Uploading, processing and downloading a text file via this web

[6] `https://github.com/clarinsi`
[7] `http://www.clarin.si/`
[8] TCF if an XML-based format used by WebLicht `http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format`

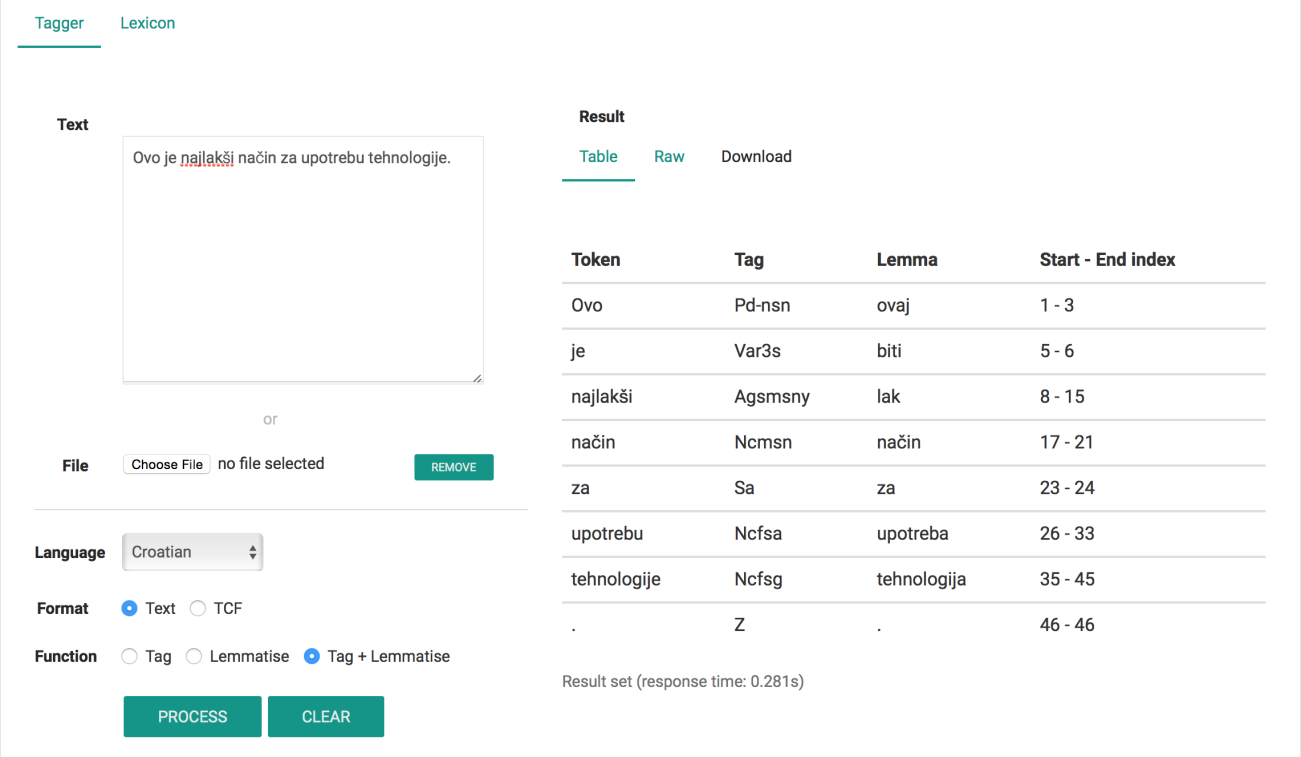


Figure 2: Screenshot of the web interface.

application should not take more than a minute of a user's time.

3.2.2. Web Services

The easiest way to use the language technologies from code is via the Python library which is available via PyPI[9] while the documentation on using the library is available from GitHub.[10] Currently all the developed technologies besides the CSMT text normaliser are available through this library. Here is a code snippet example of using the Python library for segmentation, tagging and lemmatisation:

```
from reldi.tagger import Tagger
from getpass import getpass
username="my_username"
passwd=getpass("Input password: ")
tagger = Tagger("hr")
tagger.authorize(username, passwd)
result=tagger.tagLemmatise("Obradi me.")
```

4. Future Developments

While a lot of work was already put into developing the presented technologies and ensuring their accessibility through a unified ecosystem, a lot is still to be done. Here we present the order of our planned activities.

We plan to develop additional annotation tools, namely a dependency parser and a named entity recogniser. While Slovene and Croatian are part of the Universal Dependencies (UD) project[11] (Nivre et al., 2016), we are working on adding Serbian to its repository by annotating the Serbian dataset corresponding to the Croatian SETimes.HR corpus (Agić and Ljubešić, 2014).

For named entity recognition we have a series of datasets already available and plan on expanding them and develop a CRF-based named entity recogniser.

Once a tool is developed, the procedure we follow is the following: (1) releasing it as open-source via GitHub, (2) including it in our API, (3) ensuring access to the API component through our Python library and (4) making the tool accessible via the web application, which is the final step of exposing a technology as it requires most work, the majority of which is related to the development of the user interface. We are currently working on including the CSMT normalisers of all three languages into the API and training models for UD parsers for Slovene and Croatian.

Once all the technologies have gone through the process of development, inclusion in the API, the Python library and the web application, we will deploy our whole ecosystem as a single package, enabling researchers with large data processing needs to seamlessly install and use the technologies on their own servers.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran), the Slovenian Research Agency within the national basic research project "Resources, Tools and Methods for the Research of Nonstandard Internet Slovene" (J6-6842, 2014-2017) and the Swiss National Science Foundation grant no. IZ74Z0_160501 (ReLDI).

[9] https://pypi.python.org/pypi/reldi
[10] https://github.com/clarinsi/reldi-lib
[11] http://universaldependencies.org

5. 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).

Apache Software Foundation. 2014. openNLP Natural Language Processing Library. http://opennlp.apache.org/.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (Version 6)*.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, and Miro Romih. 2015. *Morphological lexicon Sloleks 1.2*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1039.

Tomaž Erjavec, Camelia Ignat, Bruno Poliquen, and Ralf Steinberger. 2005. Massive multilingual corpus compilation: Acquis communautaire and totale. In *The 2nd Language & Technology Conference - Human Language Technologies as a Challenge for Computer Science and Linguistics*. Association for Computing Machinery (ACM) and UAM Fundacja.

Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In *Proceedings of the Eight Conference on Language Technologies*, Ljubljana, Slovenia.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *In Proc. of the Sixth Workshop on Statistical Machine Translation*.

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*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, and Nanika Holz. 2015. Training corpus ssj500k 1.4. Slovenian language resource repository CLARIN.SI.

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).

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2016a. Corpus-Based Diacritic Restoration for South Slavic Languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA).

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016b. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA).

Nikola Ljubešić. 2016a. *Inflectional lexicon hrLex 1.1*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/1135/1067`.

Nikola Ljubešić. 2016b. *Inflectional lexicon srLex 1.1*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1066`.

Nikola Ljubešić, Katja Zupan, Darja Fišer, and Tomaž Erjavec. 2016. Normalising Slovene data: historical texts vs. user-generated content. In *Proceedings of KONVENS 2016*.

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

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Strojno podprta kolacija slovenskih rokopisnih besedil: variantna mesta v luči računalniških algoritmov in vizualizacij

Matija Ogrin, Andrejka Žejn

Inštitut za slovensko literaturo in literarne vede, ZRC SAZU
Novi trg 2, 1000 Ljubljana
matija.ogrin@zrc-sazu.si
andrejka.zejn@zrc-sazu.si

Povzetek

Prispevek predstavi problematiko dveh verzij baročnega rokopisnega besedila *Jezusovo življenje v sto postavah*, kjer je mlajši, *Poljanski rokopis,* nastal kot prepis starejšega. Tekstnokritična raziskava je temeljila na metodi kolacije – primerjave obeh verzij in analize variantnih mest. Za računalniško podporo postopku kolacije smo uporabili programa Juxta in CollateX, ki sta pokazala kljub razmeroma nizki stopnji (indeks 0.35) variantne oddaljenosti prepisa od prvotnega rokopisa v obsegu treh poglavij več tisoč variantnih mest. Variante smo v analizi razdelili na tri skupine: 1. variante pri zapisovanju fonemov; 2. variante na ravni jezika: v glasovju, oblikah besed in besedju in 3. variante na besedilni ravni. Tretja skupina zaobsega številna mesta, kjer je pisec starejše baročno besedilo strnjeval in krajšal, da bi tako dosegel bolj pregledno in preprosto besedilo, četudi občasno na škodo retoričnih figur in temperamentnega sloga, ki daje temu meditativnemu besedilu značilno literarno kvaliteto. Vsekakor je strojno podprta kolacija kot metoda tekstološke analize prinesla obsežno evidenco besedilne variantnosti v tem razmeroma kratkem baročnem tekstu in s tem odprla pot za nadaljnje raziskave njegovega jezika in literarnega sloga.

Computer-aided Collation of Slovenian Manuscript Texts: Variant Readings as an Object of Algorithms and Visualization

Very few Slovenian manuscript texts from 18th century have been preserved in more than one witness. The comprehensive *Poljane manuscript*, which contains a Baroque meditative text on the life and passion of Jesus Christ, proved to be a witness to an older manuscript protograph, of which only 80 pages are preserved. This extent, however, is sufficient for a study in nature and types of differences between the two manuscript texts. The text-critical approach based on the collation of variant readings. To this scope, we used two collation programs: Juxta and CollateX, which detected several thousands of variants between the two versions, even though the general variation index was low (only 0.35). The analysis detected three groups of variant readings: 1. those related to variant writing of phonemes, 2. language variations in phonemes, word forms and lexicon and 3. variants on the textual level. The third group comprises many passages where the writer has slightly reduced and shortened the old text in order to make it more clear and simple. In several variant readings, this adaptation was not in favor of the rhetorical devices which give this meditative text a distinctive literary quality. At any rate, the computer aided collation has brought us a great amount of evidence on textual variance, which is now open for further research in language and literary style.

1 Preoddaja rokopisnih besedil

Za preučevanje besedilne tradicije rokopisnih tekstov je bistveni pogoj to, ali je določeno historično besedilo sploh ohranjeno v več kot enem rokopisu. Le če sta ohranjena vsaj dva rokopisa »istega« besedila in imamo torej dve verziji, je mogoča med njima primerjava – in le z njo se jasno pokažejo spremembe, ki nastajajo pri posredovanju ali *preoddaji* besedil iz enega rokopisa v drugega in naslednjega. Bolj ko se veča distanca med izvorom besedila in njegovim končnim bralcem in več ko je vmesnih stopenj v njegovem posredovanju, večja je možnost, da se besedilo pri tem spremeni (Inglese 2006: 13). Dejansko vsakokrat, ko je besedilo preoddano (tradirano), nastanejo v njem spremembe (Jäger 1998: 35), ki odsevajo proces in historične dejavnike same preoddaje. Ti dejavniki segajo od nehotenih napak do jezikovnih vplivov, denimo vplivov narečja na knjižni jezik, pogosto pa so tudi individualni, povezani z osebno kulturo pisca, njegovo literarno občutljivostjo, ali jih pogojuje kultura družbe ipd.

Glavni problem preoddaje ali tradicije novoveških besedil v slovenskem jeziku je v tem, da jih je iz dobe pred 19. stoletjem izjemno malo ohranjenih v več kot enem rokopisu. Zato je le v malo primerih mogoče besedila primerjati v striktnem tekstnokritičnem pomenu besede in iz primerjave sklepati glede historičnih okoliščin tradicije. Pogostejši so šele primeri iz 19. stoletja (in to je pravzaprav drug pojav), ko za določenim literarnim avtorjem ostane več verzij istega besedila – denimo pri Prešernovi *Zdravici*.

Zato je za raziskave slovenskega slovstva, jezika in kulture dragoceno, da sta se iz 18. stoletja ohranila dva rokopisa istega teksta, pri katerih je moč celo dokazati, da je mlajši nastal kot prepis starejšega. V *Registru slovenskih rokopisov 17. in 18. stoletja* (NRSS) je starejši opisan pod siglo Ms 028, mlajši pa pod Ms 023. Gre za baročno asketično-meditativno besedilo o Jezusovem življenju v sto poglavjih ali *postavah*, prevedeno oziroma predelano po nemških baročnih besedilih kapucinskega patra Martina Cochemskega. Mlajši rokopis (Ms 023) iz časa tik pred ali okrog 1800 je znan kot *Poljanski rokopis;* izjemno obsežen rokopis šteje čez 700 strani v velikosti folianta. Žal njegova predloga *Jezusovo življenje v sto postavah* (Ms 028) ni ohranjena v celoti, pač pa se je ohranilo le 40 folijev ali 80 strani rokopisa in s tem le slaba desetina starega besedila. Tudi v mlajšem, *Poljanskem rokopisu* so v tem delu, ki je na voljo za primerjavo, izgubljeni trije foliji rokopisa, kar še nekoliko zmanjša primerljivi vzporedni korpus besedila. Toda žalostni izgubi navkljub tudi ta obseg teksta zadošča za primerjavo, ki razkrije značilne tipe sprememb in s tem povezane dejavnike, ki so vplivali na besedilno tradicijo tega baročnega besedila.

2 Kolacija kot metoda, postopki in orodja

Primerjava in analiza dveh verzij besedila, ki tu nastopa kot raziskovalna metoda, obsega v procesu tekstnokritične obravnave besedila dve pomembni fazi. To sta kolacija (prim. Inglese 2006: 65–67) in recenzija (Jäger 1998: 35), ki sodita med temeljne postopke filološke oziroma tekstnokritične analize besedila.

V kolaciji obe verziji besedila vzporedno primerjamo od besede do besede, pravzaprav od črke do črke. S tem ugotovimo brezštevilne razlike v načinu zapisa glasov in slovničnih struktur, pa tudi besedilne spremembe, kakor zamenjave besednega reda, slogovne modifikacije, izpuste večjih ali manjših odlomkov, dodajanje besedila, in še marsikaj.

Recenzija pa obstoji v analizi ter dokumentiranju posameznih razlik in njihovih tipov, na osnovi česar moremo sklepati o razmerju med verzijama in o dejavnikih, ki so besedilo pri preoddaji preoblikovali. Recenzija temelji na predpostavki, da je verzije besedila, v našem primeru dve, moč razporediti v genealoškem razmerju, kjer se pokaže, kako poznejša izvira iz starejše (Edwards 1995: 190).

Za kolacijo in recenzijo, ki sta notranje smiselno povezani opravili, je bilo že ob samih začetkih digitalne humanistike pripravljenih več programov ali orodij. Do danes sta daleč najbolj razviti orodji Juxta in CollateX.

2.1 Analiza variantnosti z orodji Juxta in CollateX

Oba programa sta nastala v ožji povezavi z raznimi filološkimi projekti znanstvenokritičnih izdaj: Juxta v okviru raziskav ameriške književnosti 19. stoletja NINES, CollateX pa v okviru evropskega edicijskega konzorcija Interedition. Oba programa sta tudi prosto dostopna, saj pri njunem razvoju sodeluje večja skupina ali kar skupnost razvijalcev in uporabnikov.

Juxta lahko deluje kot samostojen lokalno nameščen program, nekoliko bolj zmogljiva inačica pa deluje kot spletni servis Juxta Commons.[1]

CollateX prav tako lahko deluje kot samostojen program, dostopna pa je tudi inačica za vgraditev v širše javansko programsko okolje.[2]

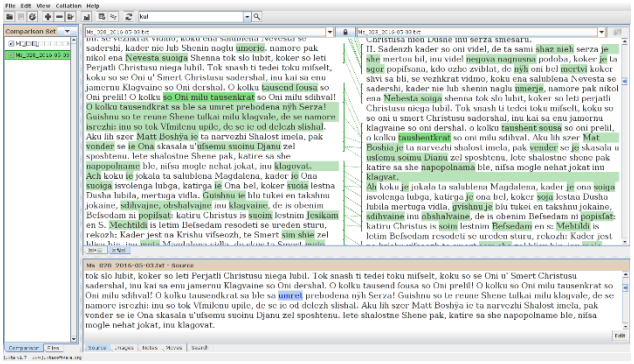
Razvijalci Juxte in CollateX so se zedinili za skupno temeljno koncepcijo elektronske obdelave variantnosti besedil. Celoten proces so razdelili na štiri faze:

1. tokenizacija, tj. segmentacija besedila vseh primerjanih verzij na pojavnice, ločila idr.;
2. vzporejanje členov, tj. izračun, katere pojavnice ali segmenti več pojavnic iz ene verzije ustrezajo (ali ne ustrezajo) pojavnicam v drugih verzijah teksta;
3. analiza doda izračunani vzporeditvi pojavnic interpretativno dimenzijo: za določen segment lahko ugotovi, da v drugi verziji ni bil (samo) spremenjen, ampak (tudi) prestavljen (transponiran) na drugo mesto v tekstu, ali pa izbrisan ali dodan;
4. vizualizacija rezultatov vzporejanja in analize je nujna končna stopnja vsake analize variantnosti – pa tudi z njo povezani izhodni podatki v raznih formatih za nadaljnjo obdelavo, zlasti v obliki kritičnega aparata.

Ta štiristopenjska zasnova je pozneje dobila ime »gothenburški model« procesiranja besedilne variantnosti (prim. Dekker in Middell, 2013).

Vhodno besedilo verzij, ki jih želimo primerjati, lahko za oba programa pripravimo bodisi v goli besedilni obliki (.txt) ali v zapisu XML-TEI. Glavna konceptualna razlika med programoma – poleg raznih manjših razlik v funkcionalnostih – je v algoritmih za vzporejanje (stopnja 2) in programju za vizualizacijo (stopnja 4).

Juxta je razvila dobro *lastno* vizualizacijo variantnih mest: omogoča kolacionirani prikaz razlik kot tudi vzporedni prikaz dveh (ali več) verzij besedila z vizualiziranjem razlik po posameznih besedah. Besedi v enem in drugem besedilu, ki se razlikujeta, sta obarvani, njuni vrstici sta povezani s črto. Razvidni so tudi izpuščeni ali dodani deli besedila (Slika 1); celoten program je zelo prijazen za uporabo tudi humanistu.



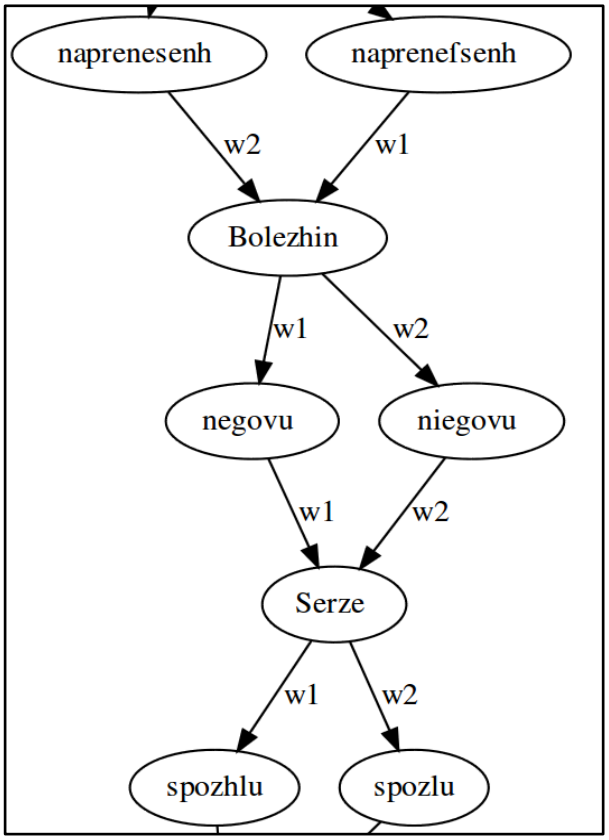
Slika 1: Vzporedni prikaz kolacije besedil v programu Juxta.

Pri CollateX, kot pravijo njeni razvijalci (Dekker in Middell, 2013), vizualizacija rezultatov trenutno ni prioriteta, pač pa program nudi večji nabor standardnih formatov izhodnih podatkov (tj. rezultatov analize – v zapisih JSON, XML-TEI, MathML idr.), ki jih je moč nato procesirati z že obstoječimi zmogljivimi orodji za grafično obdelavo, npr. s paketom Graphviz, ta omogoča mdr. generiranje variantnih grafov (Slika 2).

Glavna prednost CollateX in njena konceptualna posebnost je, da so razvijalci pomembno izboljšali stopnji 2 in 3, torej vzporejanje pojavnic in njegovo analizo, kjer tiči srčika problematike besedilne variantnosti. Razvili ali deloma prevzeli (iz računalniške biologije) so več algoritmov, ki zahtevno analizo variantnosti izboljšajo glede na tri kriterije: 1) odkrivanje transpozicij – ko so denimo stavek ali več stavkov, odlomkov, preneseni na drugo mesto v tekstu; 2) podpora za fleksibilno vzporejanje pojavnic, ko je denimo ista beseda zapisana z drugimi historičnimi črkami ali pravopisnimi razlikami, a še vedno ostaja ista beseda, kar v kombinaciji s prenosom odlomkov (kriterij 1) vzpostavlja zelo zapletene nize znakov; 3) neodvisnost razporeda verzij: razčlenjevalniki, kakšnega

[1] Servis je prosto dostopen s prijavo na naslovu http://www.juxtacommons.org/. Kot samostojen program je Juxta enostavna tako za namestitev kakor za uporabo (prim. http://www.juxtasoftware.org/).

[2] Prim. <http://collatex.net/>. CollateX je za uporabnika tehnično zahtevnejši od Juxte, zasnovan je kot javanski paket za lokalno rabo ali v spletni aplikaciji, ali v okviru knjižnic Python 3.4, njegov uporabniški vmesnik je ukazna vrstica. Vendar je tudi precej močnejši v algoritmičnem računanju razlik med verzijami.



Slika 3: Variantni graf programa CollateX.

uporablja program Juxta, pri primerjanju več kot dveh verzij najprej primerjajo glavni tekst in vsako od verzij v parih, nato pa združijo rezultate analiz parov. Vrstni red, kako združiti analize parov, vpliva na končni izračun rezultata, kar seveda ni zaželeno. Razčlenjevalnik CollateX skuša razrešiti vse te probleme z drugačno koncepcijo algoritmov (Dekker et al. 2015: 456–457).

Oba programa smo zelo uspešno uporabili za analizo variantnosti besedila *Jezusovo življenje v sto postavah,* vendar smo CollateX uporabili bolj eksperimentalno.[3] V besedilu, katerega vsaka od verzij obsega pribl. 100.000 znakov, je program Juxta našel 2.830 variantnih mest, CollateX pa kar 4.435, deloma zato, ker upošteva njegov algoritem tudi ločila in velike začetnice. Tu predstavljena analiza temelji predvsem na delu s programom Juxta, tako zaradi prijaznosti do uporabnika kakor zaradi vizualizacije razlik, saj smo variantna mesta naših dveh rokopisnih besedil lahko zelo dobro pregledovali, četudi je v eno enoto aparata Juxta pogosto združila po več dejanskih variantnih mest (zato se število dokumentiranih variant tako razlikuje).

2.2 Variantna mesta v programu Juxta in izdelava kritičnega aparata

Program Juxta temelji na glavnih konceptih tradicionalne, zlasti anglo-ameriške tekstne kritike. Vodilni koncepti so *glavno* ali *temeljno besedilo* (base text), *verzija* (witness), *variantna mesta* (readings, variant readings) in *lema* (lemma).

Temeljno besedilo je tisto, ki je med več verzijami teksta po urednikovi presoji najbližje izvirniku, avtorjevi volji ali po kakšnem drugem kriteriju najbolj adekvatno ali integralno, in je zato izbrano kot temelj, osnova izdaje.

Verzije so druge inačice istega besedila, ohranjene v raznih rokopisih, tipkopisih, izdajah ipd.

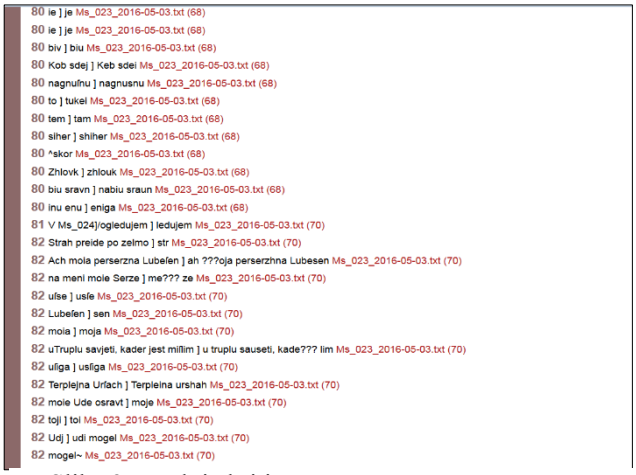
Variantna mesta so tista, kjer se verzije razlikujejo od temeljnega besedila ali med seboj. Variantno mesto ali varianta v temeljnem besedilu se v tekstni kritiki imenuje *lema*, isto mesto besedila v drugih verzijah pa se imenuje le varianta ali variantno mesto.

V teh kategorijah smo programu Juxta določili, da rokopis *Jezusovo življenje v sto postavah* iz Arhiva RS (zato *arhivski* rokopis, NRSS Ms 028) obravnava kot temeljno besedilo, *Poljanski rokopis* iz NUK (NRSS Ms 023) pa kot verzijo. Tekst, ki je ohranjen za primerjavo tako v temeljnem besedilu kakor v verziji, obsega v tem konkretnem primeru nekaj manj kot 80 strani, ker se je temeljnega besedila žal ohranilo le toliko.

Poleg dveh dobrih načinov vizualizacije variantnih mest – v kolacioniranem in vzporednem prikazu – omogoča Juxta (seveda tudi CollateX z dodatnimi možnostmi) izdelavo formalnega zapisa variantnih mest, ki je središče ekdotične problematike vseh težavnejših znanstvenokritičnih izdaj, tj. izdelavo kritičnega aparata.

Prikaz kritičnega aparata ima v filoloških vedah razne oblike in posebnosti. Juxta uporablja za prikaz način, ki je v sodobni filološki praksi najbolj razširjen (uveljavil se je mdr. z Oxfordsko izdajo Shakespearja iz tridesetih let). Temelji na načelu, da levo od oglatega oklepaja] vedno stoji *lema*, tj. varianta, ki se nahaja v temeljnem besedilu (lemmatic reading); desno od oklepaja] pa so *variante*, ki se nahajajo v verzijah (stemmatic readings; prim. Williams in Abbot 2006: 122). V našem kritičnem aparatu so torej levo od] vedno oblike zapisa v t. i. *arhivskem* rokopisu Ms 028, desno od] pa so zapisi v rokopisu Ms 023 – *Poljanskem rokopisu*, npr:

suoiga Sÿnu] soiga synu. (Slika 3)



Slika 2: Izsek iz kritičnega aparata programa Juxta.

[3] Program se uporablja v ukazni vrstici. Druga stran njegovih bolj kompleksnih in zmogljivih algoritmov je večja poraba časa za procesiranje. Vsaka verzija teksta, ki smo ga računalniško obdelali, je obsegala le okr. 100.000 znakov. Juxta je primerjavo verzij naredila v dveh sekundah, CollateX pa celote ni mogla obdelati, pač pa je bilo tekst treba v skladu z navodili razdeliti na manjše enote. Za primer: procesiranje kratkega odlomka z 10.000 znaki je trajalo na starejšem prenosniku v okolju windows 30 sekund, na zelo zmogljivem strežniku v okolju linux pa še vedno kar 10 sekund. Pri le štirikrat večjem obsegu, 40.000 znakov, se je čas procesiranja na zelo zmogljivem strežniku povečal na pet (5) minut.

V enotah aparata so tako sopostavljene besede, zaporedne besede ali daljši odlomki, ki se na kakršenkoli način razlikujejo. Če se neka beseda (ali več besed) nahaja v temeljnem besedilu, v verziji pa je ni, je označena s tildo, npr.: vidit mogu~. Če se, nasprotno, neka beseda (ali več besed) nahaja v verziji, v temeljnem besedilu pa je ni, je označena s strešico, npr.: ^she.

Številna variantna mesta se v našem aparatu izkažejo kot krajšanje in reduciranje starejšega besedila. Daljši odlomek rokopisa Ms 028 je v *Poljanskem rokopisu* lahko nadomeščen z le eno besedo. Na podlagi kolacije, izpisane v kritičnem aparatu, lahko vsaj približno tudi kvantitativno ovrednotimo posamezne tipe razlik in ugotavljamo, ali je v njih moč zaslediti kakršno koli sistematično težnjo, značilnost, zakonitost.

3 Recenzija kot jezikoslovna razčlemba

Tekstnokritična *recenzija,* podrobnejša primerjava z ugotavljanjem dejanskih razlik, ki nastopajo v programsko označenih variantnih mestih, poteka od tod naprej brez računalniške podpore – z jezikoslovno in stilno analizo pojavov v vsaki variantni besedi ali odlomku.

Med ustrezajočima besedama je lahko le ena razlika (pertiſsenla] pertisſenla, videl] vidli), pogosto pa jih je v eni besedi več (Moÿſesa] Moisesa, navſmilenu] nausmilenu, tuoie lubeſnivo] toje lubesnivu). V analizi 2.830 variant, ki jih je Juxta odkrila na nekaj manj kot 80 straneh rokopisnega besedila, smo na podlagi kolacije mogli opredeliti troje temeljnih tipov variantnih mest, ki jih predstavimo v nadaljevanju:

1. razlike v zapisu posameznih glasov;
2. razlike v jezikovni podobi;
3. razlike na besedilni ravni: krajšanja, redukcije in dodatki.

3.1 Variantnost v zapisovanju glasov

Ena od tipičnih značilnosti besedil, pisanih v bohoričici, tako rokopisnih kot tudi tiskanih, je od začetkov slovenskega knjižnega jezika v 16. stoletju dalje nedoslednost v zapisovanju črk zlasti za sičnike, šumevce, glasove u, v in dvoustnični u, i in j, sklope teh dveh glasov ter palatalna l in n. Različni avtorji v širokem časovnem razponu so te glasove zapisovali tako glede na tradicijo in uveljavljeno normo, kot tudi pod vplivom nemškega in latinskega (ter redkeje tudi italijanskega in madžarskega) črkopisa in ne nazadnje v skladu s svojim poznavanjem problematike in rešitev ter lastnimi zapisovalnimi načeli ali preferencami.

Dileme glede ustreznega zapisovanja, deloma pogojene tudi s stanjem govorjenega jezika, ki so ga avtorji prenašali v pisni prenosnik, so se na začetku druge polovice 18. stoletja še stopnjevale. Vprašanje dokončne prireditve pisave za slovenski knjižni jezik se je dokončno rešilo šele sredi 19. stoletja z zamenjavo bohoričice z gajico (Toporišič, 1998).

Pri primerjanju besedil se v *Poljanskem rokopisu* pokažejo sledeče spremembe v zapisovanju:

3.1.1 Daleč najpogostejše je nadomeščanje črke i s črko j za glas j; ta sprememba običajno zadeva samostalniške, glagolske ali pridevniške končnice ali konce osnov, kjer se j pojavi pred samoglasnikom, najpogosteje v 3. osebi ednine glagola *biti*[4] (ie] je), v 3. osebi ednine glagolov v končnici *-je* (npr. umerie] umerje, samerkuie] samirkuje) in v 3. osebi množine glagolov v končnici *-jo* (imaio] imajo, shalujeio] shalujejo). Pri pridevnikih je sprememba zelo pogosta v sklonskih oblikah svojilnih pridevnikov *moj*, *tvoj*, *svoj* (moie] moje, tuoie] toje, suoia] soja), pri samostalnikih v orodniški končnici *-jo* ednine samostalnikov druge ženske sklanjatve (Andochtio] andohtjo, Pomozhio] pomozhjo, smertio] smertjo) in imenovalniški (in obenem tožilniški) končnici ednine samostalnikov srednjega spola na *-(j)e* (Oblizhie] Oblizhje, Narozhie] narozhje).

Nasprotno je razmeroma pogosto tudi nadomeščanje črke j s črko i za glas j; najpogosteje je ta razlika v zapisu povezana z naročnim razvojem palatalnega n kot jn: v starejšem besedilu prevladujejo zapisi in, v mlajšem pa jn. Med primeri za to zamenjavo najdemo največ glagolnikov s pripono *-anje* (vſmilejna] usſmileina, shalvajne, klagvajne sdihvajne] shalvaine klagvaine, sdihvaine). Ta zamenjava je pogosta še v nekaterih drugih besedah, zelo pogosto na začetku (jskra] iskra) in na koncu besede (Farj] Fari) in največkrat ob soglasnikih d (Edjnshna] Edinshna), s (Sjdov] sidov) in l (Ljza] Liza).

Podobno za ostale črke, ki zaznamujejo glasova i in j ter njune sklope, pri nekaterih besedah opazujemo doslednost, pri drugih pa različne zamenjave v besedi z isto osnovo ali podstavo. Najbolj značilen in dosleden je zapis y namesto ÿ v oblikah besede *sin* (Sÿn] Syn), deloma tudi v osebnem zaimku *mi* in pridevniku *božji* (Mÿ] my, Boshÿ] boshi), pogosto je ÿ ali y zapisan namesto j v besedah *ljudi* in *krvi* (ludj] Ludÿ, Kervj] Kervÿ), črka i namesto ÿ je največkrat zapisana na koncu besede in je ponovno tipična zlasti za oblike pridevnika *božji* in zaimek *mi* (Boshÿ] boshi, Mÿ] mi) ter mestnik besede *naročje* (Narozhÿ] narozhi) in besedo *oči* (Ozhÿ] ozhi). Zapis ie namesto ÿ, povezan tudi s spremembo v jezikovni podobi, je običajen v sklonskih oblikah osebnega zaimka *on* (nÿm] niem, nÿh] nieh), verjetno pa je kot pomota nastala tudi zamenjava v nasprotni smeri: nieh] nyh oz. Hieronÿmus] Hÿronimus.

3.1.2 Pri zamenjavah v zapisu šumnikov in šumevcev lahko tudi opazujemo določene doslednosti, zamenjave v zapisu najpogosteje zadevajo zapis glasu s. Kjer je v starejšem besedilu ta glas zapisan z dvočrkjem ſs, je v mlajšem besedilu zapisan kot sſ skoraj dosledno v oblikah zaimkov *vse* in *ves* (uſse] usſe) in oblikah zanikanega glagola *biti* (niſso] nisſo). Poleg tega je zamenjava značilna še za oblike in tvorjenke glagolov *misliti* (premiſsel] premisſel), *viseti* (viſsi] visſi), *pisati* (popiſsat] popisſat), *pretresti* (pretreiſse] pretresſe), *prositi* (proſsem] prosſem), *nositi/nesti* (perneſsu] pernesſu), *-tisniti* (pertiſsenla] pertisſenla) ter oblike besed *beseda* (Beſsede] besſede), pri daljšanju osnove s -s- v oblikah besed *telo* (Teleſsa] Telesſa), *nebo* (Nebesſ] Nebeſs) *uho* (Usheſsa] ushesſa) ter v oblikah besed *čas* (Zhaſsu] zhasſu) in *glas* (glaſsam] glasſam). V vseh teh besedah se glas s pojavlja med samoglasnikoma. Za glas s je v mlajšem rokopisu pogost tudi zapis s namesto ſ, ponovno zlasti v oblikah in tvorjenkah k glagolom *misliti* (miſlim] mislim) in *tresti* (pretreiſle] pretresle) ter v besedah *potres* (Potreſ] potres), *Jezus* (Jeſusa] Jesusa) in z nekaj pojavitvami ali le sporadično v številnih drugih besedah. Zamenjava je značilna v vseh pozicijah, za in pred samoglasniki in

[4] Pri navajanju primerov je v besedilu običajno najprej kot osnovna oblika navedena poknjižena beseda.

soglasniki. Zapis sſ namesto starejšega ſ je omejen na oblike in tvorjenke besede *usmiljenje* (vſmilejna] usſmileina), zapis ſs namesto ſ je najpogostejši na koncu besede (npr. Derveſ] derveſs). Nekaj zamenjav pri zapisovanju glasu s je sporadičnih.

Za glas z je značilen le zapis s namesto ſ, najpogosteje v besedi *ljubezen* in njenih tvorjenkah (Lubeſen] Lubesen). Pri zapisu glasu c je sprememba v besedi *cesarske* (zeſsarske] Ceſsarske), v kateri je v mlajšem rokopisu glas c zapisan s črko c.

Pri šumnikih je v starejšem besedilu glas š položajno, po nemškem zapisu oz. po analogiji, zapisan s črko s, in sicer pred t (stir] shtir), k (skarlata] shkarlata), p (spotliva] shpotliva, Spegu] shpegu), l (slahtnim] shlahtnim) in n (Snabel] shnabel). Ravno tako je po nemškem zapisu v starejšem besedilu glas š zapisan tudi s sklopom črk sch, ta zapis je značilen zlasti na nemške izposojenke *šac* (Schaz] shaz), *šajhati* (schaiham] shaiham), *šahar* (Schaharje] shaharje), pogost pa je tudi v prislovu *še* (sche] she) in deležniku glagola *iti* (schou] shou). V mlajšem rokopisu je v vseh teh primerih dosledno zapisano dvočrkje sh.

Podobna razlika v zapisu je značilna tudi za glas č, ki je v starejšem rokopisu v določenih besedah zapisan s črko z, medtem ko je v novejšem zapisan kot zh. Ta zamenjava je najbolj tipična za glas č, ki je nastal po jotaciji, in sicer v tvorjenkah besed *srce* (Serzne] serzhne), *resnica* (reſniznu] resnizhnu) in *konec* (pokonzanh] pokonzhanh), pogost pa je tudi v oblikah besede *mrlič* (Merliza] Merlizha) ter sporadično zapisan še v nekaterih drugih besedah.

3.1.3 Zaradi še neustaljenega zapisa so značilne številne zamenjave za zapisovanje glasov u, v, pri katerih se izmenjujeta črki u in v. Zamenjave se pojavljajo skoraj v vseh položajih, opazna doslednost je le, da je v mlajšem rokopisu črka v (za dvostnični u) na koncu besede dosledno spremenjena v črko u (glav] glau).

3.1.4 Po nemščini prevzeto je tudi zapisovanje glasu h s ch v starejši verziji, kar je v mlajšem prepisu dosledno nadomeščeno s črko h. Tudi ta zapis je omejen na določene besede, in sicer dosledno na medmet *ah* (Ach] Ah) ter oblike in tvorjenke nemške izposojenke *andoht* (Andochtio] andohtjo) ter sporadično v mestniku nekaterih neprevzetih samostalnikov, npr. Miſlach] mislah, Omedlevzach] omedleuzah, in v nekaterih drugih besedah.

3.1.5 V zvezi z zapisovanjem glasov je značilno še, da je zapisovalec mlajšega rokopisa namesto dvojnih črk na več mestih zapisal le eno črko, in sicer t namesto tt (dosledno v oblikah in tvorjenkah besede *mati* (Matternu] maternu) in po enkrat v nekaterih drugih besedah), n namesto nn (Shenne] shene), redkeje l namesto ll (omejeno na pretekli deležnik na -l za ženski spol, npr. pozhella] pozhela) in f namesto ff zlasti v nemških izposojenkah. Obenem pa je avtor mlajšega *Poljanskega rokopisa* v oblikah besed *kolena* (Kolena] Kollena), *telo* (Telu] Tellu) in *želja* (shele] shelle) skoraj dosledno namesto enojne črke zapisoval dvojno.

3.1.6 Na en primer je omejen zapis th za glas t, v mlajšem rokopisu zapisan s t (threnu] trenu), medtem ko je v besedi *sobota* (Sabboto] Sabbotho) značilen nasproten proces. Med redkejšimi spremembami je tudi zamenjava latinskega æ v e, najznačilnejša za imenovalnik množine samostalnika *rana* (Ranæ] rane).

3.1.7 V zvezi z zapisovanjem je v *Poljanskem rokopisu* opazna tudi večja podomačitev v nekaterih izposojenkah (tausend] taushent, nafeuchtane] nafaihtane, Ursach] urshah).

Povzamemo lahko, da je ena glavnih ali pomembnejših sprememb v mlajšem rokopisu glede na starejšega bistven odmik od zgledovanja po nemškem zapisovanju, redkeje v zapisu besed in zlasti v zapisu posameznih glasov. Ta odmik pa ne pomeni linearnosti v smislu razvoja načina zapisovanja, saj so nekatera načela, ki jih glede na arhivsko verzijo uvaja *Poljanski rokopis*, v teoriji in zlasti v praksi v veliki meri uveljavljali že protestanti (npr. zapis h za h namesto ch ali v nekaterih primerih t namesto th za glas t). V skladu s protestantsko tradicijo je tudi zapisovanje glasu c s črko c pred i in e po zgledu latinske pisave, ki je značilno za *Poljanski rokopis*, ne pa tudi za starejšo predlogo.

Vse te spremembe, zlasti tiste, pri katerih opažamo doslednost in sistematičnost, kažejo na visoko jezikovno zavest pisca *Poljanskega rokopisa*. Za celotno besedilo *Poljanskega rokopisa* lahko ugotovimo sicer nekatere tipične nedoslednosti v zapisovanju določenih glasov (največ različnih načinov zapisovanja je značilnih ravno za glas s), nedvomno pa tudi zapisovalno sistematičnost.

3.2 Variantnost v jezikovnih pojavih

Za slovenski jezikovni prostor so ravno za obdobje od srede 18. do srede 19. stoletja (starejše rokopisno besedilo je nastalo v sredini in mlajše ob koncu 18. stoletja) značilne različne regionalne tendence. Zgodnja narečna jezikovna cepitev in stare upravnopolitične delitve so spodbudile nastanek štirih pokrajinsko narečno naravnanih knjižnih jezikovnih različic: ob kranjski tudi koroška in štajerska ter prekmurski knjižni jezik. Najpomembnejši in normativno najbolj izoblikovan je bil kranjski knjižni jezik s tradicijo od 16. stoletja, v katerem so v starejšem obdobju slovenske knjižne tradicije, nekako do tridesetih let 18. stoletja, še prevladovali dolenjsko-notranjski narečni pojavi, od druge polovice 18. stoletja pa ga vedno bolj opredeljujejo tipično gorenjske, celo rovtarske narečne značilnosti. Proti koncu 18. stoletja (1786–1802) je knjižno normo zaznamoval tudi jezik prvega katoliškega prevoda *Svetega pisma*, s katerim je kranjska različica dosegla razmeroma umerjeno ustalitev.

Vendar moramo pri presojanju jezikovne podobe besedila upoštevati tudi ali zlasti dejstvo, da se primerjani besedili uvrščata v rokopisno tradicijo druge polovice 18. stoletja, ki ni bila tako zavezana sočasni bolj ali manj uveljavljeni knjižnojezikovni normi in za katero je na sploh značilen še večji vnos sovpad z govorjenimi glasovnimi in oblikovnimi pojavi, ne glede na to, ali gre za hoteno, zavestno uporabo nekaterih posebnosti govorjenega jezika ali za nezavedno in nehoteno zapisane lastnosti in posebnosti (Orel 1998, Orel, 2003).

Razlike so obravnavane po posameznih jezikovnih ravninah, glasovna in oblikovna sta zaradi povezanosti obravnavani skupaj. Včasih najdemo primer za spremembo v obe smeri v isti besedi (npr. sim] sem in tudi sem] sim za 1. osebo ednine glagola *biti*).

3.2.1 Na glasovni in oblikovni ravnini pri samoglasnikih[5] lahko opazujemo sledeče spremembe: a > o, in sicer najpogosteje v 3. osebi množine glagola biti *so* (sa] so), večkrat v predponi *da-* > *do-* (dashezhi] dosezhi), pogosteje tudi v besedi *mogoče* (magozhe] mogozhe). Ta sprememba vpliva tudi na sklanjatveno končnico samostalnikov in pridevnikov (Imenam] imenom ali Serzna Andochtio] serzhno andohtjo). Pri sicer zelo redkem prehodu a > e, omejenem na pridevniško končnico, je značilna delna feminizacija nevter (niegoua Kolena] negove Kollena), vendar omejena le na nekaj primerov. Prehod a > e in v istem tipu tudi e > a se pojavi v nekaj primerih v dajalniški in orodniški množinski samostalniški končnici *-am(i)* (Ustem] Ustam, greshnikami] greshnikem).

Upad e > i je značilen v položaju pred r (samerkat] samirkat), kjer gre v bistvu za zapis po dejanskem izgovoru, ter v pridevniški končnici *-em* oz. *-im* (grenkem] grenkim, spizhastem] shpizhastim), vendar najdemo tudi primere za spremembo *-im* > *-em* (drugim] drugem). Le enkrat je za pridevniško rodilniško končnico značilna zamenjava -ega > -iga (narlubshega] narlubshiga), medtem ko je za rodilnik zaimka *ta* v nekaj primerih značilna nasprotna sprememba -iga > -ega (tiga] tega).

Zamenjava nenaglašenega o > e je dosledna v besedi *vendar* (vonder] vender), oblikah in tvorjenkah k *pekel* (Poklam] peklam) in besedi, zliti iz *ko bi* (kob] keb); prehod o > u je značilen za začetek (npr. obtemnele] utemnele), konec (zelmo] zelmu) in sredino besede (Kop] Kup). Obenem je v vseh položajih značilna tudi nasprotna sprememba u > o (ubuden] obuden).

Pri prehodu u > a, omejenem na tri primere zapisa besede Christusu] Christusa, gre za različno tvorbo svojilnega pridevnika (besedotvorno s pripono -ov > -u in izražanje svojine z rodilnikom). Ta sprememba je povezana tudi z asimilacijo -ou > -u (otou] otu), pri čemer je za mlajši rokopis značilna tudi sprememba v nasprotni smeri (-u > -ou: Poku] pekou).

Za diftonge je značilna dosledna monoftongizacija ie > e v besedi *že* (shie] she), ie > i v besedi *vera* oz. *verovati* (vierval] virval) ter dosledna sprememba ej > ie v samostalniku *svet* (Sveitu] svieto) in enkrat v samostalniku *leto* (lejt] Liet).

Od ostalih samoglasniških pojavov so kot najpogostejši značilni še dokaj številni upadi in redukcije, najpogosteje na koncu besede in najpogosteje i (katiri] katir), vendar tudi v drugih položajih in za druge samoglasnike; ter nasprotno – v primerjavi s starejšim besedilom – ohranitev reduciranega glasu (Shvot] shivot).

3.2.2 Med soglasniškimi pojavi je najpogostejša dosledna sprememba asimilacija sv-, tv- > s-, t- na začetku besede v pridevniških zaimkih *tvoj*, *svoj* (tuoie] toje, suoiga] soiga) in na eno besedo oz. en primer omejena redukcija bt > t (obtemnele] otemnele). Dosledna je zamenjava asimiliranega glasu z neasimiliranim v oblikah besed *sovražnik* (Shourashniki] sourashniki) in *služabnik* (Shlushabnik] slushabnik), pri čemer je prekozložna asimilacija š-s > š-š tako starejšem kot tudi v mlajšem besedilu ohranjena v oblikah glagola *slišati* (v starejšem besedilu z značilnim zapisom glasu š v položaju pred l s črko s – slishal] shlishal), Za palatalni n so značilne praktično vse možne spremembe: n je nadomeščen z nj v sklonskih oblikah osebnih zaimkov *on*, *ona* (No] nio, neh] nieh) in z jn v glagolnikih (prim. še nadomeščanje črke j z i) (pokanam] pokainam); namesto nj je zapisan n v oblikah svojilnih zaimkov *njegov* in *njen* (niegoua] negova, nenga] nienga), razmeroma pogost pa je še zapis n namesto jn tudi pri glagolnikih (Shivlejne] shiulene).

Le na besedo ali dve so omejene še naslednje soglasniške spremembe: v > b v besedi *nevesta* (Nevesta] Nebesta), nezapis zvenečnostne premene v oblikah besede *hrbet* (Herptam] herbtam), redukcija proteze v- pred u (Vudi] udi) in dodajanje vzglasnega soglasnika (obenga] nobenga) v mlajšem besedilu.

3.2.3 Na oblikovni ravnini lahko poleg že navedenih sprememb ugotavljamo spremembe, omejene na nekaj posameznih besed, npr. redukcija daljšanja osnove na -r z -j- (Romarje] Romare), redukcija daljšanja osnove z -ov- v enozložnici *rob* (Robovam] Robem) ali rodilniška končnica -a namesto -u v enozložnici *sin* (Sÿnu] sÿna).

3.2.4 Za besedje so značilne številne razlike v določanju besednih mej, ki se v večini nanašajo na zveze predloga in sledeče besede ali zveze prislova in glagola, večinoma kalke po nemških glagolih. Zelo redke so spremembe v besedotvorju, ki v nekaj primerih povzročijo spremembo glagolskega vida (pertiskala] pertisſenla), nekajkrat, v večini primerov pa ne, so okrajšave izpisane z besedo (S] svet). Opazno je tudi nadomeščanje besed s sopomenkami (klagovat] jokat, prelivala] souse tozhila, mogozhnih] slovezhih, grauſat] gnusit, obvarvu] ohranu). Kolacija, ki jo izpiše program Juxta, pokaže tudi precej zamenjav besed, ki niso sopomenke, ter veliko dodanih in izpuščenih besed in besednih zvez, katerih vlogo in vpliv na besedilo moramo opazovati glede na sobesedilo.

Ne nazadnje primerjava besedil pokaže, da je avtor mlajšega besedila pri preoddaji popravil nekatere očitne pisarske pomote iz starejšega besedila, obenem pa tudi sam napačno prepisal nekatere besede.

Za spremembe v sami jezikovni podobi besedila lahko ugotovimo, da je zlasti izstopajoča in dosledno uveljavljena asimilacija tv-, sv- > t-, s-, značilna za gorenjska in rovtarska narečja. Sodobnejše jezikovno stanje v *Poljanskem rokopisu* izkazujejo tudi vsaj delna ukinitev akanja in vsaj delna monoftongizacija ter dosledni zapisi *vender* namesto *vonder* in *peku* namesto *poku* (z vsemi oblikovnimi in besedotvornimi različicami). Bistvenih sprememb sklanjatvenih ali spregatvenih paradigem ni, tudi zamenjava besed s sopomenkami ne vpliva bistveno na jezikovno podobo, saj sopomenke niso diatopične.

3.3 Variantnost na besedilni ravni: izpusti in povzemanja

Na besedilni ravni opazimo, da je pisec *Poljanskega rokopisa* starejši tekst iz *arhivskega* rokopisa na mnogih mestih prepisal s spremembami, ki jih je moč razdeliti v tri kategorije.

3.1 Tekst, ki se nahaja v *obeh rokopisih.* Pisec je v *Poljanskem rokopisu* spremenil le posamično besedo v stavku, bodisi kot sopomenko ali pa je mislil, da popravlja

[5] Nadsegmentne lastnosti samoglasnikov (trajanje, mesto in vrsta naglasa) v analizi niso upoštevane, saj iz samega zapisa tudi niso razvidne.

napako in izboljšuje slogovno podobo besedila, denimo ob pojavih, o katerih poroča Evangelij, da so se godili v trenutkih po Jezusovi smrti na križu (popravek v(z)buditi > biti):

Kai ie moglu kul to sa en grosoviten Strah ubodit?] Kai je moglu kul to sa en grosoviten strah bit?

Sprememb te vrste je na besedilni ravni veliko, zdi se, da nakazujejo težnjo k večji razumljivosti.

3.2 Tekst, ki v *arhivskem* rokopisu je, a ga v *Poljanskem* ni. To so številne redukcije in krajšanja besedila v obsegu od ene besede do celega stavka in tudi daljših odlomkov. Tako denimo v prizoru, ko Mati Božja žaluje ob Jezusovi smrti, arhivski rokopis razvije zaporedje retoričnih členov, ki se prelije v sklepno svetopisemsko podobo iz preroka Izaije, v *Poljanskem rokopisu* pa je del tega sestava izpuščen:

Kdu si more ena taka shalost damiſselt! Kdu more ismert, kok velika ie ena taka shalost bla?] Kdu si more ena taka shalost damisſelt.

Ta primer ponazarja redukcijo retorične ali literarne strukture z namenom, da bi bil tekst krajši in preglednejši. Podobne krajšave enega ali dveh stavkov so zadele številna mesta, kjer so bili prizori Jezusovega trpljenja prikazani za okus pisca *Poljanskega rokopisa* morda preveč drastično in nazorno in jih je skušal omiliti. Tretja skupina posegov tega tipa pa so izpusti daljših odlomkov po več stavkov, tudi pol strani, ki ponekod stopnjujejo meditativno naravo besedila, duhovno ali mistično razsežnost prikazanih dogodkov; ponekod je izpuščena celotna molitev, s katero se v pasijonskem delu pripovedi sklepajo poglavja. Vse to kaže na določen vpliv razsvetljenske dobe: čeravno je pisec *Poljanskega rokopisa* cenil in ohranjal baročno duhovno izročilo, ga je vendar v manjši meri tudi prirejal, povzemal in krajšal, da bi prvotnemu baročnemu tekstu odvzel nekaj tega, kar se mu je v podoživljanju Jezusovega trpljenja utegnilo ob začetku 19. stoletja zdeti preveč spiritualistično ali v prikazovanju preveč čutno in drastično.

3 – Tretja skupina variantnih mest na besedilni ravni je najbolj presenetljiva: to je tekst, ki ga v *arhivskem* rokopisu ni, v *Poljanskem* pa je. Samo tri mesta v vsem ohranjenem vzporednem korpusu so takšna. Eno je manjše; drugi dve sta večji in zelo zanimivi. Eno od njiju, ki govori o potresu ob Jezusovi smrti, se glasi:

jnu ie bil tok strashan, de so Ludje menil, uſse more na Kop padlu] inu je bil tok strashan, de so Ludje menil, usſe more na kop past, koker je tudi dost imenitnih mest inu Tergu na kop padlu

To mesto, kot vidimo, je zelo težavno (in drugo o snemanju s križa prav tako). Vidimo namreč, da je v *arhivskem* rokopisu nastal izpust besedila, stavek se konča z nepravilno oblikovanim povedkom *more na Kop* [izpust] *padlu*. V *Poljanskem rokopisu* je to mesto popolno in se smiselno sklene z odvisnim stavkom. To vrsto besedilne variacije bi naravno razložili tako, da stoji desno od oklepaja] starejši tekst, ki je bogatejši, levo pa mlajši, ki je reduciran. Toda v resnici je obratno; mlajši tekst je na teh dveh pomembnih mestih obširnejši in popoln, starejši pa okrnjen. Če se poglobimo v naravo izpusta v *arhivskem* rokopisu, vidimo, da je pisec *arhivskega* rokopisa izpustil del stavka in naredil pokvarjen povedek, česar bi verjetno nikdar ne storil, če bi sam prevajal iz nemščine; pač pa je to prav lahko storil, če je tudi sam prepisoval iz nekega predhodnega rokopisa, protografa. Zdi se, da je bilo tako. Arhivski rokopis, za katerega smo doslej mislili, da je bil prvotni rokopis, *arhetip*, se tako izkaže za prepis, *apograf* starejše rokopisne tradicije.

Možni sta dve razlagi problema: če je po našem dosedanjem sklepanju pisec *Poljanskega rokopisa* prepisoval iz *arhivskega* rokopisa, je tu opazil napako in jo popravil bodisi tako, da je imel pri roki nemški izvirnik ter iz njega prevzel manjkajoči odvisni stavek, bodisi tako, da je imel na voljo drug, nam neznan rokopis s slovenskim prevodom in je odvisnik prevzel iz njega.

4 Sklep

Skleniti moramo, da nam je strojno podprta kolacija *arhivskega* rokopisa (NRSS Ms 028) in *Poljanskega rokopisa* (NRSS Ms 023) v evidenco prinesla množico variantnih mest, iz katerih je moč vsaj hipotetično povzeti nekatere težnje prepisovalca, ki so oblikovale njegovo verzijo teksta – *Poljanski rokopis*.

Iz primerjave starejšega in mlajšega besedila lahko ugotovimo, da glavnino razlik, tudi kvantitativno, prepoznamo v zapisu črk za posamezne glasove; posameznih razlik, ki bi lahko pomembno vplivale na spremembo same jezikovne podobe besedila, je razmeroma malo ali so omejene na nekaj besed, poleg tega nekatere potekajo v obe smeri. Značilne pa so tudi nekatere dosledne in sistematične spremembe, ki kažejo v smeri sočasnih sodobnih teženj k oblikovanju enotnega knjižnega jezika, temelječega na gorenjski in deloma rovtarski osnovi.

Izsledki te primerjave odpirajo izhodišče za natančnejšo analizo zapisa določenih glasov in za jezikoslovno analizo obeh besedil. Sorazmerno majhno število različnih primerov za jezikovne razlike lahko razložimo s tem, da besede, v katerih bi še lahko ugotavljali enake ali podobne jezikovne pojave in posledično spremembe, v besedilu niso zapisane. Dokončnejše rezultate bi dala analiza celotnega besedila obeh rokopisov, v kateri bi poleg razlik upoštevali tudi enakost jezikovnih pojavov, ki v tukajšnji primerjavi razlik seveda niso bili zajeti. Nekatere spremembe, ki so se v tej primerjavi pokazale kot sporadične ali omejene na nekaj primerov, se lahko v kontekstu celotnega *Poljanskega rokopisa* pokažejo kot sistematično zapisovalno ali jezikovno načelo avtorja ali se potrdijo kot nedoslednosti. Možno je tudi, da je avtor *Poljanskega rokopisa* s spremembami poskušal jezikovno podobo starejše predloge poenotiti, ne nujno spremeniti.

Na besedilni ravni številni krajši in daljši izpusti besedila kažejo na poskuse poenostavljanja in zgoščanja kompleksnega starejšega meditativnega besedila. Tri variantna mesta iz nenavadne skupine, kjer je starejši rokopis okrnjen, novejši pa popoln, izpričujejo, da se je pisec *Poljanskega rokopisa* zavedal nekaterih prepisovalnih izpustov arhivskega rokopisa in da je imel neki vir, iz katerega je manjkajoče besedilo prevzel, ter tako tri mesta, ki so v arhivskem rokopisu poškodovana, v *Poljanskem rokopisu* popravil.

Vsekakor odkritje teh treh variantnih mest omogoča zanimive nove raziskave, saj so povsem nepričakovano, toda jasno izpričala, da je arhivski rokopis *Jezusovega življenja v sto postavah* prepis starejše besedilne tradicije, ki se je v baročni dobi širila na Slovenskem s pomočjo rokopisne kulture.

5 Literatura

Ronald Haentjens Dekker, Gregor Middell. 2013. *CollateX – Software for Collating Textual Sources.* Documentation. http://collatex.net/doc/

Ronald Haentjens Dekker et al. 2015. Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Digital Scholarship in the Humanities*, 30(3).

A. S. G. Edwards. 1995. Middle English Literature. V: *Scholarly Editing. A Guide to Research.* Ed. by. D. C. Greetham. New York: MLA.

NRSS, 2011. *Register slovenskih rokopisov 17. in 18. stoletja.* Ur. M. Ogrin. Ljubljana: ZRC SAZU, IJS. http://ezb.ijs.si/nrss/

Irena Orel. 1998. Oblakov oblikoslovni in skladenjski prispevek v obravnavi starejših slovenskih besedil. V: *Obdobje baroka v slovenskem jeziku, književnosti in kulturi: mednarodni simpozij v Ljubljani od 1. do 3. julija 1987; Obdobja* 9, str. 183–94. Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete v Ljubljani.

Irena Orel. 2003. Slovenski pisni jezik nekoč in danes – med izročilom in govorom. V: *Slovenski knjižni jezik – aktualna vprašanja in zgodovinske izkušnje: ob 450-letnici izida prve slovenske knjige; Obdobja 20*, str. 551–62. Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete.

Martina Orožen. 1984. Slovnična in besediščna preobrazba Dalmatinovega knjižnega jezika ob Japljevem prevodu Biblije (1584–1784–1802). *Protestantismus bei den Slowenen / Protestantizem pri slovencih, Wiener Slawistischer Almanach, Sonderband*, 13:153–177.

Martina Orožen. 1985. Smernice knjižnega jezikovnega razvoja od Jurija Dalmatina do Jurija Japlja (1584–1784). *Jezik in slovstvo*, 30(7/8):217–223.

Fran Ramovš. 1924. *Historična gramatika slovenskega jezika. 2, Konzonantizem.* Učiteljska tiskarna.

Fran Ramovš. 1935. *Historična gramatika slovenskega jezika. 7, Dialekti.* Učiteljska tiskarna.

Jože Toporišič. 1986. Bohoričica 16. stoletja. V: *16. stoletje v slovenskem jeziku, književnosti in kulturi: mednarodni simpozij v Ljubljani od 27. do 29. junija 1984; Obdobja 6*, str. 271–305. Filozofska fakulteta.

Jože Toporišič. 1998. Bohoričica 17. in prve polovice 18. stoletja. V: *Obdobje baroka v slovenskem jeziku, književnosti in kulturi: mednarodni simpozij v Ljubljani od 1. do 3. julija 1987; Obdobja 9*, str. 233–52. Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete v Ljubljani.

Popisi prebivalstva Slovenije 1830–1931
Orodje za transkribiranje historičnih demografskih podatkov

Andrej Pančur*

* Inštitut za novejšo zgodovino
Kongresni trg 1, 1000 Ljubljana
andrej.pancur@inz.si

Povzetek

Avtor v prispevku predstavi orodje za transkribiranje historičnih demografskih podatkov. To je bilo razvito v sklopu projekta Popisi prebivalstva Slovenije 1830–1931, ki se odvija na Inštitutu za novejšo zgodovino v okviru raziskovalne infrastrukture Slovenskega zgodovinopisja. Projekt ima dvojen namen, saj želi zadovoljiti tako potrebe ustanov s področja varstva kulturne dediščine po čim lažjemu dostopu širše javnosti (prvenstveno rodoslovcev) do njihovega gradiva, kot tudi potrebe raziskovalcev, ki se ukvarjajo s historično demografijo, po čim širšem naboru relevantnih raziskovalnih podatkov.

Slovenian Population Censuses 1830–1931
Tool for the Transcription of Historical Demographic Information

In his contribution the author presents the tool for the transcription of historical demographic information, developed during the project entitled "Popisi prebivalstva Slovenije 1830-1931" (Slovenian Population Censuses 1830–1931), taking place at the Institute of Contemporary History in the context of the Research Infrastructure of Slovenian Historiography. The purpose of the project is twofold, as it attempts to satisfy the need of the institutions working in the field of cultural heritage protection to ensure that the wider public (especially genealogists) can benefit from easy access to these institutions' materials; as well as the need of the researchers, dealing with historical demographics, to have the widest possible collection of relevant research information at their disposal.

1 Uvod

Uporaba digitalnih orodij je sestavni del vsakršne raziskovalne metode v digitalni humanistiki. Orodja za transkribiranje lahko klasificiramo kot posebno skupino digitalnih orodij (Puhl et al., 2015, 22-23),[1] ki se jih relativno pogosto uporablja v projektih iz digitalne humanistike. Uporaba teh orodij se je zlasti razširila v zadnjih letih, ko se vse več projektov pri prepisovanju (historičnih) podatkov odloča za uporabo spletnih sistemov za izkoriščanje moči množic (crowdsourcing). Digitalna orodja za transkribiranje lahko pri tem razvrstimo v tri večje skupine (Noll, 2013, 10-11):

- vpisovanje v prost obrazec,
- XML ali HTML označevanje,
- vpisovanje v podatkovna polja.

Velika večina projektov pri tem uporablja orodja iz prve skupine, najmanj pa iz druge skupine. Digitalna orodja, pri katerih se podatke vpisuje v različna podatkovna polja se prvenstveno uporablja pri projektih, ki so povezani z rodoslovjem ali z raziskavami s področja demografije. Takšni projekti praviloma najprej poskrbijo za digitalizacijo relevantnih historičnih virov (matične knjige, popisi prebivalstva, davčni registri, domovinske knjige itd.), čemur sledi transkripcija podatkov. Ker so podatki v teh historičnih dokumentih ponavadi pisani z roko, poteka tudi njihovo prepisovanje ročno. Trenutno prepoznavanje rokopisov (Handwritten Text Recognition - HTR) pri historičnih besedil še ni zadosti razvito za množično uporabo, (Fornés et al., 2014) čeprav so se že pojavljali prvi projekti, ki so jo implementirali.[2] V rodoslovnih projektih se ponavadi prepisuje zgolj osnovne podatke, ki so potrebni za indeksacijo. Uporabnik s pomočjo indeksiranih podatkov poišče digitalizirano sliko, s katere si prebere ostale (neprepisane) podatke. V raziskovalnih projektih se ponavadi prepiše vse podatke, ki se jih nato klasificira v skladu z raziskovalnimi potrebami. Pri tem ni nujno, da se ohrani povezava med prepisanimi podatki in digitalizirano sliko.

Projekt Popisi prebivalstva Slovenije 1830–1931, ki se na Inštitutu za novejšo zgodovino (INZ) že pet let odvija v okviru raziskovalne infrastrukture Slovenskega zgodovinopisja, je primer projekta, ki združuje tako elemente rodoslovnih kot raziskovalnih projektov.

2 Popisi prebivalstva Slovenije 1830–1931

Digitalizirani historični popisi prebivalstva Slovenije so javno dostopni na portalu Zgodovina Slovenije – SIstory.[3] Projekt se izvaja v tesnem sodelovanju z Zgodovinskim arhivom Ljubljana (ZAL), ki hrani večjo število popisov prebivalstva, kateri vsebujejo mikropodatke o takratnem prebivalstvu. Tako so v celoti ohranjeni popisi prebivalstva Ljubljane (1830, 1857, 1869, 1880, 1890, 1900, 1910, 1921, 1928 in 1931). Relativno zelo dobro so ohranjeni še popisi za Idrijo, Novo mesto, Škofjo Loko in Vrhniko, delno pa tudi za različne podeželske občine. ZAL se je projektu pridružil, ker želi to gradivo dati preko spleta na voljo svojim uporabnikov, v prvi vrsti rodoslovcem.

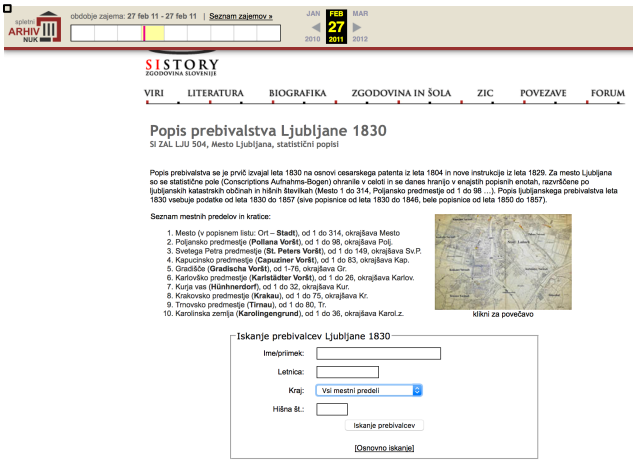
Sprva je ZAL sam digitaliziral popis prebivalstva Ljubljane 1830–1857. Ker je želel, da bi bil popis čim bolj dostopen širši javnosti, se je odločil, da bo popise objavil na portalu SIstory. V ta namen so bili v okviru dejavnosti raziskovalne infrastrukture s pomočjo optičnega prepoznavanja znakov (OCR) iz analognega indeksa (tipkopis iz leta 1934) pridobljeni osnovni podatki o prebivalcih Ljubljane iz popisa 1830–1857. Na podlagi

[1] DIRT: Digital Research Tools, http://dirtdirectory.org/tadirah/transcription.

[2] Transkribus, https://transkribus.eu/Transkribus/.

[3] Popisi prebivalstva Slovenije 1830-1931, Zgodovina Slovenije – Sistory, http://sistory.si/publikacije/?menu=510.

teh podatkov je bil nato izdelan iskalnik po popisih prebivalstva (glej Sliko 1).[4]



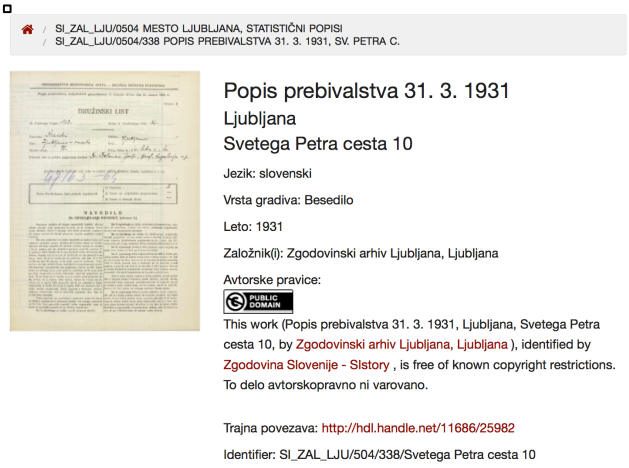
Slika1: Popis prebivalstva Ljubljane 1830: napredni iskalnik, Spletni arhiv NUK, 27. 2. 2011, http://nukrobi2.nuk.uni-lj.si:8080/wayback/20110227114943/http://www.sistory.si/popis_prebivalstva_1830_napredno-iskanje.html.

Že kmalu pa so se pokazale določene pomanjkljivosti takšnega pristopa pri objavljanju popisov prebivalstva:

- le za manjši del popisov so obstajali analogni indeksi, ki bi omogočili enostavno izdelavo elektronskih indeksov;
- obstoječi analogni indeksi niso zajemali vseh popisanih oseb (v glavnem le glave družine);
- ker so analogni indeksi nastali pred drugo svetovno vojno, so zajemali tudi osebe iz arhivskega gradiva, ki danes ni več ohranjeno.

Zato smo se odločili, da bomo po zgledu na rodoslovne projekte indeksacijo izvajali s prepisovanjem osebnih podatkov. Trenutno so na portalu SIstory objavljeni popisi prebivalstva Ljubljane (1830–57, 1869, 1921 - 1. del, 1931), občine Vrhnika (1870, 1880, 1890, 1900, 1910) in nekaterih občin okrajnega glavarstva Novo mesto iz leta 1869 (Dobrnič, Mirna, Velika Loka, Bela Cerkev, Črmošnjice, Kočevske Poljane, Prečna, Trebnje in novomeško predmestje Kandija). Za objavo se pripravlja še popis prebivalstva Ljubljane 1857. V sklopu digitalizacije vseh teh popisov je bilo narejenih 84000 slik in prepisani podatki za več kot 142000 oseb. V sodelovanju s FamilySearch[5] so bili leta 2015 digitalizirani še vsi ostali historični popisi, ki jih hrani Zgodovinski arhiv Ljubljana, skupaj več kot 270000 slik.

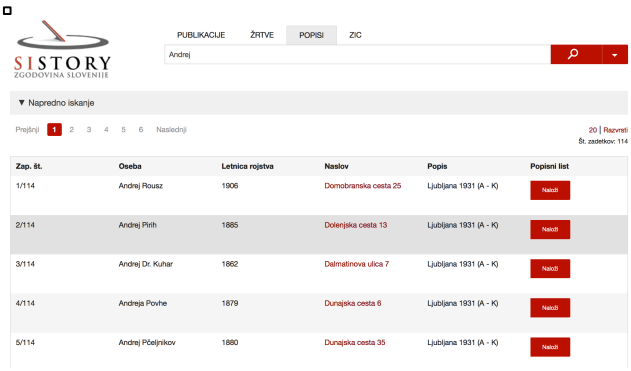
Na portalu SIstory so kot PDF publikacije objavljene popisnice za posamezne hiše. Metapodatke teh publikacij je mogoče iskati s pomočjo splošnega SIstory iskalnika. Metapodatki ne vključujejo informacij o osebah, temveč samo o kraju (naselje, ulica in hišna številka) in času popisa.



Slika 2: Popis prebivalstva kot publikacija na portalu SIstory.

Za iskanje podatkov o osebah, ki so bile popisane v okviru posameznih hiš, je predviden poseben iskalnik.[6] Sredi leti 2016 smo za iskalnik začeli uporabljati Elasticsearch, kateri temelji na Apache Lucene.[7] Iskalnik smo konfigurirali v skladu s potrebami rodoslovnih projektov:

- iskanje po osnovnih podatkih: ime in priimek, leto rojstva, bivališče (naslov) in hišna številka;
- uporaba n-gram algoritma za iskalne nize, ki se ne ujemajo povsem z zapisi v bazi;
- filtriranje rezultatov po popisih;
- povezava na sliko popisnega lista z ostalimi podatki o iskani osebi;
- povezava na PDF publikacije hiše (naslov hiše), v okviru katere je bila popisana iskana oseba.



Slika 3: Iskalnik po indeksu oseb iz popisov prebivalstva.

Iskalnik podatke o osebah zajema iz orodja za transkribiranje podatkov. V tabeli 1 so prikazani vsi digitalizirani popisi prebivalstva, katerih slike so bile že uvožene v to orodje (1), prepisani podatki o osebah pa omogočajo iskanje teh oseb na portalu SIstory. V tabeli je prikazano tudi število slik tistih digitaliziranih popisov, katere nameravamo v naslednjih letih postopoma uvoziti v orodje za transkribiranje podatkov in hkrati kot PDF publikacije tudi na SIstory.

[4] Podatki iz prvotne baze so še vedno dostopni kot Popis prebivalstva Ljubljane 1830: Tabelarni prikaz podatkov z iskalnikom in povezavami na digitalizirano gradivo, Zgodovina Slovenije – Sistory, http://sistory.si/SISTORY:ID:26731.

[5] FamilySearch, https://familysearch.org/.

[6] Popisi prebivalstva, Iskanje, Zgodovina Slovenije – SIstory, http://sistory.si/popis.

[7] https://www.elastic.co/products/elasticsearch.

popis	št. slik	št. oseb	orodje
Mesto Ljubljana 1830-1857	5069	16758	1
Mesto Ljubljana 1857	5260	15719	0
Mesto Ljubljana 1869	14627	23088	1
Mesto Ljubljana 1880	12934		0
Mesto Ljubljana 1890	22362		0
Mesto Ljubljana 1900	28198		0
Mesto Ljubljana 1910	36806		0
Mesto Ljubljana 1921	12841	2919	1
Mesto Ljubljana 1921	14559		0
Mesto Ljubljana 1928	36568		0
Mesto Ljubljana 1931	29561	59737	1
Občina Vrhnika 1870	576		1
Občina Vrhnika 1880	257		1
Občina Vrhnika 1890	992	2793	1
Občina Vrhnika 1900	657	4066	1
Občina Vrhnika 1910	705		1
Mesto Novo mesto 1870	284		0
Okr. glav. Novo mesto 1857	675		0
Okr. glav. Novo mesto 1869	10917	17609	1
Okr. glav. Novo mesto 1880	2869		0
Okr. glav. Novo mesto 1890	551		0
Okr. glav. Novo mesto 1900	506		0
Okr. glav. Novo mesto 1910	3026		0
Okr. glav. Novo mesto 1931	2947		0
Občina Čekovnik 1900	94		0
Občina Čekovnik 1910	138		0
Občina Čekovnik 1921	115		0
Občina Dole 1890	298		0
Občina Dole 1900	468		0
Občina Dole 1910	470		0
Občina Dole 1921	883		0
Občina Dole 1924	183		0
Občina Idrija 1870	1791		0
Občina Idrija 1880	1289		0
Občina Idrija 1890	1431		0
Občina Idrija 1900	247		0
Občina Idrija 1910	210		0
Občina Idrija 1921	2743		0
Občina Idrija 1931	5407		0
Občina Idrija 1936	5428		0
Občina Zminec 1880	502		0
Občina Zminec 1900	426		0
Občina Zminec 1931	1226		0
Občina Škofja Loka 1869	1133		0
Občina Škofja Loka 1880	793		0
Občina Škofja Loka 1890	640		0
Občina Škofja Loka 1900	300		0
Občina Škofja Loka 1931	951		0
	270913	142689	

Tabela 1: Popisi prebivalstva Slovenije 1830-1931 (število digitaliziranih slik; število transkribiranih oseb; popis je (1) oziroma še ni (0) uvožen v orodje za transkribiranje podatkov).

Na primeru popisa Mesto Ljubljana 1857 se vidi, da imajo lahko nekateri popisi že prepisane osnovne podatke o osebah, čeprav ti popisi še niso bili uvoženi v orodje za transkribiranje podatkov.

Za lažjo iskanje po gradivu so uslužbenci arhiva tekom let za nekatere popise že naredili indekse, v katerih so bili zajeti osnovni podatki o popisanih osebah. Indeksi niso bili izdelani po enotnem podatkovnem modelu, temveč se glede na nabor spremenljivk med seboj lahko precej razlikujejo. V glavnem vsebujejo le ime ime in priimek, leto rojstva in bivališče. Praviloma vsebujejo še podatek o tem, v katerem arhivskem dokumentu oziroma na kateri digitalizirani sliki se nahaja originalni arhivski zapis o indeksirani osebi. Nekateri arhivarji so v izdelavo indeksa vložili še dodaten trud in prepisali nekatere dodatne podatke kot so družinski stan, poklic, družinska razmerja, kraj rojstva in domovinska pravica. Zaradi lažjega iskanja oseb so stare oziroma ponemčene zapise slovenskih priimkov pogosto normalizirali v sodoben slovenski zapis.

Praviloma torej ti indeksi vsebujejo le osnovne podatke o osebah, s pomočjo katerih je nato mogoče najti še dodatne podatke o indeksirani osebi v analognem ali digitaliziranem arhivskemu gradivu. Na ta način so bili zbrani podatki za 15700 oseb popisa Ljubljane 1857 (v urejanju za uvoz v orodje) in za 10400 oseb popisov okrajnega glavarstva Novo mesto 1869 (uvoženo v orodje) ter še za ostale popise okrajnega glavarstva Novo mesto (še nismo dobili).

Na drugi strani spektra zbiranja podatkov iz historičnih popisov prebivalstva pa so raziskovalci in ostali zgodovinarji, ki prepišejo vse podatke iz popisov prebivalstva. Na ta način smo od Muzejskega društva Vrhnika dobili podatke za 6800 oseb, kateri so bili prepisani iz popisov občine Vrhnika 1890 in 1900 (Anžič, 2004).[8]

Indeksi, ki so jih naredili arhivisti, so bili pretvorjeni v XML, primeren za uvoz v relacijsko MySQL bazo orodja za transkribiranje. Te osnovne podatke o osebah se nato v orodju dopolni še z ostalimi, pred tem neprepisanimi podatki.

3 Orodje za transkribiranje

3.1 Osnovni namen

Na Inštitutu za novejšo zgodovino razvito orodje za transkribiranje ima v skladu z različnimi interesi partnerjev (Zgodovinski arhiv Ljubljana) v projektu Popisi prebivalstva dvojen namen:

- Zadovoljiti potrebe ustanov s področja varstva kulturne dediščine po čim lažjemu dostopu širše javnosti (prvenstveno rodoslovcev) do njihovega gradiva.
- Zadovoljiti potrebe raziskovalcev, ki se ukvarjajo s historično demografijo, po čim širšem naboru relevantnih raziskovalnih podatkov.

Zaradi zelo različnih interesov rodoslovnih (prepisuje se osnovne podatke vseh oseb) in raziskovalnih projektov (prepisuje se vse podatke reprezentativnega vzorca oseb) so bila do sedaj razvita orodja za transkribiranje prilagojena potrebam samo ene od teh dveh skupin projektov. Z razvojem novega orodja za transkribiranje smo to razdvojenost učinkovito presegli na način, ki maksimalno koristi obema partnerjema v projektu. Arhivisti tako prispevajo digitalizirano gradivo in podatke o indeksiranih osebah, raziskovalci prepišejo manjkajoče podatke o že indeksiranih osebah, hkrati pa prepisujejo

[8] SI_ZAL_VRH/0016 Matični urad Vrhnika, 1870–1959, Zgodovina Slovenije – Sistory, http://sistory.si/publikacije/?menu=737.

tudi podatke o osebah, ki jih arhivisti še niso indeksirali. Posledično se s tem širi tudi nabor indeksiranih oseb, do katerih preko iskalnika dostopajo uporabniki rodoslovnih projektov.

3.2 Skupine uporabnikov

Orodje za transkribiranje podatkov je prosto dostopno za vse raziskovalce, vendar se morajo zainteresirani raziskovalci (in ostali zainteresirani uporabniki) najprej registrirati.[9] Glede na pravice in dolžnosti se registrirani uporabniki delijo na tri skupine:

- navadni uporabniki,
- napredni uporabniki,
- uredniki popisov.

Z registracijo uporabniki najprej dobijo status navadnega uporabnika. Navadni uporabniki v orodju ne vidijo podatkov o že prepisanih osebah, temveč le tiste podatke, ki so jih sami prepisali. Posledično lahko prepisujejo podatke o osebah samo iz tistih hiš, iz katerih ni prepisoval še nihče drug. Podatke, ki so jih prepisali, lahko izvozijo v XLS datotekah. Šele ko (pravilno) prepišejo podatke za 300 oseb, jim urednik popisov lahko status nadgradi v naprednega uporabnika.

Napredni uporabniki v orodju vidijo ne le podatke, ki so jih sami prepisovali, temveč vse podatke, ki so bile pred tem vneseni v bazo tudi s strani ostalih uporabnikov. Hkrati pridobijo pravico do korigiranja vseh že obstoječih zapisov in pravico do izvoza celotne baze podatkov.

Urednike popisov lahko določi le administrator orodja. Glede dostopa do že prepisanih raziskovalnih podatkov ima enake pravice kot napredni uporabnik. Urednik popisov ima pravico, da ustvari XML datoteko za uvoz novega popisa, da novemu popisu preko administracije določi dodatna polja za prepisovanje in da nadzira pravilnost prepisov navadnega uporabnika.

Velike razlike med pravicami navadnih in naprednih uporabnikov glede prostega dostopa do vseh raziskovalnih podatkov so bile določene iz dveh razlogov:

- Ker uporaba orodja ni omejena le na ožjo projektno skupino, temveč je odprta za vse zainteresirane raziskovalce in študente, smo se z zahtevo, da mora uporabnik najprej pravilno prepisati podatke za 300 oseb, preden dobi pravico do dostopa do celotne baze podatkov, hoteli izogniti problemu prostega strelca (free rider problem).
- Hkrati smo presodili, da so posamezni popisi prebivalstva kot zgodovinski vir lahko tako zelo specifični, da je za čim bolj pravilno interpretacijo spremenljivk potrebno prepisati vsaj nekaj originalnih podatkov.

3.3 Osnovna načela

Relacijska baza podatkov, v katero se preko orodja za transkribiranje prepisuje podatke iz popisov, je bila zgrajena v skladu s sledečimi načeli:

1. Struktura popisov prebivalstva se je tekom časa zelo spreminjala, zato ni mogoče za vse popise vnaprej točno določiti vsa podatkovna polja.
2. Poleg popisov prebivalstva mora orodje omogočiti prepisovanje podatkov še iz matičnih knjig in drugih podobnih tabelarnih historičnih osebnih podatkov (domovnice, vojaške konskripcije ipd.).
3. Vsi popisi imajo skupna samo sledeča osnovna polja, podatke katerih zajema tudi iskalnik: ime in priimek, leto rojstva in širša popisna enota, znotraj katere je bila oseba popisana.
4. Osebe so med seboj povezane v razmerju, pri katerem se na eno izhodiščno osebo veže nič ali več odvisnih oseb.
5. Osebe morajo biti popisane v okviru ene enote, enota pa lahko ima nič ali več podenot, na katero je vezana ena ali več izhodiščnih oseb.
6. Prepisuje se lahko podatke za enote, podenote in osebe.
7. Enote morajo imeti povezavo na eno ali več slik (digitaliziranega arhivskega gradiva), osebe morajo imeti povezavo na eno sliko.
8. Popisi, enote, podenote, osebe in slike imajo unikatne identifikacijske oznake
9. Enote, podenote in osebe ni mogoče brisati.
10. Prepisanih podatkov enot, podenot in oseb ni mogoče brisati. Popravljen prepis je shranjen kot nova verzija. Za vsake shranjeno verzijo prepisanih podatkov se shrani tudi podatek o uporabniku in časovna znamka.

3.4 Izgradnja orodja

V skladu s temi načeli je bilo izdelano orodje za transkribiranje historičnih demografskih podatkov, ki omogoča čim hitrejše prepisovanje iz slik v relacijsko MySQL bazo podatkov. Prva, poskusna verzija orodja (Popisi 1.0) je bila izdelana leta 2011. Ta verzija orodja je bila v glavnem namenjena le dodatni indeksaciji oseb. Na podlagi pridobljenih izkušenj z delom na prvi verziji orodja se je leta 2012 začelo izdelovati novo verzijo (Popisi 2.0), v kateri so bila upoštevana vsa prej našteta načela. Ta verzija je v naslednjih letih doživela le manjše modifikacije. Lastnik kode je Inštitut za novejšo zgodovino, kodo pa so napisali zunanji izvajalci. Orodje je bilo zavestno zgrajeno s pomočjo nekaterih temeljnih tehnologij: PHP, MySQL, JavaScript, CSS in HTML. Ker je veliko programerjev, ki obvlada te tehnologije, je vzdrževanje orodja relativno poceni. V naslednjih letih načrtujemo večje posodobitve (Popisi 3.0), predvsem glede uporabe obstoječih JavaScript knjižnic (mdr. YUI library). Ob tej priložnosti načrtujemo objavo pod odprtokodno licenco.

3.5 Upravljanje orodja

Velikost popisov v orodju ni vnaprej določena, temveč je odvisna od konkretnih potreb uporabnikov in posameznih parcialnih projektov. V praksi sta se pri tem izoblikovali dve pravili:

- posamezen popis v orodju naj ustreza sestavi analognega arhivskega gradiva;
- večje popise (več kot 30000 oseb) naj se razdeli na manjše, bolj obvladljive dele, ki omogočajo hitrejši izvoz podatkov.

Ker vsak popis praviloma vsebuje na stotine slik in enot smo se zavestno določili, da ne potrebujemo uporabniškega vmesnika za shranjevanje slik in ustvarjanje novih enot. Že ob digitalizaciji arhivskega gradiva se uredi osnovne podatke o enotah (minimalno naslov in število enote) in katere slike so vezane na

[9] http://sistory.si/admin.

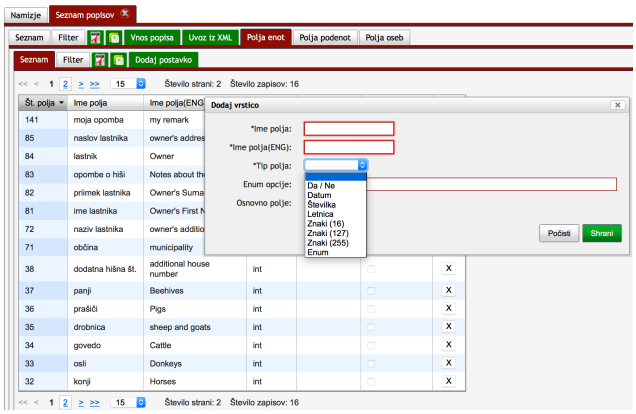
posamezno enoto. Nato pa je te slike in osnovne podatke nujno potrebno množično uvoziti v orodje. Pri popisih prebivalstva so enote posamezne hiše (naslov je ulica, število je hišna številka), znotraj katerih se je popisovalo njihove prebivalce. V skladu z osnovnim načelom 2 pa lahko urednik za enoto določi katero koli poljubno enoto, ki vsebuje podatke o osebah. Ker je iz praktičnih razlogov težje prepisovati podatke iz hiše, ki ima več kot 20 slik, se lahko posamezno hišo razdeli na več smiselnih delov, pri katerih ima sicer vsak del svojo unikatno identifikacijo, vendar imajo vsi deli (enote) iste podatke o hiši. Pri matičnih knjigah ali domovnicah je enota lahko samo ena stran, na kateri se nahaja obrazec o osebi, vse enote pa imajo posledično isti naslov in samo drugo število (stran v matični knjigi).

Glede na načelo 4 mora biti za vsako prepisano osebo določeno:

- oseba je izhodiščna oseba;
- oseba je vezana na točno določeno posamezno izhodiščno osebo.

Če med osebami ni nobene relacije, je vsaka oseba izhodiščna oseba. Pri popisih prebivalstva je izhodiščna oseba t. i. glava družine, pri krstni matični knjigi je izhodiščna oseba krščenec, pri poročni matični knjigi je izhodiščna osebe mož, pri mrliški pa umrli. Na izhodiščno osebo so vezane osebe v točno določenem razmerju (glede na šifrant: žena, mož, sin, hči, mati, oče, brat, sestra) ali v poljubnem razmerju (polje ostala razmerja).

Podenote se uporablja samo v primerih, ko osebe niso bile popisane samo znotraj enote, temveč tudi znotraj njegove podenote. Podenote se v praksi uporablja predvsem za prepisovanje podatkov o stanovanjih (podenota) v posameznih hišah (enota). Ker je bila znotraj ene podenote lahko popisana ena ali več družin, je za vsako podenoto potrebno določiti pripadajoče izhodiščne osebe.

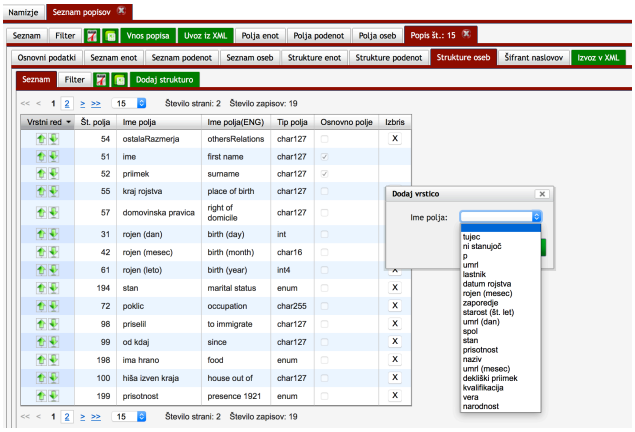


Slika 4: Uporabniški vmesnik za ustvarjenje dodatnih (extra) polj za vnos podatkov.

V skladu z načeli 1, 3 in 6 lahko urednik popisov poleg obveznih osnovnih podatkov za enote (naslov in št.), podenote (izhodiščna oseba) in osebe (ime in priimek, izhodiščna oseba, razmerje) preko uporabniškega vmesnika (Polja enot, Polja podenot, Polja oseb) kreira poljubna dodatna (extra) polja za vnašanje podatkov (glej Sliko 4). Vsako polje mora dobiti ustrezno slovensko in angleško ime. Ker se ime polja izpiše v obrazcu za prepisovanje podatkov, naj bo smiselno in čim krajše. Za vsako polje je potrebno določiti njegov podatkovni tip:

- Booleova spremenljivka (bool; da / ne),
- celo število (int; številka),
- štiri cela števila (int 4; letnica),
- datum (date),
- največ 16 znakov (char 16; znaki 16),
- največ 127 znakov (char 127; znaki 127),
- največ 255 znakov (char 255; znaki 255),
- naštevni (enum).

Urednik popisov nato preko uporabniškega vmesnika (Struktura enot, Struktura podenot, Struktura oseb) za vsak nov popis določi dodatna (extra) popisna polja in njihov vrstni red (glej Sliko 5).



Slika 5: Uporabniški vmesnik za določanje strukture polj posameznega popisa.

Pri tem je priporočljivo, da urednik pred tem skrbno analizira vsebino popisa in šele na podlagi te analize izbere najbolj primerna polja. Če bo npr. za prepisovanje rojstnega datuma izbral tip polja datum, se mora dobro zavedati, da je v to polje poleg letnice obvezno potrebno vpisovati še dan in mesec rojstva. Zlasti skrbno mora izbrati enega od podatkovnih tipov znaki. Če bo izbral tip s premajhnim številom znakov, višek znakov ne bo shranjen v bazo podatkov. Če bo izbral tip s prevelikim številom znakov, bo vizualno preveč razširil obrazec za prepisovanje podatkov in s tem otežil preglednost in hitrost prepisovanja. Polje s podatkovnim tipom znaki 255 je namreč šestkrat daljše od polja s tipom znaki 16.

Pri prepisovanju historičnih demografskih podatkov poznamo tri načine prepisovanja:

1. Uporabnik prepisuje podatke natančno takšne kot so bili prvotno zapisani.
2. Uporabnik prepisuje podatke natančno takšne kot so bili prvotno zapisani, vendar pri tem sproti popravlja očitne napake.
3. Uporabnik pri prepisovanju podatke sproti normalizira na način, ki ustreza njegovi zastavljeni raziskovalni nalogi.

Prvi in drugi način se večinoma uporabljata pri rodoslovnih in dolgoročno naravnanih raziskovalnih projektih, tretji način se večinoma uporablja pri enkratnih raziskovalnih projektih. Pri Popisih prebivalstva Slovenije smo se odločili za prvi način prepisovanja. Z našega stališča ima ta način prepisovanja dve veliki prednosti:

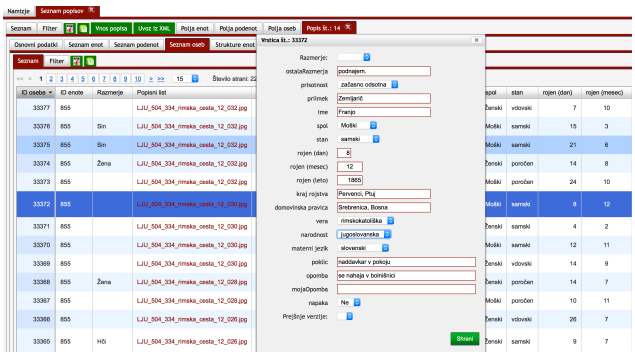
- Ker uporabniku med prepisovanjem ni potrebno izvajati normalizacijo zapisa, poteka prepisovanje sorazmerno veliko hitreje.
- Ker projekt Popisi prebivalstva predvideva ponovno uporabo raziskovalnih podatkov je podatke potrebno prepisati samo enkrat. Prepisane

podatke lahko nato raziskovalci vedno znova drugače interpretirajo in klasificirajo.

Toda v nekaterih primerih se je ta način prepisovanja izkazal za povsem kontraproduktivno, saj je imel za posledico manjšo hitrost prepisovanja in nobene dodatne interpretativne vrednosti. V primeru popisa prebivalstva Ljubljane 1931 je bilo pri opisni spremenljivki stan vrednost poročen zapisana na 53 različnih načinov in vrednost samski na 42 načinov, pri spremenljivki vera je bila vrednost rimskokatoliška zapisana na 46 načinov, pri spremenljivki narodnost pa vrednost jugoslovanska na 45 načinov. Še večja raznolikost načina zapisa istih vrednosti spremenljivke je bila v primeru večjezičnih popisov (slovenski in nemški). Pri popisu prebivalstva Ljubljane 1869 je bilo tako pri opisni spremenljivki stan vrednost poročen zapisana na 134 različnih načinov in vrednost samski na 51 načinov, pri spremenljivki vera pa je bila vrednost rimskokatoliška zapisana na kar 190 načinov. Zato smo sprejeli nova pravilna za način prepisovanja podatkov:

- Opisne spremenljivke, ki vsebujejo veliko število različnih vrednosti (imena, priimki, poklici, kraji ipd.), se prepisuje nespremenjene.
- Opisne spremenljivke, ki vsebujejo malo število različnih vrednosti in kateri so bili že prvotno ustrezno klasificirani (spol, vera, družinski stan, narodnost, jezik ipd.) se pri prepisovanju sproti klasificira na način, ki ustreza prvotni klasifikaciji.

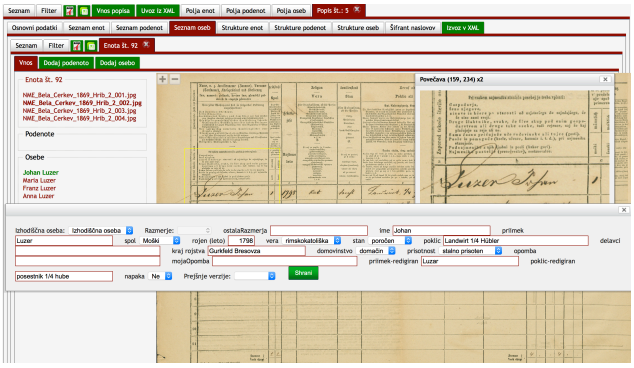
Za slednje primere smo uvedli naštevni (enum) tip podatkov. Urednik popisov predhodno preko vmesnika za ustvarjanje dodatnih polj določi vrednosti (med seboj ločene z vejico), ki jih bo uporabnik preko spustnega menija lahko vnašal v bazo. Ker se v MySQL bazi te vrednosti zapisujejo v varchar tip podatkov, jih lahko urednik popisov vedno naknadno spreminja, dodaja ali briše glede na potrebe projekta.



Slika 6: Urejanje zapisov v seznamu oseb.

Do že prepisanih podatkov o enoti, podenoti ali osebi uporabnik dostopa preko Seznama enot, Seznama podenot in Seznama oseb. Uporabnik lahko ureja in dopolnjuje obstoječi seznam (glej Sliko 6).

Uporabnik začne prepisovati nove podatke preko Seznama enot (gumb Podatki). Podatke prepisuje v obrazec s polji za prepisovanje podatkov, ki lebdi nad digitaliziranо sliko. Obrazec lahko uporabnik prosto premika na način, da si z njim podčrtuje podatke, katere prepisuje. Sliko je mogoče povečati ali pomanjšati, podatke je možno brati tudi s pomočjo povečevalne lupe.



Slika 7: Obrazec s polji za prepisovanje podatkov o osebi.

4 Uvoz in izvoz podatkov

Uporabnik lahko sezname enot, sezname podenot in sezname oseb izvozi v XLS format in izvožene podatke nato izven orodja za transkribiranje podatkov dalje ureja v skladu s svojimi raziskovalnimi potrebami.

Vse podatke posameznega popisa lahko napredni uporabnik izvozi tudi v XML zapisu (glej Sliko 8). Za format XML je izdelana posebna shema, ki se čim bolj dosledno naslanja na relacije v SQL bazi.[10] Podatki enot, podenot in oseb so zapisani v okviru ločenih elementov. Vsak od njih ima svojo identifikacijsko številko (id), verzija shranitve podatkov (version), identifikacijska številka osebe, ki je shranila to verzijo zapisa (user_id_added) in kdaj je bil zapis shranjen (date_added). Podatki enot, podenot in oseb imajo nekatere za njih specifične elemente. Podatek enot ima identifikacijsko številko mesta in številko mesta, podatek podenot ima identifikacijsko številko poglavarja (če jih je več, so med seboj ločene z znakom |), podatek oseb ima identifikacijsko številko razmerja, slike (file) in poglavarja. Vsi pa imajo še dodatna (extra) polja. Vsako dodatno (extra) polje ima v atributu id zapisano identifikacijsko oznako tega polja. Lastnosti dodatnih (extra) polj so na enoten način zapisani v okviru elementov polja_enot, polja_podenot in polja_oseb. V sklopu elementa povezave so zabeležene relacije med identifikacijskimi številkami enot in popisnih listov, enot in oseb ter naposled še enot in podenot. Identifikacijske številke in nazivi popisnih listov (slik), mest (naslov enote) in razmerij (med izhodiščno in odvisno osebo) se nahajajo v sklopu elementa sifranti.

To XML shemo se uporablja tudi za uvoz podatkov v orodje za transkribiranje. Na ta način poteka ne samo uvoz podatkov za slike in enote, temveč tudi za osebe, za katere so arhivarji in ostali zgodovinarji že izdelali indekse. Ti so te podatke prepisovali v XLS datoteke ali celo v tabele in odstavke DOC datotek. Zato je potrebno pred uvozom v orodje za transkribiranje te podatke še dodatno urediti in pretvoriti v XML. Zaradi specifičnosti vsakega popisa je pred vsakim uvozom novih podatkov le-te potrebno pretvoriti z na novo napisanimi XSLT stili.

[10] Imena XML elementov so večinoma izbrana glede na zasnovo prve verzije orodja Popisi 1.0 in jih kasneje nismo več spreminjali. Zato nekatera od njih ne ustrezajo več njihovemu novemu pomenu. Mapiranje: poglavar = izhodiščna oseba, popisni list = slika, mesto = naslov enote, mesto_st = št. enote.

```
<popis>
    <id>3</id>
    <naziv>Ljubljana 1869</naziv>
    <polja_enot>
        <polje>
            <id_extra>84</id_extra>
            <order>3</order>
            <ime_polja>lastnik</ime_polja>
            <tip_polja>char127</tip_polja>
        </polje> ...
    </polja_enot>
    <podatki_enot>
        <podatek>
            <id>1</id>
            <version>1</version>
            <user_id_added>11</user_id_added>
            <date_added>2014-12-11 09:14:30</date_added>
            <mesto>1</mesto>
            <mesto_st>1</mesto_st>
            <extra_polja>
                <polje id="84">Magist. Laibach</polje> ...
            </extra_polja>
        </podatek> ...
    </podatki_enot>
    <polja_podenot>
        <polje> ... </polje> ...
    </polja_podenot>
    <podatki_podenot>
        <podatek> ...
            <poglavar_id>1234|4567</poglavar_id> ...
        </podatek> ...
    </podatki_podenot>
    <polja_oseb>
        <polje> ... </polje> ...
    </polja_oseb>
    <podatki_oseb>
        <podatek> ...
            <razmerje>2</razmerje>
            <file>209</file>
            <poglavar_id>245</poglavar_id> ...
        </podatek>
    </podatki_oseb>
    <povezave>
        <enote_popisni_listi>
            <povezava>
                <enota_id>1</enota_id>
                <file_id>1</file_id>
            </povezava> ...
        </enote_popisni_listi>
        <enote_osebe>
            <povezava>
                <enota_id>1</enota_id>
                <oseba_id>3013</oseba_id>
            </povezava> ...
        </enote_osebe>
        <enote_podenote>
            <povezava>
                <enota_id>6</enota_id>
                <podenota_id>1</podenota_id>
            </povezava> ...
        </enote_podenote>
    </povezave>
    <sifranti>
        <popisni_listi>
            <popisni_list>
                <id>1</id>
                <file>Mesto_1_001.JPG</file>
            </popisni_list> ...
        </popisni_listi>
        <mesta>
            <mesto>
                <id>2</id>
                <naziv>Gradišče</naziv>
            </mesto> ...
        </mesta>
        <razmerja>
            <razmerje>
                <id>2</id>
                <naziv>Mož</naziv>
            </razmerje> ...
        </razmerja>
    </sifranti>
</popis>
```

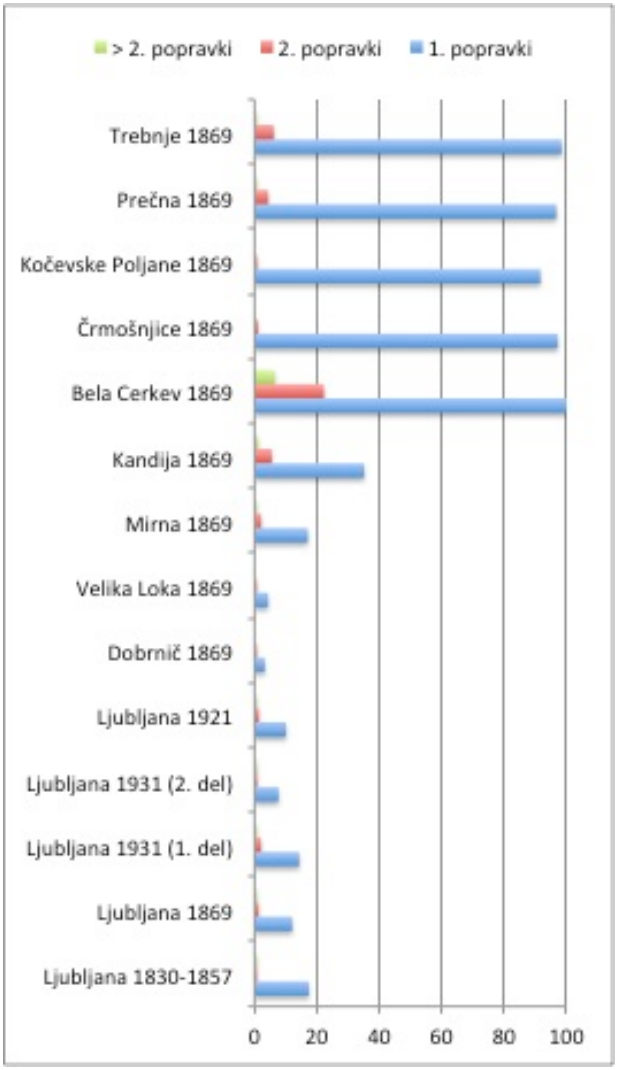
Slika 8: Zapis popisa v XML.

5 Analiza uporabe orodja

Poleg članov raziskovalne infrastrukture Slovenskega zgodovinopisja orodje uporabljajo tudi študentje Filozofske fakultete (in delno FDV) Univerze v Ljubljani. Čeprav projekt Popisi prebivalstva torej ni tipičen 'crowdsourcing' projekt, menim, da bi statistična analiza uporabe tega orodja pomagala ostalim raziskovalcem pri načrtovanju podobnih projektov. Takšnih analiz (Causer in Terras, 2014, 6-9) namreč do sedaj ni bilo veliko predstavljenih. Za analizo sem uporabil v XML izvožene podatke. Za razliko od XLS izvoza, ki vključuje le zadnjo verzijo zapisa podatkov, se v XML izvozi celotna baza podatkov: vse verzije zapisov podatkov ter kdo in kdaj jih je shranil. Z analizo bom zajel samo podatke o osebah, saj so uporabniki do sedaj le redko dopolnjevali podatke o enotah in podenotah. Ker smo v bazi dodali časovno znamko šele marca 2015, bom časovno analizo moral omejiti le na obdobje po tem datumu.

Za vsakogar, ki načrtuje raziskovalni projekt, pri katerem je nujno potrebno prepisati podatke o osebah, je najbolj zanimiv podatek o povprečni hitrosti prepisovanja. Pri tem se je potrebno zavedati, da bo ta hitrost odvisna tudi od jezika in vrste pisave, s katero je zapisan originalni popis prebivalstva. Popis prebivalstva Ljubljane 1869 (18 spremenljivk) je večinoma zapisan v nemščini in v nemški kurenti. Zato pri njem znaša srednja vrednost (mediana) vpisovanja novih oseb v bazo 1 minuta in 56 sekund. Nasprotno znaša mediana pri popisi prebivalstva Ljubljane 1931 (20 spremenljivk, slovenščina, latinica) samo 1 minuta in 15 sekund. Pri tem pa lahko hitro opazimo precejšnje razlike med začetnimi in izkušenimi uporabniki. Medtem, ko začetniki potrebujejo za vpis ene osebe 1 minuto in 44 sekund, jo izkušeni uporabniki prepišejo v 1 minuti in 13 sekundah. Zelo velike razlike so tudi med samimi izkušenimi uporabniki. Ena uporabnica je tako v povprečju za vpis ene oseba porabila samo 36 sekund.

Velik vpliv na končno hitrost prepisovanja podatkov imajo nujni popravki podatkov. V grafikonu na sliki 9 je prikazano, koliko odstotkom oseb so uporabniki naknadno popravljali podatke. Pri tem je tudi prikazano, koliko odstotkom so bili podatki popravljeni enkrat, koliko dvakrat in koliko več kot dvakrat. Pri tem je potrebno razlikovati med popisi, za katere so bili v orodje uvoženi tudi indeksi (Trebnje, Prečna, Kočevske Poljane, Črmošnjice in Bela Cerkev) in med ostalimi popisi. Pri prvi skupini popisov je bilo pri skoraj vsaki osebi potrebno osnovne podatke iz indeksov najprej dopolniti še z ostalimi podatki. Pri ostalih popisih pa so uporabniki praviloma že prvič prepisali vse podatke o osebah. Pri popisih Ljubljana 1830, 1868 in 1931 (1. del), kateri so bili najprej uvoženi v orodje za transkribiranje, je nekoliko večje število prvih popravkov tudi posledica razvoja orodja (iz verzije 1.0 na 2.0) in ustvarjanja novih podatkovnih polj. Hkrati moramo tudi vedeti, da sta popisa Bele Cerkve in Kandije edina popisa, ki sta bila naknadno tako temeljito dvojno pregledana, da v prihodnosti praktično ne bosta več potrebovala popravkov.



Slika 9: Odstotki popravljenih podatkov oseb glede na popis.

Če upoštevamo te dejavnike, lahko pri analizi grafikona pridemo do sledečih zaključkov:

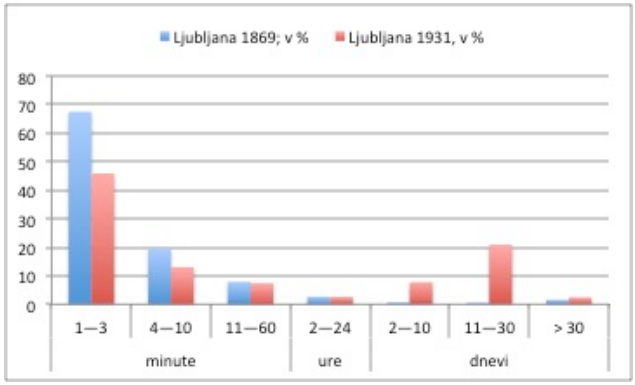
- Pri prepisovanju podatkov o osebah lahko pričakujemo, da bodo uporabniki morali popravljati podatke pri manj kot 10 % osebah.
- Ta odstotek popravkov je večji v primerih, ko mora uporabnik brati različne rokopise (ljubljanski popisi so zaradi različnih rokopisov veliko težje berljivi kot popisi okrajnega glavarstva Novo mesto).
- Preden bo osnovni prepis popisa postal uporaben za raziskovalne namene, ga bo uporabnik moral še enkrat skrbno pregledati in razrešiti vse morebitne dileme glede pravilnosti prepisa. Pri tem lahko pričakuje, da bo moral popraviti podatke za dosti več kot petino oseb.

V spodnji tabeli 3 je prikazana analiza načina popravljana besedila. Pri tem sem popravke razvrstil v tri večje skupine: besedilo je bilo delno popravljeno, v prej prazno polje so bili vpisani podatki, besedilo je bilo v celoti izbrisano. Dobljeni rezultati analize so lahko pri različnih popisih povsem različnih. Te razlike so posledice samo enega dejavnika: v primeru, da uporabnik ni prepisal vseh podatkov o osebi, je veliko večja verjetnost, da bo te podatke dopolnila druga oseba.

	popravljeno besedilo	vpisano besedilo v prazno polje	povsem izbrisano besedilo	popravek naredila druga oseba
Bela Cerkev 1869	5,1	94,9	0,0	98,3
Ljubljana 1931 (1)	7,8	91,4	0,9	87,8
Ljubljana 1869	45,6	48,6	5,7	31,8
Ljubljana 1931 (2)	41,2	28,8	30,0	18,4

Tabela 3: Načini popravljanja besedila glede na izbrane popise; v %.

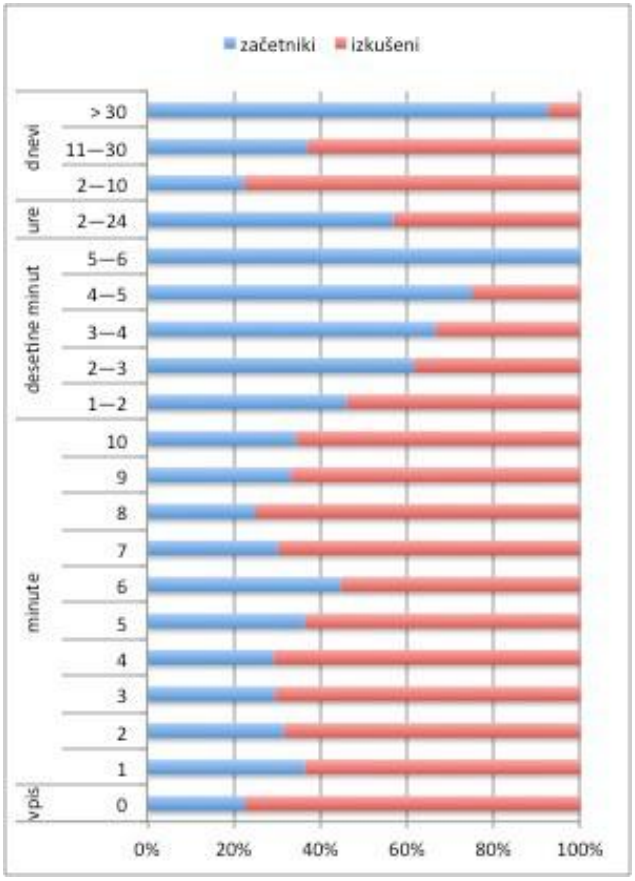
Iz teh podatkov je tudi razvidno, da uporabniki pogosto sproti popravljajo svoje prepisane podatke. Da bi lahko bolje razumeli to prakso, sem pripravil še časovno analizo popravkov. Na žalost pri tem razpolagamo samo s podatki od marca 2015, zaradi česar ti podatki niso primerni za časovno analizo popravkov, ki so se zgodili po daljšem časovnem obdobju.



Slika10: Koliko časa je preteklo od shranitve izvorne in popravljene verzije podatkov; popisi prebivalstva in časovne skupine v %.

Iz zgornjega grafikona (slika 10) je tako jasno razvidno, da lahko prakso popravljanja zapisanih podatkov razdelimo na dve večji skupini. Naknadno popravljanje in dopolnjevanje besedil po preteku 10 in več dni, ter na sprotno popravljanje napačnih zapisov. V slednjem primeru je tako velika večina popravkov opravljena že v prvih treh minutah, večinoma kar v prvi minuti.

Pri tem so zanimive precejšnje razlike med izkušenimi uporabniki in začetniki (slika 11). Glede na število opravljenih vpisov namreč začetniki pogosteje popravljajo (svoje) vpisane podatke kot pa izkušeni uporabniki. Sorazmerno največ je teh popravkov zlasti po prvih desetih minutah. Razlog za to tendenco je lahko samo eden. Zaradi svoje neizkušenosti se začetni uporabniki pri prepisovanju podatkov relativno bolj pogosto kot izkušeni srečajo z dilemo kako prebrati ali kako zapisati kakšen podatek. Pri tem reševanju te dileme namenijo dosti več časa kot njihovi izkušeni kolegi. V primeru morebitne večje vključenosti študentov v bodoče projekte bo vsekakor potrebno večjo pozornost nameniti reševanju teh problemov (npr. intenzivnejše in dolgotrajnejše delavnice).



Slika 11: Razlike med izkušenimi in začetnimi uporabniki glede popravljanja podatkov.

6 Zaključek

V prispevku sem predstavil orodje za transkribiranje historičnih demografskih podatkov, ki ga na Inštitutu za novejšo zgodovino uporabljamo v sodelovanju z arhivi, društvi in univerzo. Ker bi si želeli, da bi k uporabi orodja pritegnili še nove uporabnike, nameravamo temu primerno orodje razvijati tudi v prihodnje. Tako si bodo parcialni raziskovalni projekti v bodoče lahko rezervirali želeni sklop digitaliziranega gradiva (kateri še ni uvožen v orodje) in dobili zanj izključno pravico za obdelavo. Šele po preteku projekta bodo ti podatki v skladu s politiko odprtega dostopa do raziskovalnih podatkov dani na razpolago tudi ostalim raziskovalcem.

7 Literatura

Sonja Anžič. 2004. Prebivalstvo občine Vrhnika na prelomu 19. in 20. stoletja. *Vrhniški razgledi*, 5: 95-100. http://www.dlib.si/details/URN:NBN:SI:spr-LQTKRUDL.

Tim Causer in Melissa Terras. 2014. Crowdsourcing Bentham: beyond the traditional boundaries of academic history. *International Journal of Humanities and Arts Computing*, 8(1): 46-64. http://dx.doi.org/10.3366/ijhac.2014.0119.

Alicia Fornés, Josep Lladós, Joan Mas, Joana Maria Pujades in Anna Cabré. 2014. A Bimodal Crowdsourcing Platform for Demographic Historical Manuscript. V: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, str. 103-108, New York, NY. DOI: 10.1145/2595188.2595199.

Aaron G. Noll. 2013. Crowdsourcing Transcription of Archival Materials. V: *Graduate History Conference: Interdisciplinary Approaches to Historical Inquiry*, Boston, MA. http://scholarworks.umb.edu/ghc/2013/panel6/4/.

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. http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2015-4-4.

Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI

Andrej Pančur*

* Inštitut za novejšo zgodovino
Kongresni trg 1, 1000 Ljubljana
andrej.pancur@inz.si

Povzetek

Inštitut za novejšo zgodovino sodeluje z Državnim zborom RS pri digitalizaciji sejnih zapisov. Besedila iz zbirke sejnih zapiskov zasedanj Skupščine (socialistične) republike Slovenije so v okviru raziskovalne infrastrukture Slovensko zgodovinopisje kodirane po smernicah Text Encoding Initiative (TEI). Avtor se jim v prispevku posveča s posebno pozornostjo, saj so v digitalni humanistiki *de facto* standard za kodiranje tekstovnih besedil. Podatki, ki jih je mogoče na ta način pridobiti, pa omogočajo odgovore na številna raziskovalna vprašanja.

Encoding the Slovenian Parliament Session Minutes in Line with the TEI Guidelines

The Institute of Contemporary History cooperates with the National Assembly of the Republic of Slovenia in the process of digitising its session minutes. In the context of the Research Infrastructure of Slovenian Historiography, the texts from the collection of session minutes from the (Socialist) Assembly of the Republic of Slovenia are encoded in accordance with the Text Encoding Initiative (TEI) guidelines. In his contribution, the author pays special attention to these guidelines, as they are the *de facto* standard of text encoding in digital humanities. The information that can be acquired in this manner provides answers to many research questions.

1 Uvod

V zadnjih letih se v demokratičnih državah na spletu vedno pogosteje javno objavljajo podatki (open data), ki so jih ustvarile različne državne in javne službe. Dolgo, tudi sto in večletno tradicijo javnega objavljanja podatkov o svojem poslovanju, imajo zlasti različne parlamentarne ustanove. Splošna javnost, mediji in raziskovalci so vedno kazali velik interes predvsem za zapisnike sej različnih parlamentarnih teles. To gradivo uporabljajo raziskovalci z različnih področij: zgodovinarji, sociologi, politologi, komunikologi, jezikoslovci, psihologi ipd.

Prvotno je bilo to gradivo dano na voljo javnosti v analogni obliki (večinoma kot tiskane knjige), v zadnjem času pa je vedno pogosteje objavljeno na spletu kot dokumente PDF, HTML in XHTML. Obenem vedno več organizacij poudarja prednosti odprtih formatov XML (Global Centre for ICT, 2014).

XML se je kot zelo primeren format uveljavil tudi v različnih raziskovalnih projektih, v katerih so obdelovali to gradivo. V formatu XML so tako mdr. dostopna zasedanja britanskega parlamenta (Hansard) od leta 1803,[1] nizozemskega od leta 1803 (Marx in Schuth, 2010), španskega od leta 1977 (Martin-Dancausa in Marx, 2010), češkega od leta 1993 (Jakubíček in Kovář, 2010), poljskega od leta 1993 (Ogrodniczuk, 2012) in bolgarskega (v okviru korpusa političnih govorov) od leta 2006 (Osenova in Simov, 2012).

2 Zapisniki sej zakonodajnih teles v Sloveniji

Javnosti in raziskovalcem je na voljo tudi gradivo parlamentarnih ustanov z ozemlja današnje Slovenije oziroma parlamentarnih ustanov, člani katerih so bili tudi poslanci iz Slovenije. Veliko gradiva je sicer še vedno dostopna le v analogni obliki, vedno večji del pa je tudi že digitaliziran in dan na voljo javnosti na različnih portalih:

- ALEX, Historische Rechts- und Gesetzestexte Onilne: avstrijski državni zbor (1861–1918);[2]
- Landtag Steiermark: štajerski deželni zbor (1848–1914);[3]
- Zgodovina Slovenije – SIstory:
  - kranjski deželni zbor 1861–1869;[4]
  - jugoslovanska zakonodajna telesa 1919–1939, 1942–1953;[5]
  - Ljudska skupščina Ljudske republike Slovenije (1947–1963);[6]
  - Skupščina Socialistične republike Slovenije (1963–1990);[7]
- Državni zbor Republike Slovenije od leta 1990 do danes.[8]

Razen slednjih, ki so objavljeni v formatu HTML, so vsi ostali obljavljeni kot PDF.

Na sliki 1 je glede na posamezen sklic Skupščine oziroma mandat Državnega zbora prikazano število besed govorov, ki se nahajajo v PDF publikacijah (zasedanja skupščine) na portalu SIstory (skupaj 36 milijonov besed) in kot HTML na spletni strani Državnega zbora (skupaj 69 milijonov besed). Do leta 1974 so sejni zapiski vsebovali še obsežne priloge (skupaj 10,5 milijonov besed), katere so kasneje začeli objavljati v posebni publikaciji (Poročevalec). Le-ti so po letu 2006 dostopni na spletni strani Državnega zbora. Raziskovalna infrastruktura Slovenskega zgodovinopisja, ki upravlja portal SIstory, se je v sodelovanju z Državnim zborom odločila v naslednjih

[1] Hansard archive (digitised debates from 1803), http://www.hansard-archive.parliament.uk/.

[2] http://alex.onb.ac.at/sachlichegliederung.htm.

[3] http://www.landesarchiv.steiermark.at/cms/ziel/111284715.

[4] http://sistory.si/publikacije/?menu=719.

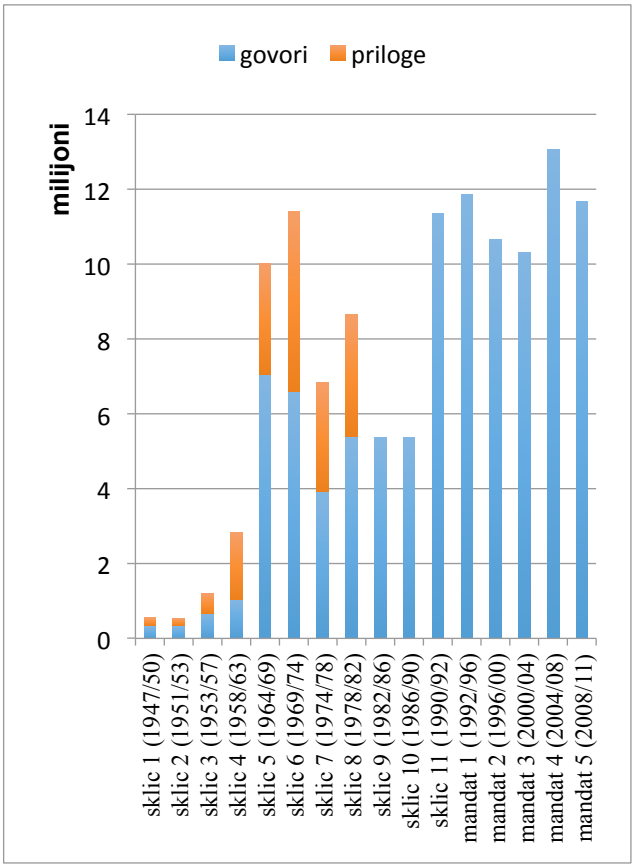
[5] http://sistory.si/publikacije/?menu=396

[6] http://sistory.si/publikacije/?menu=407.

[7] http://sistory.si/publikacije/?menu=408.

[8] Republika Slovenija, Državni zbor, https://www.dz-rs.si/wps/portal/Home/deloDZ/seje/sejeDrzavnegaZbora/PoDatumuSeje/.

treh letih digitalizirati še vse manjkajoče Poročevalce. Trenutno so digitalizirani za leta 1974/82 (skupaj 6,2 milijona besed). Če poleg zapisnikov sej, prilog in poročevalcev upoštevamo še seje različnih delovnih teles ter morebitno ostalo gradivo, pridemo do tako velikih količin besedila, ki jih noben raziskovalec ne more obdelati na klasičen način – z branjem. Zato je v teh primerih nujna strojna obdelava vsebine.



Slika 1: Število besed parlamentarnih govorov v sejnih zapiskih, objavljenih na portalu SIstory (1947–1990) in na spletni strani Državnega zbora (1990–2011).

Za nadaljnjo uporabo v raziskovalne namene se je do sedaj uporabljalo samo zapisnike sej državnega zbora, ki so v HTML formatu. Večje količine tega gradiva so uporabili v raziskovalnem projektu SloParl. Pisni del korpusa, ki je nastal v okviru tega projekta tako vsebuje 23 milijonov besed iz obdobja 1996–2005 (Žgank et al., 2006). To gradivo uspešno uporabljajo tudi različne iniciative, ki se zaradi učinkovitejšega nadzora državljanov nad delom Državnega zbora, zavzemajo za čim lažji dostop do tega gradiva.[9] Za razliko od zgoraj naštetih tujih raziskovalnih projektov pa nobeden od dosedanjih slovenskih projektov pri označevanju zapisnikov sej ni uporabil XML.

3 Označevanje in kodiranje v XML

Označevalne sheme teh korpusov se ponavadi med seboj razlikujejo. Največja korpusa (britanski in nizozemski) uporabljata lastno XML shemo, ki je prilagojena različni strukturi besedil obeh korpusov. Španski korpus je shemo prevzel po nizozemskem. Češki korpus je bil kodiran v skladu s potrebami jezikoslovnih raziskav. Delno bolgarski in predvsem poljski korpus pa sta uporabila Smernice Text Encoding Initiative (TEI) (TEI Consortium, 2015).[10]

Smernice TEI so predvsem v digitalni humanistiki *de facto* standard za kodiranje tekstovnih besedil. Zato smo te smernice uporabili tudi pri označevanju zbirke sejnih zapiskov zasedanj Skupščine (socialistične) republike Slovenije, ki jih izvajamo v okviru raziskovalne infrastrukture Slovensko zgodovinopisje na Inštitutu za novejšo zgodovino.

3.1 Vsebinska struktura zapisnikov sej

Pri pretvorbi zapisnikov sej iz formatov PDF in HTML v XML je potrebno izluščiti vsebino strukture besedila. Sejni zapisniki imajo namreč povsem enotno standardizirano strukturo besedila, kar omogoča avtomatično prepoznavanje sledeče strukture: dokument → seje → govori (govorci) → odstavki.

Na začetku vsake seje so zapisani podatki o vrsti seje, številki seje, datumu seje, predsedujočemu seje in o času začetka seje. Govori so med seboj praviloma ločeni z nekoliko večjim razmikom med zadnjim odstavkom predhodnega in prvim odstavkov naslednjega govora. Hkrati se prvi odstavek govora vedno začne z navedbo imena in priimka govornika (skupaj z morebitnimi dodatnimi informacijami o govorniku), ki je od besedila govora ločena z dvopičjem. Kljub tako jasni strukturi, pa je potrebno upoštevati še nekatere odklone. Do leta 1984 so bila npr. imena in priimki govornikov pisana z malimi, kasneje z velikimi črkami. Največ težav pa pri avtomatični pretvorbi povzroča neenoten razmik med govori. Lahko se zgodi, da razmika med govori sploh ni. Obenem se lahko zgodi, da se razmik nahaja pred začetkom novega vsebinskega sklopa. Hkrati so tudi naslovi vsebinskih sklopov lahko zapisani v veliki črkami.

Iz strukture besedila je mogoče pridobiti tudi podatke o časovnem poteku seje, številu prisotnih članov (kvorum), izidu glasovanj ter opise različnih dejanj med govori (ploskanje, vzkliki, govorjenje iz klopi ipd.). Ti podatki so kot komentarji zapisnikov magnetogramov sej ponavadi navedeni v oklepajih. Namesto oklepajev so lahko uporabljali tudi poševno črto. Hkrati je potrebno zelo paziti, da ne bomo vseh besedil v oklepajih avtomatično označili kot komentarje, saj so bili v oklepajih lahko zapisani tudi deli govorov (predvsem v okviru daljših, strokovnih poročil).

Zaradi vseh teh nedoslednosti ni nujno, da bo avtomatično označevanje strukture besedila povsem pravilno, kar je potrebno upoštevati pri naslednjih korakih označevanja besedila v XML.

[9] V okviru Kiberpipe je potekal projekt *Delajo zate!*, s katerim so želeli delo državnega zbora narediti bolj transparentno. V primerjavi s spletno stranjo Državnega zbora je na spletni strani http://www.delajozate.si/ objavljeno gradivo mogoče še dodatno filtrirati glede na poslance. Podobnim ciljem sledi tudi projekt Parlamentaria, ki je trenutno v fazi izdelave. Gl. Parlameter, https://parlameter.si/.

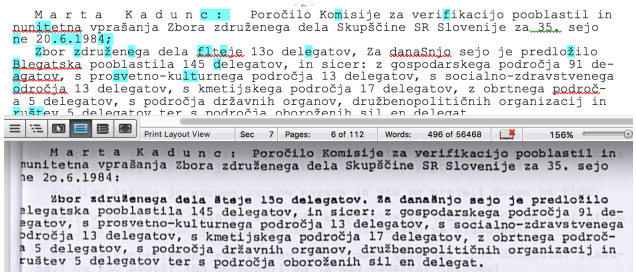
[10] Pri bolgarskem korpusu so TEI uporabili samo za označevanje dokumentov in odstavkov, drugače pa so uporabili Text Corpus Format (TCL). Pri izdelavi poljskega korpusa so govore označili v skladu z modulom za transkribiranje govorov.

3.2 Pretvorba iz PDF in HTML v osnovni TEI

Zaradi deloma zelo različne kvalitete OCR zapisa poteka pretvorba v XML na tri načine, ki zahtevajo različno stopnjo dodatnega urejanja:

1. PDF → DOCX → TEI,
2. PDF → XML (Adobe Acrobat) → TEI,
3. HTML → TEI.

Glede količine vloženega dela je najbolj zahteven prvi postopek. Ta postopek se najbolje obnese v primerih, ko z digitalizacijo ni bilo mogoče zajeti celotnega besedila. Državni zbor namreč hrani analogne zapisnike sej 1984-1990 kot tipkopis samo v enem izvodu. Pri vezavi tipkopisov je bil lahko levi rob besedila tako ozek, da je pri digitalizaciji odrezalo črko ali dve. Obenem so lahko bile črke na začetku vrstice delno rezane. Posledično so zato v OCR zapisu zelo pogoste napake. Na sliki 2 je spodaj prikazan izsek iz dokumenta PDF in zgoraj adekvatni izsek iz dokumenta DOCX, pri katerih so jasno razvidne napake, ki so nastale zaradi manjkajočega dela levega robu besedila. Po naših izkušnjah je takšne napake najlažje popravljati tako, da se PDF pretvori v DOCX z ohranjeno postavitvijo strani. Pri ročnem popravljanju napak se nato dopolnjuje manjkajoče črke na levem robu in obenem še popravlja morebitne druge napake. Za pretvorbo DOCX v TEI uporabljamo zadnjo verzijo XSLT (TEI) stilov, za katere je Tomaž Erjavec napisal dodaten profil. Pred pretvorbo je potrebno govore označiti v Wordu s slogom tei:sp.[11]



Slika 2: Rezan levi rob digitaliziranega besedila. Primerjava med PDF spodaj in DOCX zgoraj.

Veliko manj napora zahteva pretvorba zapiskov sej iz let 1947/84. Ker so bili ti zapiski tiskani v knjižnih izdajah, se je do danes ohranilo več kopij tega gradiva. Državni zbor nam je zato lahko odstopil odvečne dvojnike, ki so bili med digitalizacijo razrezani. Posledično je tudi OCR zapis teh zapisnikov skoraj brez napak. Najbolj pogoste napake kodiranja znakov se popravi po pretvorbi v XML s XSLT stilom, ki vsebuje adekvatne regularne izraze (npr. na začetku stavkov *Zeli* = *Želi*, *Ce* = *Če*; besede s številko *2* namesto z *Ž*). Vsebinsko sporne primere (npr. namesto *m* pravilno *in* ali *ni*) se popravi z iskalnikom najdi-zamenjaj, ki vsebuje regularne izraze.

Sprva se je v teh primerih za pretvorbo iz PDF v TEI kot vmesni XML zapis uporabljala pretvorba iz PDF v Adobe Acrobat XML. Vendar se je kmalu kot bolj enostaven in zanesljivejši izkazal postopek, pri katerem se z ABBYY FineReader že ob digitalizaciji poleg PDF datoteke hkrati naredil še HTML. Tako nastali HTML ima dve pomembni prednosti pred ABBYY FineReader XML:

- Ker so v ABBYY FineReader XML shranjene informacije celo o posameznih znakih, so HTML datoteke veliko manjše in preglednejše, hkrati pa še vedno dokaj verno odražajo postavitev originalnih strani.
- V primeru morebitnih kasnejših ročnih popravkov je HTML datoteka zelo pregledno in lahko berljivo referenčno besedilo.

HTML se pretvori v osnovni TEI z različnimi XSLT stili, napisanimi posebej za ta projekt. Nekoliko drugačne verzije teh stilov se uporablja tudi za pretvorbo HTML zapisnikov sej iz spletnih strani Državnega zbora. Besedila sejnih zapisnikov smo pridobili z luščenjem podatkov (Beautiful Soup)[12], v primeru širše zastavljenega projekta pa jih raziskovalci lahko pridobijo neposredno od Državnega zbora. Ker ti zapisniki sej nimajo napak v kodiranju znakov, je njihova pretvorba sorazmerno najhitrejša. To velja zlasti za zapisnike sej po letu 1996, katere se je sproti objavljajo v elektronski obliki.

3.3 Kazala vsebine in seznami govornikov

Sejni zapisniki vse do leta 1996 vsebujejo tudi kazala vsebine in sezname govornikov. Kazala vsebine kasneje zamenjajo dnevni redi, z uvedbo elektronskega glasovanja pa namesto seznama govornikov veliko večji pomen pridobijo kvorumi glasovanja.

Zlasti kazala vsebine prinašajo dodatne informacije (dnevni red in potek seje), katere iz same strukture besedila govorov niso jasno razvidne. Prehod med enim in drugim vsebinskim sklopom (točko dnevnega reda) vedno najavi predsedujoči. Če točki dnevnega reda ne sledi razprava, lahko znotraj istega govora najavi novo točko itn. Če se v vsebinsko razpravo vmešajo ostali govorniki, predsedujoči v enem od svojih naslednjih govorov zaključi razpravo in nato znotraj istega govora najavi novo vsebinsko točko. Predsedujoči lahko razpravo tudi prekine, začne znova ali zaključi. Ti vsebinski sklopi so v najboljšem primeru lahko znotraj govora predsedujočega označeni s poudarjenim naslovom (znotraj odstavka ali med dvema odstavkoma). Avtomatsko označevanje strukture teh vsebinskih sklopov je zato podvrženo pogostim napakam, zaradi česar je nujno potrebno opraviti ročno korekcijo vseh zapisnikov.

Dodatno vsebinsko vrednost prinašajo še seznami govornikov. Po naših izkušnjah je njihova glavna vrednost v bolj preglednem in pravilnejšem označevanju govornikov. Zaradi relativno pogostih napak in drugih nedoslednosti v zapisovanju imen in priimkov govornikov, je avtomatska dvojna kontrola imen in priimkov (imena v seznamih se morajo ujemati z imeni govornikov v govorih) zelo koristna, saj opozori na morebitne sporne primere. Ker smo na podoben način kot seznam govornikov označili še seznam predsedujočih posamezne seje, lahko pri analizah tako označenih govorov primerjamo (pogosto pomembne) razlike med govori predsedujočih in ostalih govornikov.

[11] DOCX to TEI to HTML Conversion, http://nl.ijs.si/tei/convert/. Kot JSI profil je vključena tudi v zadnje verzije oXygen XML urejevalnika. Pri nas uporabljamo nekoliko prilagojeno verzijo JSI profila.

[12] Beautiful Soup, https://www.crummy.com/software/BeautifulSoup/.

3.4 Text Encoding Initiative (TEI)

Pri kodiranju sejnih zapisnikov v TEI smo najprej nameravali uporabiti modul za transkribiranje govorov. Toda pravkar opisana struktura zapisnikov sej je v resnici zelo podobna elementom dramskih besedil: scena, govori in didaskalije (stage-direction) (Marx, 2009, 3). Zato smo raje uporabili TEI modul za dramska besedila.

Zapisniki sej so bili objavljani v različnih publikacijah (monografije s prilogami ali brez njih, vezani tipkopisi, spletne strani) z različnim obsegom vsebine. Razlike med temi publikacijami smo morali upoštevati tudi pri osnovni strukturi delitve besedila v dokumentu TEI. Ker smo pri tem upoštevali tudi strukturo zapisov sej starejših izvršnih in zakonodajnih teles na lokalnih (dežele, republike) in širših državnih ravneh (Jugoslavija, Habsburška monarhija), bo ta poenotena struktura primerna tudi za njih.

Vsak TEI dokument ustreza eni enoti izvornega gradiva. Zato lahko nekateri TEI dokumenti vsebujejo več sej različnih zborov, drugi pa samo eno sejo enega dne. Metapodatki v teiHeader so narejeni avtomatično pri pretvorbah z različnimi XSLT stili. Vedno vsebujejo titleStmt (naslov dokumenta title, podatke o ustvarjalcih dokumenta respStmt), publicationStmt s krajem objave in avtorsko pravico (licenca Creative Commons Priznanje avtorstva 4.0), čim bolj natančne podatke o izvirnem besedilu in njegovih avtorskih pravicah (sourceDesc) ter ob vsaki novi pretvorbi v revisionDesc še natančne podatke o opravljenih dodatnih kodiranjih.

```
<text>
   <front>
      <!-- možen titlePage, docImprint -->
      <div type="contents">
         <!-- kazalo vsebine; lahko tudi v back -->
      </div>
      <div>
         <!-- seznam govornikov; lahko tudi v back -->
      </div>
   </front>
   <body>
      <!-- govori -->
   </body>
   <back>
      <div type="appendix">
         <!-- priloge -->
      </div>
      <div>
         <!-- seznam govornikov; lahko tudi v front -->
      </div>
      <div type="contents">
         <!-- kazalo vsebine; lahko tudi v front -->
      </div>
      <!-- možen div[@type='colophon'] -->
   </back>
</text>
```

Slika 3: Osnovna delitev besedila zapisnikov sej v front, body in back.

Znotraj telesa besedila body so lahko kodirani samo govori (glej sliko 3). Ker se kazalo vsebine in seznam govornikov lahko nahajata pred ali za govori jih lahko kodiramo znotraj uvodnega razdelka front ali znotraj zaključnega razdelka back. Kazalo vsebine se mora nahajati v okviru razdelka div, ki ima atribut type z vrednostjo contents. Kazalo je kodirano kot seznam list. Sezname govornikov je potrebno zapisati v sklopu posebnega div razdelka, kateremu ni potrebno dajati atributa type, saj morajo biti seznami kodirani v okviru elementa castList (seznam nastopajočih). Vse morebitne priloge so kodirane v back v posebnem div razdelku z vrednostjo atributa type appendix.[13]

Podatki o sejah so kodirani v telesu besedila body. Ker so se sestave parlamentov tekom zgodovine lahko precej spreminjale, je te spremembe potrebno upoštevati tudi pri kodiranju v TEI. Današnji slovenski parlament (po letu 1992) se tako npr. deli na Državni zbor in na Državni svet. Skupščina (socialistične) republike Slovenije pa je med leti 1963–1992 v različnih kombinacijah sestavljalo kar 11 različnih zborov in 6 različnih (samoupravnih skupščin).

```
<body><!-- govori -->
   <div><!-- vrsta zbora oz. skupščine -->
      <div><!-- seja oz. zborovanje -->
         <docDate><date when="1990-07-02">dan zasedanja</date>,
            <date>možen drugi dan zasedanja</date></docDate>
         <timeline>
            <!-- navezava na tei:stage[@type='time'] -->
         </timeline>
         <castList>
            <!-- predsedujoči seje -->
         </castList>
         <stage type="time">
            <!-- čas začetka govorov -->
         </stage>
         <div><!-- pred dnevnim redom -->
            <div>
               <!-- vsebinski sklopi -->
            </div>
         </div>
         <div><!-- dnevni red -->
            <div>
               <!-- vsebinski sklopi oz. točke -->
            </div>
         </div>
      </div>
   </div>
</body>
```

Slika 4: Kodiranje podatkov o sejah.

Razdelek div znotraj telesa besedila body zato vsebuje informacijo o vrsti zbora oz. skupščine (glej sliko 4). Naslovi so kodirani v elementu head. Če je bila kot vir uporabljena knjiga, je teh razdelkov lahko več. Ta razdelek div nujno vsebuje nov razdelek div s podatki o seji oz. zborovanju. Z atributom n je označena številka seje. Vsaka seja vsebuje podatek o dnevu seje (kodirano v okviru datuma dokumenta docDate). Ker so seje lahko trajale več dni, pri čemer je med posameznimi dnevi zasedanja lahko minilo precej časa, je vsak datum kodiran kot date/@when. Vedno je zapisan tudi podatek o začetku seje. Ta podatek je zapisan v elementu didaskalija stage (več o tem glej spodaj). Pred začetkom govorov so vedno navedeni tudi predsedujoči (ena ali več oseb). Podatke o teh osebah se kodira v skladu s seznamom govornikov (glej sliko 5). Vsak govornik ima unikatni identifikator @xml:id, ki je avtomatično skonstruiran iz njegovega imena in priimka. Če je predsedujoči naveden tudi v seznamu govornikov, dobi atribut sameAs, ki ga navezuje na seznam predsedujočih.

Posamezne seje so nato označene v skladu s kazalom oziroma dnevnim redom (glej sliko 4). Ta se ponavadi deli na dva večja razdelka: Pred dnevnim redom in Dnevni red. Prvi div razdelek je opcijski, drugi je nujni. Vsak od teh razdelkov mora nujno vsebovati enega ali več

[13] Kodiranje različnih prilog še nismo povsem poenotili, zato v tem prispevku tega ne predstavljam. Pri tem načrtujemo, da bomo to lahko dokončno storili šele v sklopu poskusnega kodiranja Poročevalca.

razdelkov, ki zajemajo posamezne vsebinske sklope. Ti vsebinski sklopi so ponavadi usklajeni z originalnim kazalom vsebine, katere so izdelale osebe, zadolžene za izdelavo magnetogramov. Oseba, ki izvaja kodiranje v TEI, lahko vsebino govorov označi drugače iz dveh razlogov:

- če presodijo, da določena vsebina ni bila označena v originalnem kazalu;
- če presodijo, da so podpostavke v kazalu preveč nadrobne, zaradi česar označijo samo skupno postavko.

```
<front>
   <!-- ... -->
   <div>
      <!-- seznam govornikov -->
      <castList>
         <castItem>
            <actor xml:id="sp.DolinšekDrago">Dolinšek
               Drago</actor> 26, 46</castItem>
         <!-- ... -->
         <castItem>
            <actor sameAs="#sp.ZupančičJože">Jože
               Zupančič</actor> 56, 59</castItem>
         <!-- ... -->
      </castList>
   </div>
</front>
<body><!-- govori -->
   <div><!-- vrsta zbora oz. skupščine -->
      <div><!-- seja oz. zborovanje -->
         <!-- ... -->
         <castList>
            <!-- predsedujoči -->
            <castItem>
               <roleDesc>Sejo je vodil</roleDesc>
               <actor xml:id="sp.ZupančičJože">Jože Zupančič</actor>,
               <role>predsednik Zbora združenega dela</role>
            </castItem>
         </castList>
```

Slika 5: Seznam govornikov in seznam predsedujočih.

Oseba, ki izvaja kodiranje, vse razdelke div tudi naknadno označijo z globalnim atributom ana, ki se navezuje na skupno taksonomijo taxonomy v kolofonu dokumenta TEI. Razdelki, ki vsebujejo vsebinske sklope, se preko atributa corresp navezujejo na vsebinsko ustrezen item element v kazalu vsebine.

```
<div><!-- vsebinski sklop -->
   <sp who="#sp.PriimekIme">
      <speaker>Ime in priimek govornika in
         morebitne oznake govornika</speaker>
      <p>Besedilo govora, besedilo <title>naslov
         sklopa</title> besedilo govora,
         <stage>komentar zapisnikarja magnetograma
            razprave</stage> besedilo.</p>
   </sp>
   <stage type="time">
      <!-- čas prekinitve in ponovnega začetka razprave -->
   </stage>
   <sp who="#sp.PriimekIme">
      <speaker>Ime in priimek</speaker>
      <ab>Govor iz klopi, ki ga je komentator zapisal v sklopu
         govora govornika iz govorniškega odra.</ab>
   </sp>
   <stage type="time">
      <!-- čas konca razprave -->
   </stage>
</div>
```

Slika 6: Kodiranje govorov.

Znotraj posameznega vsebinskega sklopa so govori označeni v skladu s TEI modulom za dramska besedila (slika 6): Govor je označen z elementom sp, govorec z elementom speaker, govorjeno besedilo z elementi odstavek p ali anonimni blok ab. Atribut sp/@who se navezuje na omembo tega govorca v seznamu govornikov castList. Anonimni bloki se nahajajo samo znotraj tistih govorov, ki niso bili opravljeni z govorniškega odra, temveč kot medklici iz poslanskih klopi. V magnetogramih so zapisnikarji takšne govore kot medklice označili znotraj odstavkov govorov iz govorniškega odra. Kot vse druge komentarje so jih od ostalega govora zamejili z oklepaji. Zato mora oseba, ki opravlja kodiranje v TEI, takšne govore kodirati na roko. Vsi ostali komentarji znotraj odstavkov so označeni kot didaskalije stage. Znotraj odstavkov se z elementom title kodira še morebitne naslove vsebinskih blokov. To so naslovi, ki jih je ob napovedi novega vsebinskega sklopa napovedal predsedujoči.

```
<!-- čas prekinitve in ponovnega začetka razprave -->
<stage type="time">(Seja je bila <time to="1990-07-02T11:30:00"
   xml:id="stage.t.2">prekinjena ob 11.30 uri</time> in se je <time
      from="1990-07-02T19:45:00" xml:id="stage.t.3">nadaljevala ob
      19.45 uri</time>.)</stage>
```

Slika 7: Kodiranje časa začetka in konca govorov.

Komentarji zapisnikov megnetogramov, ki vsebujejo podatke o času, ko so se govori začeli oz. končali, so kodirani kot element stage, ki ima vrednost atributa type time (glej sliko 7). Seja je bila lahko večkrat prekinjena. Prekinitve so bile lahko le krajši predahi ali nekajdnevne preložitve. Znotraj elementa stage je začetek časovnega bloka kodiran s time/@from in konec s time/@to. Na unikatni identifikator elementa time se navezuje časovnica timeline/when. Slednja je narejena avtomatsko iz kodiranih podatkov v time.

Delovni proces kodiranja besedila poteka v skladu s smiselno predvidenimi koraki, ki so prilagojeni posebnostim izvornega besedila zapisnikov sej. Oseba, ki opravlja kodiranje, si za vsak novi korak najprej izbere XSLT stil, s katerim naredi avtomatsko pretvorbo. Po končani pretvorbi popravi morebitne nedoslednosti in ročno kodira nekatere dele besedila. Na ta način se krajše seje lahko kodira v pol ure, za daljše seje se običajno porabi do dve uri, za najdaljše (tudi več kot 200000 besed) pa do 4 ure.

4 Dostopnost, uporaba in načrti

Delovne verzije TEI zapisnikov sej so dostopni na GitHub.[14] Trenutno smo kodirali 53 sej zapisnikov sej do leta 1990, ki vsebujejo 4382 govorov 828 različnih govornikov, ki so skupaj izgovorili 1.150.000 besed, v prilogah pa smo kodirali še 200.000 besed. Te seje so naključno izbrani vzorci, s pomočjo katerih smo preizkušali različne načine kodiranja. Toda za uporabo različnih historičnih analiz so ti vzorci povsem neprimerni. Zato smo se odločili, da v celoti kodiramo vsaj en sklic. Izbrali smo zgodovinopisno zelo zanimiv 11. sklic ”osamosvojitvene” skupščine (1990–1992). Trenutno smo kodirali skoraj celotno besedilo (manjka Zbor občin): 41.131 govorov, 7.574.000 besed.

To besedilo nato shranimo v nov GitHub repozitorij SlovParl,[15] kjer osnovne dokumente TEI še dodatno kodiramo v skladu z nameni zgodovinopisne raziskave. Trenutno imamo v tem repozitoriju dodatno kodirano celotno besedilo Zbora združenega dela (2.739.000 besed). Največ dela smo vložili v izdelavo novih

[14] Sejni zapiski, https://github.com/SIstory/Sejni_zapiski; Seje Državnega zbora, https://github.com/SIstory/Seje_DZ.
[15] SlovParl, https://github.com/SIstory/SlovParl.

dokumentov TEI, ki omogočajo dodatno analizo TEI korpusa.

Obstoječi seznami govornikov v castList so namreč neprimerni za resno historično analizo. Ker smo njihove unikatne identifikatorje narejeni z avtomatsko pretvorbo zapisanih imen in priimkov, so iste osebe, ki imajo v drugih primerih drugače zapisano svoje ime, označene kot različne osebe. Obenem se različne osebe z istim imenom in priimkom[16] avtomatično smatra za isto osebo. Zato je na podlagi različnih zgodovinskih virov potrebno sezname poslancev, ministrov, poročevalcev in drugih govornikov ustrezno preveriti in prečistiti. To je tudi idealna priložnost, da tako narejenemu seznamu dodamo čim več javno dostopnih osebnih podatkov. Te podatke smo zapisali v ločenem dokumentu TEI speaker.xml.

Osebe smo kodirali v seznamu oseb listPerson/person. Znotraj person elementa smo označili njihova imena (persName), spol (sex), datum in kraj rojstva (birth), datum in kraj smrti (death), izobrazbo (education), poklic (occupation), bivališče (residence) in različna službovanja (affiliation) v službah, na funkcijah, v političnih strankah in drugih organizacijah, nenazadnje v parlamentu. Ker se je pripadnost osebe organizacijam čez čas spreminjala, smo veliko pozornost posvetili prav kodiranju teh sprememb. Zato so elementi affiliation preko atributov ref in ana navezani na ustrezno kodirane ustanove org v seznamu organizacij listOrg. Te organizacije (zbori, stranke, ministrstva) smo kodirali na način, ki ne omogoča le njihove analize na podlagi obstoja teh organizacij skozi čas, temveč tudi na podlagi njihovega razvoja skozi čas (preimenovanja, združevanja in razdruževanja).

Kodirani govori se preko atributa sp/@who ne navezujejo več na seznam govornikov v castList, temveč na poenoten seznam oseb listPerson. Na castList so po novem navezane preko atributa corresp. Na podlagi teh povezav lahko zastavljamo različna bolj ali manj kompleksna raziskovalna vprašanja (Pančur in Šorn, 2016). Tako npr. hitro izvemo, da sta bila v Zboru združenega dela najbolj zgovorna poslanca Jože Zupančič (735.000 besed) in Bogo Rogina (114.000), kar niti ni presenetljivo, saj je bil prvi predsednik in drugi podpredsednik tega zbora. Med navadnimi poslanci je zato rekorder Jože Arzenšek (106.000), najmanj zgovoren pa je bil Jože Košak, kateremu je uspelo izreči le 14 besed. Z le nekoliko bolj zapleteno poizvedbo lahko tudi ugotovimo, da so poslanci, ki so ob začetku mandata pripadali koaliciji DEMOS, spregovorili 22 % vseh besed, poslanci iz opozicijskih strank 23 % in neodvisni poslanci kar 55 %.

Odgovore na bolj zapletena raziskovalna vprašanja omogoča tudi na novo izdelan dokument TEI, ki v gnezdenem elementu list vsebuje tematsko kazalo točk dnevnega reda. Pri izdelavi tega kazala se je uporabilo podatke iz obstoječih vsebinskih sklopov in kazal vsebine. Postavke item tematskega kazala se pri tem navezujejo na ustrezne unikatne identifikatorje vsebinskih sklopov govorov. V enotnem kazalu smo najprej povezali vse točke dnevnega reda, ki so bile pred tem lahko razbite na različne dneve zasedanj. Potem smo med seboj povezali sorodne vsebinske sklope, npr. sprejem določenega zakona ter njegovih sprememb. Nato smo na podlagi poslovnika skupščine izdelali posebno vsebinsko shemo. Najbolj obsežen sklop *Akti in postopki* smo razvrstili v skladu s tematskim kazalom pravnega reda Republike Slovenije.[17] S pomočjo tako povezanih podatkov lahko hitro izvemo, da je bil zakon, o katerem so poslanci Zbora združenega dela najbolj razpravljali, Zakon o lastninskem preoblikovanju podjetij (103.870 besed). Med sprejemom tega zakona se je zvrstilo 520 govorov, obravnavali pa so ga v štirinajstih terminih. Kot zanimivost lahko navedem še podatek, da je v povprečju razprava nepretrgoma (do prvega odmora ali konca seje) trajala 100 minut.

Že na podlagi te stopnje kodiranja lahko torej odgovorimo na zelo različna raziskovalna vprašanja. Z dopolnjevanjem obstoječega seznama poslancev in ostalih govornikov, lahko ta vprašanja še dodatno razširimo. Z uporabo drugačnega tematskega kazala lahko raziskovalna vprašanja prilagodimo svojim potrebam. V načrtu imamo še kodiranje imenskih entitet v govorih. Poskusno smo pri tem že uporabili Stanford NER za slovenščino (Ljubešić et al., 2013). Trenutno za analizo v glavnem uporabljamo XSLT stile, ko pa bo zbirka dokumentov TEI narasla na več deset milijonov besed, se bo v glavnem uporabljalo (XQuery) aplikacije NoSQL baze dokumentov eXist, ki med drugim omogoča tudi indeksacijo (Siegel in Retter, 2015). Ko se bo naposled zbirka sejnih zapisnikov v celoti pretvorila v XML, bo primerna za uporabo povsem novih metod v historičnih raziskavah kot so različne raziskave zgodovine konceptov ali raziskave glede sprememb v odnosu do druge svetovne vojne v povojnih parlamentarnih razpravah (Piersma, 2014).

Zapisnike sej, ki smo jih kodirali v skladu s TEI modulom za dramska besedila, smo naknadno pretvorili še v dokumente TEI, kjer je besedilo označeno v skladu s TEI modulom za transkribiranje govorov. Ta pretvorba je shranjena v GitHub repozitoriju CLARIN.SI.[18] Pri tem smo za govore uporabili sledeče mapiranje:

- sp/p → div[@type='sp']/u
- sp/ab → div[@type='inter']/u
- stage → div/note.

Ker smo se pri tem odločili, da elementi u ne vsebujejo nobenih drugih elementov, temveč samo besedilo, jih lahko naknadno avtomatsko jezikovno označimo. Trenutno so partnerji iz Inštituta Jožefa Stefana že izvedli tokenizacijo, oblikosladenjsko označevanje in lematizacijo. Korpus so uvozili v spletni konkordančnik noSketchEngine (Erjavec, 2013), vsi dokumenti TEI pa so dostopni v repozitoriju CLARIN.SI (Erjavec et al., 2014).

5 Literatura

Tomaž Erjavec. 2013. Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0*, 1(1): 24-49. http://slovenscina2.0.trojina.si/arhiv/2013/1/Slo2.0_2013_1_03.pdf.

Tomaž Erjavec, Jan Jona Javoršek in Simon Krek. 2014. Raziskovalna infrastruktura CLARIN.SI. V: Ninth Language Technologies Conference, str. 19-24. Ljubljana: IJS. http://nl.ijs.si/isjt14/proceedings/isjt2014_03.pdf.

Global Centre for ICT (2014). Technological Options for Capturing and Reporting Parliamentary Proceedings.

[16] V 11. sklicu skupščine sta bila npr. dva Jožeta Smoleta.

[17] PIS: Pravno-informacijski sistem, http://www.pisrs.si/Pis.web/pravniRedRSDrzavniNivoKazalaTematskoKazalo.

[18] https://github.com/DARIAH-SI/CLARIN.SI.

http://www.ictparliament.org/sites/default/files/handbook-proceedings_1.pdf.

Miloš Jakubíček in Vojtěch Kovář. 2010. CzechParl: Corpus of Stenographic Protocols from Czech Parliament. V: Petr Sojka in Aleš Horák, ur., *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010*, str. 41-46, Tribun EU. http://www.muni.cz/research/publications/914313.

N. Ljubešić, M. Stupar, T. Jurić, Ž. Agić. 2013. Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene. *Slovenščina 2.0*, 1(2): 35-57. http://www.dlib.si/details/URN:NBN:SI:DOC-VSWXF4CE.

Carlos Martin-Dancausa in Maarten Marx. 2010. Parliamentary documents from Spain. V: Proceedings of the International Conference on Language Resources and Evaluation, LREC, Valetta, Malta.

Maarten Marx. 2009. Advanced Information Access to Parliamentary Debates. *Journal of Digital Information*, 10(6): 1-11. https://journals.tdl.org/jodi/index.php/jodi/article/view/668.

Maarten Marx in Anne Schuth. 2010. DutchParl: The Parliamentary Documents in Dutch. V: Proceedings of the International Conference on Language Resources and Evaluation, LREC, Valetta, Malta.

Maciej Ogrodniczuk. 2012. The Polish Sejm Corpus. V: *LREC 2010, Eight International Conference on Language Resources and Evaluation*, str. 2219-2223, Istanbul, Turkey. http://www.lrec-conf.org/proceedings/lrec2012/index.html.

Petya Osenova in Kiril Simov. 2012. The Political Speech Corpus of Bulgarian. V: *LREC 2010, Eight International Conference on Language Resources and Evaluation*, str. 1744-1747, Istanbul, Turkey. http://www.lrec-conf.org/proceedings/lrec2012/index.html.

Andrej Pančur in Mojca Šorn. 2016. Digitalni pristop k parlamentarni zgodovini: uporaba gradiva Državnega zbora v digitalni humanistiki. V: *Četrt stoletja Republike Slovenije – izzivi, dileme in pričakovanja*, str. 115-126. Inštitut za novejšo zgodovino, Ljubljana.

Hinke Piersma, Ismee Tames, Lars Buitinck in Maarten Marx. 2014. War in Parliament: What a Digital Approach Can Add to the Study of Parliamentary History. *DHQ: Digital Humanities Quarterly*, 8(1). http://www.digitalhumanities.org/dhq/vol/8/1/000176/000176.html.

Erik Siegel in Adam Retter. 2015. eXist: A NoSQL Document Database and Application Platform. O'Reilly, Cambridge.

TEI Consortium. 2015. TEI P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/Guidelines/P5.

Andrej Žgank, Tomaž Rotovnik, Matej Grašič, Marko Kos, Damjan Vlaj in Zdravko Kačič (2006). Slovenska govorna in tekstovna baza parlamentarnih razprav za avtomatsko razpoznavanje govora. V: Mednarodna multi-konferenca Informacijska družba IS, str. 115-118, Ljubljana. http://nl.ijs.si/isjt06/proc/22_Zgank_2of2.pdf.

Raba vejice v uporabniških spletnih vsebinah

Damjan Popič,* Darja Fišer,*† Katja Zupan,† Polona Logar‡

* Oddelek za prevajalstvo Filozofske fakultete v Ljubljani
Aškerčeva 2, 1000 Ljubljana
damjan.popic@ff.uni-lj.si, darja.fiser@ff.uni-lj.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Mednarodna podiplomska šola Jožefa Stefana
Jamova cesta 39, 1000 Ljubljana
katja.zupan@ijs.si
‡ Oddelek za slovenistiko Filozofske fakultete v Ljubljani
Aškerčeva 2, 1000 Ljubljana
polona.logar1@gmail.com

Povzetek

V prispevku predstavimo pilotno študijo nestandardne stave vejice v slovenskih uporabniških spletnih vsebinah. Najprej predstavimo razvoj tipologije za označevanje nestandardne stave vejice v slovenščini, v nadaljevanju pa predstavimo rezultate analize 500 naključno izbranih tvitov različnih stopenj standardnosti. Analiza je pokazala, da je v slovenskih uporabniških vsebinah problematična predvsem manjkajoča vejica, in sicer med odvisniki in nadrednimi stavki ter pri pri- ter pastavčnih strukturah.

Comma Use in User-Generated Content

The article presents a pilot study of non-standard use of the comma in Slovene user-generated content. Initially, the development of the typology for annotating non-standard comma use in Slovene is presented. Afterwards, the results of an analysis of 500 randomly selected tweets (with two varying levels of standardness) are given. The results show that in Slovenian user-generated content comma use is problematic mostly in regard to the missing comma, especially between dependent and independent clauses, and after and before small clauses.

1 Uvod

S primerjalno raziskavo korpusov Janes (Erjavec et al. 2015) in Kres (Logar et al. 2012) je bilo pokazano, da raba vejice na družbenih omrežjih ne peša in da so razlike med rabo vejice v tradicionalnih in novomedijskih besedilih manjše, kot pregovorno velja (Popič in Fišer 2015). Vendar je bila raziskava omejena le na ožji nabor tipičnih atraktorjev vejic (Verovnik 2003, Žibert 2006) in na samo pogostnost ter razporejenost vejic, zato jo želimo nadgraditi s pričujočim prispevkom in pripraviti izhodišče za celovito analizo rabe vejice, predvsem pa razviti čim bolj univerzalen kategorizacijski sistem za označevanje nestandardne rabe vejice in ga preizkusiti na širšem vzorcu.

V pričujoči raziskavi preverimo, v čem raba vejice v uporabniških spletnih vsebinah odstopa od norme. Analiza temelji na označevanju napačno stavljene vejice v naključnem vzorcu 500 tvitov iz korpusa Janes v0.4. Pri tem smo kot merilo upoštevali še stopnjo jezikovne in tehnične standardnosti besedil (Ljubešić et al. 2015) ter v vzorec zajeli po 250 tvitov z naslednjima oznakama: T1L3 (nestandarden jezik, a standardna raba velikih začetnic, presledkov, ločil) in T3L3 (tako tehnično kot tudi jezikovno povsem nestandardni tviti). Ker nas zanima stava vejice v nestandardni slovenščini, tvitov, zapisanih v standardni slovenščini (L1), v analizo nismo zajeli, vendar v prihodnje načrtujemo razširitev raziskave tudi na standardne tvite in druge tipe uporabniških spletnih vsebin, kot so komentarji na novice, forumska sporočila, blogi in pogovorne strani na Wikipediji.

Na podlagi označenega korpusa smo opravili kvantitativno ter preliminarno kvalitativno analizo manjkajočih in odvečnih vejic v uporabniških spletnih vsebinah ter tako dobili vpogled v najbolj problematične kategorije.

2 Metodologija

Označevanje nestandardne stave vejic je potekalo v orodju WebAnno (Eckart de Castilho et al. 2014), uporabili pa smo kategorije, predstavljene v Tabeli 1. V nadaljevanju najprej predstavimo zasnovo in izpopolnjevanje tipologije za označevanje nestandardne rabe vejic, zatem pa opišemo še proces označevanja in delotok.

2.1 Tipologija

Za opis nestandardne (ne)rabe vejice smo razvili obsežno tipologijo, pri tem pa smo se oprli na aktualni jezikovni predpis (Slovenski pravopis 2001, Pravila), z nekaj dopolnitvami in prilagoditvami. Za podporo sorodnim raziskavam in vpogled v sistem označevanja pri pričujoči raziskavi smo smernice za označevanje objavili na spletu.[1] Pri sestavi kategorizacije smo se primarno držali načela, da naj bo ta čim bolj univerzalna (da torej lahko s čim bolj podobnimi kategorijami opiše čim več primerov nestandardne rabe vejice (tj. odvečne vejice, manjkajoče vejice itd.)). Seveda to ni bilo povsem izvedljivo, zato tipologija vsebuje tudi kategorije, vezane predvsem na specifične primere rabe (tako denimo kategorija Stavčni člen opisuje predvsem nestandardno »stavčnočlensko« rabo vejice, ki je vezana zgolj na odvečno vejico), povečini

[1] Smernice za označevanje so na voljo na spletni strani http://nl.ijs.si/janes/wp-content/uploads/2014/09/vejice-smernice.pdf.

pa so kategorije vendarle univerzalne in namenjene opisu katerega koli segmenta jezika – tipologija torej ni namenjena zgolj opisu nestandardne stave vejice v spletni rabi, temveč je bil namen sestavljavcev prav ta, da lahko zadosti opisu katerega koli žanra ali besedilnega tipa.

Poleg jezikovnosistemskega opisa stave vejice smo se pri pripravi tipologije oprli tudi na že opravljene empirične, izgradivne preglede stave vejice v različnih okoljih, in sicer v spletnem (Popič in Fišer 2015), šolskem (Logar in Popič 2015, Kosem et al. 2012) ter v okolju poklicnih piscev (Popič 2014). Z začetno različico tipologije smo označili testni nabor tvitov in na podlagi pridobljenih spoznanj tipologijo dopolnjevali. Bistvene razlike med začetno in končno tipologijo so se nanašale predvsem na nestandardno odvečno vejico, pri kateri tipična razdelitev glede na jezikovnosistemski opis odpove, saj temelji na skladenjski razčlenitvi – odvečnost vejice pa temelji ravno na tem, da uporabnik skladenjske razčlenitve ne (pre)pozna. Tipologijo podajamo v nadaljevanju, brez ponavljajočih se podkategorij – pri vseh segmentih skladenjske rabe vejice smo namreč uvedli še podkategoriji levo- in desnosmerne vejice (Korošec 2003), saj v tem prepoznavamo različne (potencialne) vzgibe in razloge za nestandardno stavo vejice.

Manjkajoča/odvečna vejica/napačen znak
1 Skladenjska
1.1 Priredje
1.1.1 Vezalno *@xxx jst sm se že zdavni odločla,da bo prej!se je pa tut že zdavni okol obrnu,in prehiteva po rasti..tko,da držim pesti! @xxx*
1.1.2 Stopnjevalno *@xxx tip nikakor ne sodi v odbore kaj sele v parlament...ta je bolj za debate pr'Kovac...tam je bolj po njegovem stilu...*
1.1.3 Ločno *@xxx @xxx @xxx @xxx Zlata, men se odpre le uradna stran RTV. Kaj moram nardit, oz za katero oddajo gre?*
1.1.4 Protivno *@xxx Razumem ampak point rapa na začetku je bil predvsem v tem kakšno zgodbo boš povedal btw ma Kendrick tud bangerja...*
1.1.5 Vzročno *@xxx za urgenco #UKC in ZD zelo strinjam..lah bi še mal fasado poštimal pa kak oddelk dogradil drgač pa super ja..#domzale*
1.1.6 Posledično *@xxx not even close.. ampak vseen dobr vedet :) .. zdej bom ene dva tedna na zlo low budget tko da se ne branim marelic :)*
1.1.7 Pojasnjevalno *@xxx jao, to so nardil možu, ko je bil v ZDA in to pred startom dirke. Na srečo je šla ekipa dol kasneje in da sem bila pooblaščena.*
1.2 Odvisnik
1.2.1 Osebkov *@xxx pri moji ghetto mikrovalovki tm enih 20sek.mal sčekirej umes. je dost da je topla, za čez šmorn glih kul. :) @xxx*
1.2.2 Predmetni *@xxx neki sm vidu sproti, drugač pa ne vem kaj dogaja in kaj je to..:) razen komada, pa da je "snovalec" le tega otvoril zadevo..*
1.2.3 Krajevni *@xxx o/ mel sem srečo. Če nebi blo nesreče zjutraj na Ac bi imel avto tam kjer je bilo meter vode. Sm se pa peljal do Ajdušne*
1.2.4 Časovni *@xxx Vse živo ... trust me !! :) Preden pa jaz grem na Švedsko pa tebe verjetno čaka kofi pri Slamiču hehe :) Kaj pravš ? ;)*
1.2.5 Načinovni *Tud macki so po lastnikih. K rai se tko igra da lezi na istem mestu pa zamahne s tacko sam takrt ko igraca pride cist do njega #lenoba*
1.2.6 Vzročni *Fakt št. 1: Skor sm zgubila glavo na tekočih stopnicah ker sm oprezala za enim tipom. Fakt št. 2: V torbici nosim cedevito. #normalday*
1.2.7 Namerni /
1.2.8 Pogojni *@xxx @xxx @xxx @xxx @xxx @xxx ja ce bote organiziral v 'otrokuprijaznim' uram pa tud jz pridem :)*
1.2.9 Dopustni *@xxx je dobra in rešuje življenje poštenim ljudem, ki drgač ne bi imeli možnosti, mi je vseen če tud kak opravičil napiše @xxx*
1.2.10 Prilastkov *drgač pa ivan ni nestrpen do srbov , bosancev in hrvatov. ma enga prjatla k ma prjatla k ma prjatla k ma vodovodarja iz prijedora #truestory*
1.3 Polstavek *Jst bi tud najdu kovanec vreden veliko denarja. Kje je drgač sploh logika? Gor piše neki, vrednost pa veliko večja. Težave modernega sveta.*
1.4 Stavčni člen *@xxx poznam komb.proja+kneipp+franck za "belo kavo". sem rad pil, ko sem bil mali:) za muckefuck, pa sem slišal 3 dni nazaj :D*
1.5 Besedna zveza *Medtem, ko vsi brenčite o stricih, MK in BP jaz razmišljam le o #Gaza. In razmišljam kdaj je Twitter postal just another social network.*
1.6 Pri-, pa- in dostavek ter izpostavek *sej sm nasla v smeteh pol ful dobr res haha jao* 👐👍
2 Neskladenjska
2.1 X → vejica /
2.2 Vejica → X /
2.3 Tipkarska napaka /
2.4 Drugo (gl. Tabela 6)

Tabela 1: Pregled skladenjskih in neskladenjskih oznak po kategorijah.

Kot lahko vidimo, kategorije pretežno temeljijo na razdelitvi, podani v jezikovnem predpisu, dodani sta kategoriji Stavčni člen in Besedna zveza, pri neskladenjski rabi vejice pa smo želeli čim podrobneje razčleniti, za katero vrsto nestandardne stave gre, zato so vključene še štiri podkategorije. Za neskladenjsko rabo vejice je rezerviran tudi parameter »napačen znak«, saj v tovrstnih primerih kategoriji manjkajoče in odvečne vejice nista (nujno) relevantni.

V določeni meri smo jezikovnosistemske kategorije tudi združevali (gl. kategorijo 1.6), in sicer na mestih, kjer drobljenje informacij po naših pričakovanjih ne bi bistveno izboljšalo podatkov o nestandardni rabi vejice, bi pa nesorazmerno povečalo število kategorij.

2.2 Označevanje

Med procesom označevanja je bilo dvojno označenih 500 tvitov. Pri označevanju je bilo mogoče za posamezni primer izbrati le eno kategorijo, večkategorialnih oznak – npr. ko je utemeljitev za stavo vejice lahko tako levo- kot desnosmerna – nismo dopuščali. Če je bilo mogočih več interpretacij, smo izbrali kategorijo, ki je bila v danem kontekstu bolj povedna oz. verjetnejši razlog za napačno stavo vejice, pri čemer smo favorizirali odvisnike. Pogost primer je npr. oziralni odvisnik sredi tro- ali večstavčnih povedi, kjer pogosto umanjka levosmerna vejica odvisnika.

Po koncu označevanja je kuratorka pregledala vseh 500 dvojno označenih tvitov in pri neskladjih med označevalcema sprejela dokončno odločitev. Po podatkih orodja WebAnno je Cohenov koeficient ujemanja med označevalcema kappa znašal 0,57, kar je srednje dobro ujemanje. Zelo dobrega ujemanja zaradi težavnosti problema, sprotnega vzpostavljanja tipologije in smernic ter neizkušenosti označevalcev ni bilo mogoče pričakovati – tudi zato smo se odločili za dvojno označevanje. Neskladja med označevalcema so bila najpogostejša predvsem pri naslednjih segmentih:

- ločevanje posameznih medmetov znotraj pastavka;
- prepoznavanje vezalnosti, protivnosti in posledičnosti veznikov *pa* ter *in*;
- obravnava smeškov in drugih metajezikovnih struktur, tipičnih za uporabniške vsebine;
- prepoznavanje vrst odvisnikov, zlasti v primerih, pri katerih so vezniške strukture zapisane nestandardno.

Pri prvih treh problematičnih mestih smo se odločili za princip minimalne intervencije, zato se je kuratorka izogibala posegov v izvorno stavo vejic, razen če je poseg nedvoumno predvidevalo določeno pravopisno pravilo ali kontekst, medtem ko je bila pri zadnjem problematičnem torišču potrebna vsebinska odločitev. Za čim ustreznejšo obravnavo težavnih segmentov smo sprejeli natančne smernice, da bo označevanje v prihodnje enostavnejše in bolj usklajeno.

Skupno je bilo pri 500 tvitih označenih 405 mest, kar pomeni v povprečju 0,8 oznake na posamezni tvit.

3 Analiza

V nadaljevanju predstavimo izsledke analize nestandardne rabe vejic v uporabniških spletnih vsebinah. Najprej predstavljamo kumulativne rezultate po posameznih kategorijah, ki jih podajamo v Tabeli 2, v nadaljevanju pa se osredotočimo na posamezne (pod)kategorije, zlasti tiste, ki so glede na rezultate empirične raziskave v uporabniških spletnih vsebinah še posebno problematične.

Kategorija	Odveč	Manjka	Nap. znak
1 SKLADENJSKA	**19 (4,7 %)**	**382 (95,3 %)**	**0**
1.1 Priredja	**8 (2 %)**	**35 (8,7 %)**	**0**
1.1.1 Vezalno	**6**	8	**0**
1.1.2 Stopnjevalno	0	2	**0**
1.1.3 Ločno	2	0	**0**
1.1.4 Protivno	0	**13**	**0**
1.1.5 Vzročno	0	1	**0**
1.1.6 Posledično	0	10	**0**
1.1.7 Pojasnjevalno	0	1	**0**
1.2 Odvisniki	**2 (0,5 %)**	**230 (57,4 %)**	**0**
1.2.1 Osebkov	0	26	**0**
1.2.2 Predmetni	1	**89**	**0**
1.2.3 Krajevni	0	1	**0**
1.2.4 Časovni	0	20	**0**
1.2.5 Načinovni	1	11	**0**
1.2.6 Vzročni	0	16	**0**
1.2.7 Namerni	0	0	**0**
1.2.8 Pogojni	0	25	**0**
1.2.9 Dopustni	0	1	**0**
1.2.10 Prilastkov	0	41	**0**
1.3 Polstavek	**0**	**2 (0,5 %)**	**0**
1.4 Stavčni člen	**3 (0,7 %)**	**0**	**0**
1.5 Besedna zveza	**6 (1,5 %)**	**0**	**0**
1.6 Pa-, pri- do- in izpostavek	**0**	**115 (28,7 %)**	**0**
2 NESKLADENJSKA	**0**	**3 (75 %)**	**1 (25 %)**

Tabela 2: Pregled skladenjskih in neskladenjskih oznak po kategorijah. Pri nadkategorijah so podani deleži oznak znotraj skladenjskega in neskladenjskega segmenta.

Kot lahko razberemo iz Tabele 2, so oznake med kategorijami razporejene zelo heterogeno, zelo neenakomerno pa so posejane tudi oznake glede na vrsto skladenjskega razmerja. Ta heterogenost je značilna za obe ravni tehnične standardnosti tvitov, ki sta zajeti v analiziranih podatkih.

Kot lahko vidimo, je bilo v 500 tvitih zanemarljivo malo (4) primerov napačno stavljene neskladenjske vejice, prav tako pa se je pri izbranem vzorcu pokazalo, da je v uporabniških spletnih vsebinah skorajda zanemarljiv pojav odvečna vejica (19 primerov), saj je velika večina (382 primerov) oznak vezana na manjkajočo vejico. Pri tem je treba poudariti, da zaradi majhnega vzorca te ugotovitve ni primerno posploševati. Velika večina popravkov se nanaša na skladenjsko rabo vejice (401 od skupno 405 primerov oz. 99 %). Glede na rezultate empirične analize lahko vidimo, da sta še posebno akutni kategoriji manjkajočih vejic pri odvisnikih (230 primerov oz. 57 % vseh skladenjskih oznak) in pastavčnih strukturah (115 primerov oz. 28,7 % vseh skladenjskih oznak).

3.1 Manjkajoča vejica

Kot lahko vidimo v Tabeli 2, je stava vejice z vidika manjkajoče vejice problematična predvsem pri odvisniških in pristavčnih strukturah, zato se v nadaljevanju nekoliko podrobneje posvetimo prav tema kategorijama. V Tabeli 3 je podan prikaz najpogostejših odvisniških vrst z označeno manjkajočo vejico.

Odvisnik	Število oznak
Predmetni (D)	**80 (19,8 %)**
Prilastkov (D)	29 (7,1 %)
Osebkov (D)	24 (5,9 %)
Vzročni (D)	16 (4,0 %)
Pogojni (D)	13 (3,2 %)
Časovni (D)	13 (3,2 %)
Pogojni (L)	12 (2,9 %)

Prilastkov (L)	12 (2,9 %)
Načinovni (D)	10 (2,5 %)
Predmetni (L)	9 (2,2 %)
Časovni (L)	7 (1,7 %)
Osebkov (L)	2 (0,5 %)
Načinovni (L)	1 (0,2 %)
Krajevni (D)	1 (0,2 %)
Dopustni (D)	1 (0,2 %)

Tabela 3: Pregled manjkajočih vejic pri odvisnikih po frekvenci in deležu znotraj skladenjske kategorije.

Kot lahko vidimo, je od skupno 230 manjkajočih vejic pri odvisnikih edina zares problematična kategorija desnosmerne vejice pri predmetnih odvisnikih, pri katerih najdemo več kot tretjino primerov z manjkajočo vejico. Tudi na splošno se je izkazalo, da je desnosmerna vejica precej bolj akutna kot levosmerna, saj zaseda šest najpogostejših mest z manjkajočo vejico (175 od skupno 230 manjkajočih vejic). Edini odvisniški vrsti, pri katerih smo zaznali nekaj več manjkajočih levosmernih vejic, sta kategoriji prilastkovih in vzročnih odvisnikov. Za slednje lahko sklepamo, da pogosto začenjajo povedi, pri prilastkovih odvisnikih pa uporabniki pogosto pozabijo skleniti oziralni odvisnik.

	Odvečna	**Manjkajoča**
Levosmerna	/	79 (19,7 %)
Desnosmerna	/	36 (9,0 %)

Tabela 4: Pregled oznak pri pa-, pri-, do- in izpostavčnih strukturah po frekvenci in deležu skladenjskih oznak.

Pri pristavkih, dostavkih, izpostavkih in pastavkih je problematična predvsem vejica na začetku, relativno pogosta pa je tudi nestandardna stava brez vejice na koncu stavka. V številnih primerih gre za nestandardno stavo vejice pri denimo medmetih (npr. *hahaha*), ki jih – tako domnevamo – uporabniki spletnih omrežij ne dojemajo kot pastavčne tvorbe, temveč kot metajezikovno pojavnost oz. verbalizirane emotikone. Zelo pogosta je tudi pojavnost nestandardne rabe vejice pri členkovnih pastavkih na začetku in koncu stavka, zlasti v navezavi s sklicem na uporabniško ime. Te rabe v korpusu sicer nismo označevali kot odmik od standarda, predvsem zaradi velike pogostosti in zaradi tehničnih zahtev uporabe Twitterja, saj lahko stavljenje vejice v sklice povzroči težave s samim sklicem. V vsakem primeru pa način uporabe sklicev potrjuje domnevo, da uporabniki izražajo metajezikovne prvine z uporabo jezikovnih sredstev.

3.2 Odvečna vejica

Vseh primerov odvečne vejice od skupno 401 skladenjske oznake v zbirki podatkov je bilo 19, kar pomeni manj kot 5 odstotkov vseh oznak, njihova porazdelitev po kategorijah pa je naslednja:

Oznaka	**Število**
Vezalno priredje	6
Besedna zveza	6
Stavčni člen	3
Ločno priredje	2
Predmetni odvisnik	1
Načinovni odvisnik	1

Tabela 5: Pregled odvečnih vejic glede na kategorijo po frekvenci in deležu skladenjskih oznak.

Kot lahko vidimo, je bila vejica največkrat odvečna v besednih zvezah (denimo med sestavljenimi vezniki), med strukturami v vezalnem ali ločnem priredju, za stavčnim členom in med enakovrednima odvisnikoma. Sodeč po skupnem številu označenih nestandardnih mest, bi pričakovali večje število odvečnih vejic, predvsem stavčnočlenskih, zlasti glede na raziskave na standardnojezikovnem gradivu (prim. Popič 2014). Preden lahko zaključimo, da je neskladenjska raba vejice pretežno manjkajoča, ker je to ena od strategij krajšanja sporočil v tovrstnem načinu komuniciranja in ker je vejico psihološko in tehnično lažje izpustiti kot pa napisati po nepotrebnem, bi bilo nujno treba opraviti analizo na večjem vzorcu besedil tega žanra.

3.3 Neskladenjska vejica

Kot smo pokazali v predhodnih poglavjih, so bili v celotnem naboru podatkov označeni zgolj štirje primeri nestandardne neskladenjske vejice, ki so se pojavili v treh tvitih. Te izpisujemo v celoti v Tabeli 6.

Setamo po Lj in pride Zoki mim, se ustavi pa da Emanuelu petko. Tamalmu nc jasn. Recem**(,)** to je Zoki kralj**(,)** in se tamal zadere: Zoki kralj! :)
@xxx Ja bi mogla tud jst naletet na dobrega :) ampak ima res ogromno folka sfaljenega (ne ostri do **2.8**). Je, kar je.
@xxx "Maybe yes, maybe no, maybe go home**(,)**" so radi rekli Šerpe na odpravi na Annapurno, kadar smo jih vprašali, če bo vreme OK

Tabela 6: Pregled primerov z označeno napačno rabljeno neskladenjsko vejico.

Kot lahko vidimo, so od tega tri vejice v resnici skladenjske, in sicer gre za tri primere soredja (1. in 3. tvit; v oklepaju so podane manjkajoče vejice na mestih, ki so bila označena tudi med označevanjem), ki ga namenoma nismo vključili v tipologijo, saj nismo pričakovali primerov premega govora. Pri drugem primeru (krepko je tiskana pojavnica, ki je bila označena med označevanjem) pa je dejansko uporabljena neskladenjska vejica, vendar lahko v tem primeru oznako problematiziramo, saj gre – kolikor lahko razberemo iz sobesedila – za navedbo odprtosti zaslonke fotografske leče, ki se tipično podaja s piko. Podobno kot pri nekaterih drugih kategorijah smo glede na dosedanje raziskave, izvedene na standardnem gradivu, pričakovali bistveno več primerov nestandardne rabe neskladenjske vejice.

4 Sklep

V prispevku smo predstavili rezultate empirične analize nestandardne rabe vejice v uporabniških spletnih vsebinah v slovenščini, raziskavo pa smo izvedli na naboru 500 tvitov. Glavni namen raziskave je bil začrtati nadaljnje raziskave stave vejice v slovenščini, zlasti v primerjavi s

standardnojezikovnim gradivom, v okviru te zasnove pa predvsem določiti, v kolikšni meri raba vejice na Twitterju odstopa od jezikovnega standarda.

Rezultati pretežno pritrjujejo dosedanjim raziskavam, obenem pa izpostavljajo nekaj novih težišč pri tej problematiki, in sicer lahko vidimo, da je nestandardna raba vejice na Twitterju vezana predvsem na skladenjsko rabo. Pri tem se kot najbolj problematični izkazujeta kategoriji pristavčnih struktur in odvisnikov (z manjkajočo vejico). Medtem ko lahko pri pristavčnih strukturah del razloga za pogostnost nestandardne stave iščemo pri strukturi Twitterja kot medija in pri novem razumevanju tovrstnih struktur, zlasti pastavkov, pa ugotavljamo, da odsotnost vejice pri predmetnih odvisnikih izhaja iz težav uporabnikov pri prepoznavanju skladenjskih struktur – to pa je v slovenščini univerzalna težava, ki ni vezana na interaktivnost ali formalnost medija. Raziskava je dala precej presenetljive rezultate z vidika rabe odvečne vejice, saj se ta pojavlja precej redkeje kot v primerljivih študijah na standardnojezikovnem gradivu, kar nakazuje na potrebo, da se pričujoča raziskava obogati še z analizo drugih stopenj standardnosti tvitov in drugih vrst uporabniških vsebin, zajetih v korpusu Janes, ter jih neposredno primerja s standardnojezikovnim gradivom. Glede na izsledke analize lahko sklenemo tudi, da bo tovrstne raziskave mogoče opraviti s tipologijo, predstavljeno v pričujočem prispevku.

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) in programa Mladi raziskovalec (37487), ki ju financira ARRS.

6 Literatura

Richard Eckart de Castilho, Chris Biemann, Iryna Gurevych in Seid Muhie Yimam. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. V: *Zbornik letne konference CLARIN (CAC) 2014*, Soesterberg, Nizozemska.

Tomaž Erjavec, Darja Fišer in Nikola Ljubešić. 2015. Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes. V: *Zbornik konference Slovenščina na spletu in v novih medijih*, str. 20–26, Ljubljana. Znanstvena založba Filozofske fakultete.

Tomaž Korošec. 2003. K pravilom za skladenjsko vejico v Slovenskem pravopisu 2001. *Slavistična revija*, 51(2): 247–266.

Iztok Kosem, Mojca Stritar, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012. *Analiza jezikovnih težav učencev: korpusni pristop*. Trojina, zavod za uporabno slovenistiko, Ljubljana.

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. V: *Zbornik konference RANLP 2015, 7.–9. september 2015*, str. 371–378, Hisar, Bolgarija.

Nataša Logar in Damjan Popič. 2015. Vejica: rezultati anketne raziskave med dijaki in študenti. *Jezikoslovni zapiski*, 21(2): 45–59.

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; Fakulteta za družbene vede Univerze v Ljubljani.

Damjan Popič. 2014. *Korpusnojezikoslovna analiza vplivov na slovenska prevodna besedila*. Doktorska disertacija, Filozofska fakulteta Univerze v Ljubljani.

Damjan Popič in Darja Fišer. 2015. Vejica je mrtva, živela vejica. V: *OBDOBJA 34: Slovnica in slovar – aktualni jezikovni opis*, str. 609–618, Ljubljana. Znanstvena založba Filozofske fakultete.

Tadeja Rozman, Mojca Stritar in Iztok Kosem. 2012. Šolar – korpus šolskih pisnih izdelkov. V: *Empirični pogled na pouk slovenskega jezika.* Trojina, zavod za uporabno slovenistiko, Ljubljana.

Slovenski pravopis. 2001. *Pravila.* Ur. Jože Toporišič et al. Založba ZRC in ZRC SAZU, Ljubljana.

Tina Verovnik. 2003. Vejica premalo, vejica preveč (2. del). *Pravna praksa*, 22(21): 51.

Živa Žibert. 2006. *Slovenska vejica: balast ali skladenjska nujnost slovenskega knjižnega jezika?* Diplomsko delo, Fakulteta za družbene vede Univerze v Ljubljani.

EAGLE − Medomrežje *Europeana* antične grške in latinske epigrafike. Digitalni dostop do antičnih napisnih spomenikov

Anja Ragolič

ZRC SAZU, Inštitut za arheologijo
Novi trg 2, 1000 Ljubljana
anja.ragolic@zrc-sazu.si

Povzetek

V prispevku je predstavljen evropski projekt *Eagle* – Medomrežje *Europeana* antične in grške epigrafike, katerega glavni cilj je bil zbrati in omogočiti širši javnosti dostop do latinskih in grških napisnih kamnov antičnega sveta. Nabor napisov je predstavljen v enotni bazi dostopni na spletu, z dvema mobilnima aplikacijama pa si lahko mimoidoči pridobi ključne informacije o najdbi na mestu hrambe spomenika. Epigrafska dediščina je z omenjenim projektom prvič postala dosegljiva ne samo strokovnjakom, ampak tudi laikom.

Eagle – *Europeana* Network of Ancient Greek and Latin Epigraphy. Digital Access of Ancient Inscriptions

The principal aim of the European Eagle project (*The Europeana network of Ancient Greek and Latin Epigraphy*) described in this article is to collect and present Greek and Latin inscriptions of the Ancient World to the broad public. The inscriptions are accessible within a single free user-friendly portal. Basic information about an ancient monument could also be uploaded with two Mobile Applications on a smart phone, at the place, where it is displayed. However, with the Eagle project is the epigraphic heritage for the first time accessible not only to scholars, but also to the curious.

1 Uvod

Antični napisi, ki so vklesani v kamen, vrezani v kovino ali druge materiale, predstavljajo pomemben historični vir, s katerimi si arheologi in zgodovinarji pomagajo pri interpretaciji preteklosti. Ti napisi so razpršeni po raznovrstnih korpusih in publikacijah, ki so mnogokrat dostopni le posameznikom. Ena glavnih težav je tudi razumevanje napisov na spomenikih v grškem ali latinskem jeziku, saj prevodi le-teh običajno manjkajo.

2 Kaj je *Eagle* in čemu služi?

Eagle (*The Europeana network of Ancient Greek and Latin Epigraphy*)[1], ustanovljen pod okriljem Evropske komisije, je evropski projekt namenjen zbiranju in predstavitvi napisov antičnega sveta, vklesanih na kamen ali zapisanih na kovino in druge materiale, ki so se ohranili do danes. Osrednji cilj projekta je bil povezati strokovnjake s področja klasične latinske in grške epigrafike, da bi skupaj uredili in vzpostavili spletno bazo s kar se da velikim številom antičnih napisnih spomenikov. V uporabniku enostavnem spletnem brskalniku je tako prvič na voljo več kot 1,5 milijonov metapodatkov in slikovnega gradiva s spremljajočimi komentarji o napisih, ki so razpršeni v 25 evropskih državah in ki predstavljajo pribl. 80 % vseh ohranjenih spomenikov Sredozemlja.

3 Konzorcij – vloga in sodelovanje posameznih ustanov

Konzorcij projekta je sestavljalo 19 partnerjev iz 13 evropskih držav. Med njimi je 14 partnerjev, predstavnikov evropskih univerz in raziskovalnih centrov, v projektu sodelovalo kot posredovalci vsebine napisov. Njihova glavna naloga je bila zbiranje, dopolnjevanje in posredovanje metapodatkov kot tudi digitalnega gradiva v eno od štirih spletnih baz (glej spodaj). Štirje tehnološki partnerji (CNR-ISTI, EUREVA, GOGATE in univerza Leuven) so skrbeli za tehnično podporo pri projektu, vzpostavljali infrastrukturo in spletno stran, posredovali metapodatke v Europeano, iskali znotraj posameznih baz dvojnice napisov in razvijali tehnološko podlago za razvoj aplikacij za pametne telefone. S podporo *Wikimedia Italia* so obstoječe zbirke napisov in prevodi postali dostopni širši javnosti. Projektni koordinator je bila rimska univerza La Sapienza, tehnični koordinator pa italijanska družba Promoter Srl.[2]

4 Elektronske baze znotraj projekta *Eagle*

Baze, tako elektronske kot takšne, ki jih raziskovalne ustanove in univerze izdelujejo za lastno uporabo, so bile temelj, na katerem je konzorcij projekta Eagle pričel graditi svojo vizijo dostopnosti do antičnih napisnih spomenikov. Med partnerji zadolženimi za vsebino so bile doslej na voljo štiri spletne baze: Epigraphic Database Roma – EDR,[3] Epigrafska podatkovna baza Bari,[4] epigrafska podatkovna baza Heidelberg - EDH[5] in Hispania Epigraphica online,[6] katerim pa so se kasneje priključili še drugi arhivi:

- Arachne[7]
- The British School at Rome digital collections[8]
- Archaia Kypriaki Grammateia Digital Corpus - Inscriptions/STARC collection[9]
- PETRAE[10]

[1] http://www.eagle-network.eu/

[2] Orlandi, Giberti in Satucci, 2014.

[3] http://www.eagle-network.eu/

[4] http://www.edb.uniba.it/

[5] http://edh-www.adw.uni-heidelberg.de/home

[6] http://eda-bea.es/

[7] http://arachne.uni-koeln.de/drupal/

[8] http://www.bsrdigitalcollections.it/

[9] http://www.eagle-network.eu/collections/archaia-kypriaki-grammateia-digital-corpus-inscriptionsstarc-collection/

[10] http://petrae.huma-num.fr/index.php/en/

- The Last Statues of Antiquity[11]
- UBI ERAT LUPA[12]

5 Delovne skupine v okviru projekta *Eagle*

Ideja projekta je bila poiskati pot med strokovnjaki za epigrafiko oz. klasične študije na eni strani ter turisti in drugimi laiki na drugi. Za dosego cilja so bile znotraj konzorcija ustanovljene tri delovne skupine.[13]

5.1 Obdelava vsebine in prevodi

Prva skupina je bila zadolžena za zbiranje podatkov o napisih, za posredovanje metapodatkov v baze ter za pisanje oz. dodajanje prevodov. *Wikimedia Italia*, italijansko združenje znotraj Wikimedie, je skrbelo za pripravo platforme, v kateri so se zbirali podatki in slikovno gradivo, ki jo je *Wikimedia Italia* posredovala *Wikimedia Commons*.

5.2 Usklajevanje, GIS in terminologija

Druga skupina je skrbela za kontinuirano pošiljanje metapodatkov, razvijala in opredelila uporabo enotnega besedišča in terminologije pri partnerjih zadolženih za vsebino ter razvijala postopke in pripomočke za ustrezno georeferenciranje spomenikov.[14]

5.3 IPR in uporabniki

Zadnja, tretja skupina, je analizirala in vrednotila vsebine, ki so se stekale v skupno platformo z zornega kota uporabnikov. Njena naloga je bila raziskati, kdo so uporabniki vsebin in čemu. Ugotovila je, da uporabniki antičnih napisnih kamnov prihajajo iz raznovrstnih skupin: laiki, otroci, turisti, posamezniki, ki se zanimajo za kulturno dediščino in/ali filologijo, akademiki, epigrafiki. Ker je bila objava slikovnega gradiva ena od temeljnih ciljev (in zahtev) projekta, pravice avtorjev fotografij kot lastnikov spomenikov pa se od države do države razlikujejo, so bile raziskave obstoječih opredelitev intelektualnih pravic na začetku projekta ključen podatek za nadaljnje zbiranje in posredovanje podatkov v repozitorij in kasneje v *Europeano*. To je bila še ena od nalog te skupine.[15]

6 Tehnološka plat projekta

Partnerji, zadolženi za vsebine, so posredovali naslednje podatke: natančen kraj najdbe (antično in moderno ime) in hrambe spomenika/najdbe, njegove/njene mere, vrsta uporabljenega materiala, tip spomenika (nagrobnik, oltar, miljnik, mejnik …), jezik v katerem je zapisan napis, prepis napisa (ali fragmenta napisa, ki je na spomeniku ohranjen) ter datacija. Podatki so se zbirali v štirih osnovnih bazah. Metapodatki spomenikov iz Emone, s katerimi se je projektu priključil ZRC SAZU, so se vpisovali v repozitorij epigrafske baze v Rimu, saj je Emona (današnja Ljubljana) v cesarskem obdobju sodila v upravno območje Desete italske regije (*regio X. Venetia et Histria*).

Na spletni strani projekta Eagle je brskalnik napisov dostopen širši javnosti. Uporabnik lahko izbira med osnovnim iskalnikom po ključnih besedah, ki se pojavljajo v napisu, ali pa uporabi napredni iskalnik, znotraj katerega je Eagle razvil sedem iskalnih polj. Ta so povezana s ključnimi podatki o antičnem spomeniku: tip napisa, tip objekta, material, napis, okras, hramba in datacija najdbe. Uporabnik lahko svoje iskalne nize shrani v »osebno bazo«, s čimer se izogne ponovnemu iskanju. Shranjene zadetke lahko opremi s komentarji in pripombami.[16]

6.1 Mobilni aplikaciji

Dve mobilni aplikaciji, ki so jih razvijali italijanski partnerji iz skupine EUREVA s podporo CNR-ISTI, ne služita samo boljšemu dostopu do informacij o spomeniku, ampak hkrati težita k promociji projekta *Eagle*. Uporabniki, zlasti turisti, lahko z mobilnima aplikacijama dostopajo do osnovnih informacij o spomeniku na kraju hrambe.

6.1.1 Osnovna/Vodilna mobilna aplikacija

T. i. *Flagship Mobile Application* dovoljuje uporabnikom pametnih telefonov (Androidi, iPhone, telefoni z windowsi), da se spomenik fotografira, aplikacija prepozna fotografiran spomenik znotraj podatkovne baze, v odgovor pa uporabniki prejmejo iz Eaglovega strežnika vse osnovne informacije o najdbi: prevod, okoliščine odkritja, sorodne tipe napisa, turistične informacije, … Aplikacija ni primerna samo za turiste, ampak tudi za zgodovinarje, epigrafike, arheologe, da z obiskom muzejev in/ali arheoloških najdišč dobijo osnovne informacije o novoodkritih spomenikih.[17]

6.1.2 Aplikacija z zgodbami

Ker smo partnerji projekta Eagle težili k temu, da bi k ogledu napisov pritegnili čim širši krog uporabnikov spletne baze in aplikacij, ki bi jim lahko bili v pomoč pri ogledih, študijah in projektih, smo skušali razviti različne strategije, s katerimi bi zadostili raznolikim potrebam uporabnikov. Na tem mestu se je pričela razvijati t. i. *Storytelling Application*. Zavedali smo se, da se zlasti učitelji pri didaktični predstavitvi spomenikov poslužujejo

[11] http://www.ocla.ox.ac.uk/statues/team.shtml
[12] http://www.ubi-erat-lupa.org/simplesearch.php
[13] http://www.eagle-network.eu/about/working-groups/
[14] http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D2.2.1_Content-harmonisation-guidelines-including-GIS-and-terminologies.pdf; http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D2.2.2_Content-harmonisation-guidelines-including-GIS-and-terminologies-Second-Release.pdf
[15] http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D2.3.1_Best-practices-on-user-engagement-with-epigraphic-content-including-IPR-requirements.pdf; http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D2.3.2_Best-practices-on-user-engagement-with-epigraphic-content-including-IPR-requirements-Second-Release_v3.2.pdf
[16] http://www.eagle-network.eu/resources/search-inscriptions/.
[17] http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D5.3.1_First-release-of-the-flagship-mobile-application-and-SDK_v1.0.pdf; http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D5.3.2_Second-release-of-the-flagship-mobile-application-and-SDK_v1.0.pdf

opisne in pripovedne metodologije. Spomenike, ki so si bili po vsebini in napisu sorodni, smo tako povezali v večje skupine: npr. napise z omembo osvobojencev, napise, ki omenjajo cesarjeva gradbena dela, napise, ki omenjajo visoko starost pokojnikov itd. Uporabnikom smo skušali približati preteklost, da bi se seznanili z rimsko kulturo in zgodovino in razumeli, kako si je treba razlagati navedbe stoletnikov na napisih v rimskem imperiju in kje se najdejo takšni spomeniki, kakšen je bil položaj osvobojenca in ob kakšni priložnosti je cesar postavil javno zgradbo. Zgodbe, ki smo jih partnerji pošiljali skupaj z zemljevidom in osnovnimi podatki o spomeniku, so dostopni tudi na spletni strani projekta Eagle.[18]

Odziv na aplikacijo je bil izjemen. Poleg publikacije[19] in razstave, ki sta nastali na podlagi omenjenih zgodbic (za virtualno razstavo glej spodaj), smo v zadnjem letu projekta (2015) v branje napisov in pisanje zgodb skušali pritegniti tudi uporabnike. Rezultati in nagrajenci prvega tekmovanja v pisanju zgodb so bili razglašeni na zadnji konferenci Eagla, ki se je odvijala med 27. in 29. januarjem 2016 v Rimu (glej spodaj). Zmagovalne tri zgodbe si je mogoče prebrati v manjši knjižici, dostopni na strani projekta Eagle v .epub ali .pdf obliki.[20]

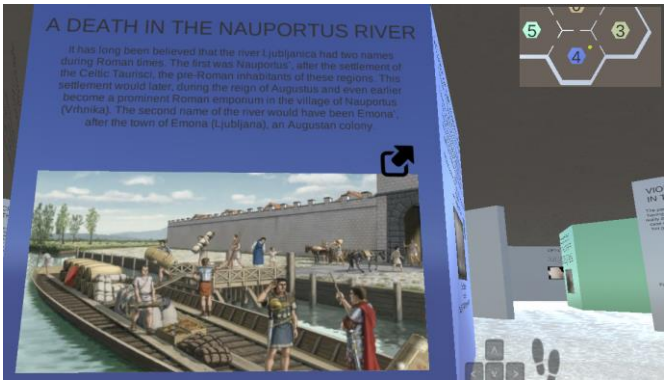
7 Konference in delavnice, organizirane v okviru projekta *Eagle*

V okviru projekta je bilo organiziranih osem mednarodnih dogodkov in srečanj. Konzorcij projekta se je sestal 2. aprila 2013 v Rimu, da bi se dogovoril o poteku dela in prvih korakih. Prva mednarodna delavnica, ki so se je lahko udeležili tudi zunanji udeleženci, je ZRC SAZU organiziral 19. in 20. februarja 2014 v Narodnemu muzeju Slovenije v Ljubljani. Z raznolikimi temami in vodstvom dr. Marjete Šašel Kos po lapidariju Narodnega muzeja Slovenije, so se udeleženci seznanili s projektom, s pravnimi določili posameznih državi pri objavi in fotografiranju spomenikov, s problemi pri zajemanju in posredovanju metapodatkov, z medmrežnimi bazami, s praktičnim prikazom fotografiranja spomenikov lapidarija Narodnega muzeja Slovenije pa je Ortolf Harl opozoril na prednosti in slabosti digitalnega dokumentiranja. Prikaz dobrih praks in poudarjanje pomena digitalne tehnologije pri ohranjanju in promociji kulturne dediščine, sta bili temi delavnice v Rimu, 16. maja 2014. Prikaz uporabe Epi-doca, programa, ki se uporablja pri digitalnih objavah napisov in na katerem gradijo vse večje epigrafske spletne baze (tudi Eagle), je bil predstavljen v Bologni 29. maja 2014. Na mednarodni konferenci v Parizu, ki je potekala med 29. septembrom in 1. oktobrom, je bil poudarjen pomen medsebojnega sodelovanja med partnerji in zunanjimi ustanovami, ki so se projektu lahko priključile kot zunanji partnerji. Ponovno so bila predstavljena pravna določila posameznih držav in različne skupine uporabnikov rimskih epigrafskih spomenikov. Peti mednarodni dogodek v Nikoziji, 11. in 12. marca 2015, je bil posvečen digitalnemu zajemanju in širjenju napisov, šesti v Bariju, med 23. in 25. septembrom, pa napisom, ki se pojavljajo na raznovrstnih podlagah, v drugih jezikih kot grščini in latinščini. Izpostavljeni so bili negrafični znaki in poudarjena vloga upodobitev na spomenikih. Na zadnji mednarodni konferenci, ki se je odvijala med 27. in 29. januarjem v Rimu, so bili predstavljeni zadnji rezultati projekta, prikazana virtualna razstava in izpostavljeni nekateri zanimivi primeri napisov iz Italije.[21]

8 Virtualna razstava

V virtualni razstavi posvečeni epigrafiki in dostopni preko Eagle portala, so zbrani izstopajoči napisi, ki so jih partnerji zadolženi za vsebine posredovali *Europeani*. Razstava je bila ustvarjena z namenom, da bi se izbrani pomembni napisi predstavili široki javnosti na zanimiv način, hkrati pa skušali podati informacije o tem, kaj vse lahko iz napisov izvemo, kakšna vrsta napisov in spomenikov se proučuje in kakšne zgodbe nam lahko ti spomeniki preteklosti posredujejo.

Razstava ima obliko panja, sestavljeno iz sedmih sob šestkotne oblike. Iz osrednje sobe, ki je locirana v sredini, je mogoče vstopiti v katerokoli sobo okoli nje. Vsaka soba je posvečena določenemu področju epigrafike: uvod (soba 0), napisi in zgodovina (soba 1), pisava (soba 2), objekti in povezave med upodobitvami, napisom in kontekstom (soba 3), v napisih izražena čustva (soba 4), delo kamnosekov, metode pridobivanja kamnine in napake na napisih (soba 5) in digitalno tehnološka orodja pomembna za epigrafiko (soba 6).[22]



Sl. 1: Primer razlagalne table v virtualni razstavi. V desnem zgornjem kotu je shema »panja« in označena soba, v kateri se napis nahaja (http://webgl-eagle.d4science.org/).

9 Vloga Inštituta za arheologijo, ZRC SAZU v okviru projekta *Eagle*

Inštitut za arheologijo ZRC SAZU je v projektu sodeloval kot partner zadolžen za vsebino. Njegova naloga je bila zbrati, obdelati in posredovati 400 metapodatkov (napisov in fotografij). Geografsko so sem sodili napisi, ki so bili odkriti v Ljubljani (antični Emoni), ter napisi iz njenega upravnega območja. Napisi iz Emone in njenega teritorija so vklesani v kamen in vrezani v druge materiale: pasne okove, svečnike, posode, svinčene etikete (*tesserae*), mozaike … Končna številka posredovanih podatkov iz Emone je presegla ciljno

[18] Mambrini, 2014: 36–39.
[19] Šašel Kos, 2015.
[20] http://www.eagle-network.eu/wp-content/uploads/2016/02/EAGLE-Storytelling-App-EAGLE-project.pdf
[21] http://www.eagle-network.eu/about/events/
[22] http://webgl-eagle.d4science.org/

številko (v *Europeano* je bilo posredovanih 569 enot) (sl. 2).[23]

Epigraphic Archives of Slovenia (ZRC-SAZU)						
Date	D-Net		Total	Europeana		
	artefact	visual		CHO	WebResources	Total
Mar	269	300	569	269	76	345

History and comments:

- The ZRC-SAZU content due to the MS16 is of 400 items.
- At March ZRC-SAZU sent to EAGLE 269 artefacts and 300 visual representations for a total of 569 items.
- In Europeana (Europeana ID: 2058817) have been published 269 artefacts and 300 visual representations.
- The CP reached and overcame the amount of data declared in the DoW.

Fig. 5.9 shows one of the important inscriptions provided by ZRC-SAZU to the EAGLE portal and to Europeana.

Sl. 2: Končno poročilo zbranih napisov in digitalnih podatkov, ki jih je ZRC SAZU prispeval v okviru *Eagla* (prevzeto po: http://www.eagle-network.eu/wp-content/uploads/2016/04/EAGLE_D3.3.4_Report-on-the-contributions-to-Europeana_v1.0.pdf).

10 Zaključek projekta in ustanovitev združenja IDEA

S 30. marcem 2016 se je projekt *Eagle* zaključil. Konzorcij se je odločil, da z delom nadaljuje, zato so spletna stran projekta z vsemi dokumenti, iskalnikom napisov, novicah o preteklih dogodkih in povezava na virtualno razstavo še vedno dosegljivi. Prav tako se lahko v sklopu *Wikipedia Commons* še vedno posredujejo prevodi napisov v moderne jezike za antične spomenike, ki so že vključeni v *Europeano*.

9. maja 2016 je skupina partnerjev Eagla ustanovila združenje imenovano IDEA – *The International Digital Epigraphy Association* (Mednarodno digitalno epigrafsko združenje), da bi ohranili, dopolnjevali in vzdrževali izkušnje in znanja pridobljena s projektom Eagle. Cilj združenja je bil promocija uporabe naprednih metodologij v proučevanju in objavljanju napisnih spomenikov, začenši z antičnimi, da bi s tem povečali vedenje o napisih tako pri strokovnjakih kot laikih. Prvo srečanje združenja je načrtovano 28. septembra v Pisi, na katerem se želi združenje predstaviti širši publiki.[24]

11 Sklepna beseda

Cilj sodelovanja vrhunskih strokovnjakov s področja epigrafike in humanistike v projektu Eagle je bil ustvariti osrednji repozitorij antičnih napisov Sredozemlja, ki bi bil dostopen javnosti različnih profilov: strokovnjakom, turistom, učiteljem, otrokom... Napisi so postali dostopni v evropski digitalni knjižnici *Europeana*, kjer je bilo ob zaključku projekta vključenih skoraj 374.000 enot novih napisov in fotografij. Sprotni rezultati in spoznanja so bili predstavljeni na številnih mednarodnih delavnicah in publikacijah. Z razvojem dveh aplikacij za pametne telefon in z vzpostavitvijo virtualne razstave pa je triletno delo poseglo tudi na področje digitalne humanistike.

[23] http://www.eagle-network.eu/wp-content/uploads/2016/04/EAGLE_D3.3.4_Report-on-the-contributions-to-Europeana_v1.0.pdf.

[24] http://www.eagle-network.eu/founded-idea-the-international-digital-epigraphy-association/.

12 Literatura

Francesco Mambrini. 2014. La Flagship Storytelling Application di Eagle. V: *Forma urbis*, n. 1, str. 36–39.

Silvia Orlandi, Luca Marco Carlo Giberti in Raffaella Santucci. 2014. EAGLE: Europeana Network of Ancient Greek and Latin Epigraphy. Making the Ancient Inscriptions Accessible. V: *Lexicon Philosophicum, International Journal for the History of Texts and Ideas*, str. 315–326.

Proceedings of the First EAGLE International Conference "*Information Technologies for Epigraphy and Cultural Heritage*" (Paris, 29 September – 1 October 2014) http://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf

Marjeta Šašel Kos. 2015. *The Disappearing Tombstone and other Stories from Emona* (*Izginjajoči nagrobnik in druge zgodbe iz Emone*). ZRC SAZU, Ljubljana.

http://www.eagle-network.eu/

http://www.edb.uniba.it/

http://edh-www.adw.uni-heidelberg.de/home

http://eda-bea.es/

http://arachne.uni-koeln.de/drupal/

http://www.bsrdigitalcollections.it/

http://www.eagle-network.eu/collections/archaia-kypriaki-grammateia-digital-corpus-inscriptionsstarc-collection/

http://petrae.huma-num.fr/index.php/en/

http://www.ocla.ox.ac.uk/statues/team.shtml

http://www.ubi-erat-lupa.org/simplesearch.php

http://www.eagle-network.eu/about/working-groups/

http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D2.2.1_Content-harmonisation-guidelines-including-GIS-and-terminologies.pdf;

http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D2.2.2_Content-harmonisation-guidelines-including-GIS-and-terminologies-Second-Release.pdf.

http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D2.3.1_Best-practices-on-user-engagement-with-epigraphic-content-including-IPR-requirements.pdf;

http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D2.3.2_Best-practices-on-user-engagement-with-epigraphic-content-including-IPR-requirements-Second-Release_v3.2.pdf.

http://www.eagle-network.eu/resources/search-inscriptions/.

http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D5.3.1_First-release-of-the-flagship-mobile-application-and-SDK_v1.0.pdf;

http://www.eagle-network.eu/wp-content/uploads/2013/06/EAGLE_D5.3.2_Second-release-of-the-flagship-mobile-application-and-SDK_v1.0.pdf.

http://www.eagle-network.eu/wp-content/uploads/2016/02/EAGLE-Storytelling-App-EAGLE-project.pdf.

http://www.eagle-network.eu/about/events/.

http://webgl-eagle.d4science.org/.

http://www.eagle-network.eu/wp-content/uploads/2016/04/EAGLE_D3.3.4_Report-on-the-contributions-to-Europeana_v1.0.pdf.

http://www.eagle-network.eu/founded-idea-the-international-digital-epigraphy-association/.

ARIADNE: povezani odprti podatki (LOD) v praksi

Benjamin Štular,* Franco Niccolucci,† Julian Richards‡

* ZRC SAZU Inštitut za arheologijo
Novi trg 2, 1000 Ljubljana
bstular@zrc-sazu.si
† PIN S.c.r.l - Polo Universitario "Città di Prato"
Piazza Giovanni Ciardi 25, 59100 PRATO (PO)
franco.niccolucci@unifi.it
‡ Department of Archaeology, University of York
The King's Manor, York YO1 7E
julian.richards@york.ac.uk

Povzetek

V prispevku so prikazane izbrane spletne storitve projekta "Napredna raziskovalna infrastruktura za arheološke podatkovne mreže v Evropi" (ARIADNE), ki je financiran v okviru sedmega okvirnega programa evropske skupnosti. Osrednja storitev je "ARIADNE portal", ki deluje ne samo kot glavna vstopna točka za iskanje in pregledovanje, temveč tudi kot platforma za objavljanje in analizo raziskovalnih arheoloških podatkov. Pri načrtovanju portala smo za raziskavo uporabljenih podatkovnih standardov in metapodatkovnih shem izdelali metapodatkovni register. Predstavljamo še spletne storitve za objavljanje kompleksnih vizualnih medijev: trirazsežni modeli, dvoinpolrazsežni modeli in slike visoke ločljivosti.

ARIADNE: Linked Open Data (LOD) in practice

This paper presents web services developed within the "Advanced Research Infrastructure for Archaeological Dataset Networking in Europe" (ARIADNE). The project is funded under the European Community's Seventh Framework Programme. The web services are not the only result of the project but are among the most important ones. The first to be mentioned is the ARIADNE portal providing the main point of access for, on the one hand, searching and browsing and, on the other hand, processing and publishing archaeological datasets online. As a supporting tool ARIADNE metadata registry has been used to survey currently used data standards and metadata schemas. In addition ARIADNE Visual Media Services have been developed to provide easy publication and presentation on the web of complex visual media assets. The following services are presented: 3D models, RTI images and High-resolution images.

1 Uvod

Dandanes je arheologom dostopna ogromna količina digitalnih podatkov, ki segajo prek različnih arheoloških obdobij, raziskovalnih področij in regij. Količina teh podatkov z uporabo informacijskih tehnologij strmo narašča. Gre za zbir rezultatov dela posameznikov in institucij, ki je razdrobljen ter nehomogen in zato težko dostopen. Podatki so raztreseni v različnih podatkovnih zbirkah, težko dostopnih strokovnih poročilih (t. i. sivi literaturi) in v znanstvenih objavah. Slednje so še vedno najpomembnejše sredstvo za objavljanje rezultatov raziskav (Selhofer in Geser, 2015).

Cilj projekta "Napredna raziskovalna infrastruktura za arheološke podatkovne mreže v Evropi" (ang. *Advanced Research Infrastructures for Achaeological Dataset Networking in Europe*, dalje ARIADNE) je združiti in povezati arheološke podatkovne zbirke v enotno raziskovalno infrastrukturo. Hkrati gojimo kulturo prostega dostopa in ponovne uporabe podatkov (Richards, 2012; Niccolucci in Richards, 2013a; Niccolucci in Richards, 2013b; Aspöck in Geser, 2014).

2 Namen prispevka

Namen prispevka je predstaviti izbrane spletne storitve projekta ARIADNE. Med ostalimi rezultati projekta, ki jih na tem mestu ne bomo predstavljali, velja izpostaviti analizo potreb uporabnikov (Selhofer in Geser, 2015), pregled metapodatkovnih standardov (Ronzino et al., 2013a), pregled politik dostopa do podatkov (Fernie, 2014), priročniki za dobre prakse (Niven in Wright, 2014) in raziskave na področju podatkovnega rudarjenja (Wilcke, 2015) ter procesiranja naravnih jezikov (Vlachidis et al., 2015).

V prispevku prikazujemo spletne storitve s stališča končnega uporabnika. Tehnične rešitve so (glej navedeno literaturo v nadaljevanju) in bodo podrobneje predstavljene na drugih mestih.

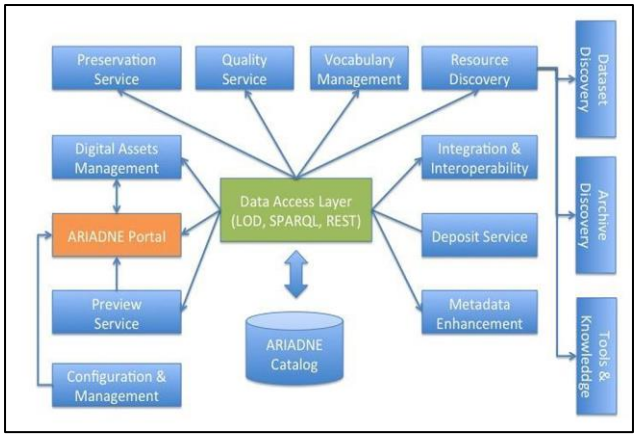
3 Spletne storitve ARIADNE

Osrednja spletna storitev projekta je *ARIADNE portal*. Portal ima dvojen namen. Na eni strani je glavna vstopna točka za iskanje in pregledovanje arheoloških podatkov; na drugi strani deluje kot platforma za objavljanje in analizo raziskovalnih podatkov. Pri načrtovanju portala smo za raziskavo obstoječih podatkovnih standardov in metapodatkovnih shem razvili metapodatkovni register. Predstavljamo še spletne storitve za objavljanje izbranih kompleksnih vizualnih medijev: trirazsežni modeli, dvoinpolrazsežni modeli in fotografije visoke ločljivosti.

Slika 1 predstavlja procesogram storitev izdelanih v okviru projekta.[1] Pri načrtovanju so bili uporabljeni podatki iz kataloga ARIADNE (ang. *ARIADNE Catalog*), do katerega so načrtovalci dostopali preko sloja za dostop podatkov (ang. *Data Access Layer*). Vnos metapodatkov je omogočal metapodatkovni register; s pomočjo le-tega so registrirani uporabniki (projektni partnerji in vsi zainteresirani) opisali lastne podatke in metapodatke po

[1] Prispevek pišemo v času trajanja projekta, zato do zaključka projekta lahko pride do manjših odstopanj.

shemi ACDM (glej dalje). Z zbranimi metapodatki upravlja temu namenjena storitev (ang. *Digital Assets Management*). Storitev za odkrivanje in pridobivanje podatkov (ang. *Resource Discovery Service*), predvsem indeksiranje in priklic podatkov, uporabnikom omogoča dostop do podatkovnih virov in integriranega ogleda preko portala. Seznam slovarjev in tezavrov, ki so kartirani v SKOS, je vzdrževan v upravljalniku slovarjev (ang. *Vocabulary Management*). Storitev krepitve metapodatkov (ang. *Metadata Enhancement Service*) – npr. rudarjenje povezav ter avtomatsko povezovanje s slovarji in tezavri - omogoča avtomatsko krepitev metapodatkov, ki so v ACDM (Wright, 2014).



Slika 1: Procesogram spletnih storitev ARIADNE.

3.1 Podatkovni model kataloga (ACDM)

Cilj projekta ARIADNE je torej povezati arheološke raziskovalne podatke vseh organizacij in posameznikov, ki to želijo. To seveda pomeni povezati številne raznorodne metapodatkovne sheme, slovarje in tezavre (Ronzino et al., 2013a).

Prvi izziv pri tem je bil razviti globalno shemo v obliki formalne ontologije, ki bo omogočila združevanje podatkov brez izgube pomenov in jo bo možno nadgrajevati (Felicetti, 2014; Doerr, 2014). To formalno ontologijo smo poimenovali referenčni model ARIADNE (ang. *ARIADNE Reference Model*; Spletni vir 1; Aloia et al., 2015). V izhodišču smo se oprli na CIDOC konceptualni referenčni model (prim. Doerr in Schaller, 2008; LeBoeuf et al., 2015), ki smo ga nadgradili za potrebe arheoloških podatkov. Gre za izjemno kompleksno področje, ki v projektu ARIADNE je in bo še predmet več specializiranih znanstvenih člankov (npr. Geser in Niccolucci, 2012; Ronzino et al., 2013b; Felicetti et al., 2013; Amico et al., 2013; Aloia et al., 2014; Masur et al., 2014; Ronzino et al., 2016).

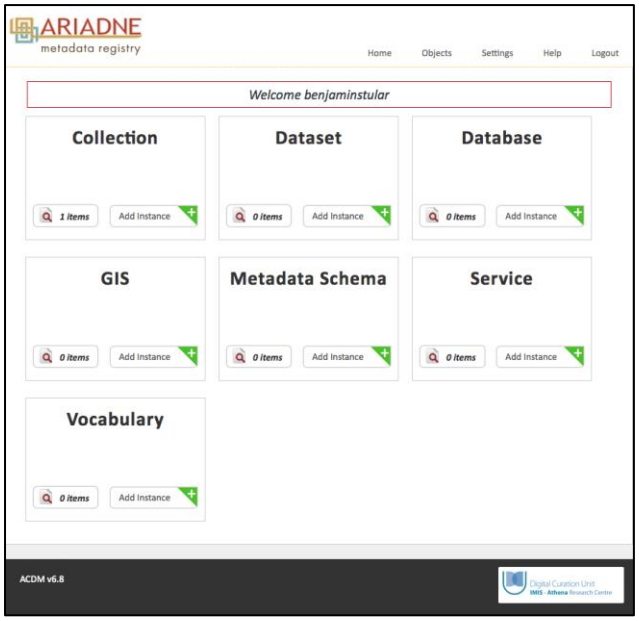
Kot naslednji nujen korak smo v projektu razvili podatkovni model, ki predstavlja arheološke podatkovne vire (ang. *ARIADNE Catalog Data Model*, dalje ACDM). ACDM tehnologija integracije podatkov temelji na skupnih lastnostih obstoječih podatkovnih zbirk in je eden pomembnejših dosežkov projekta..

V postopku izdelave ACDM smo najprej razvili spletno orodje, s katerim smo zbrali metapodatke projektnih partnerjev in podatke o obstoječih shemah, slovarjih in formatih. To orodje je metapodatkovni register (ang. *ARIADNE metadata registry*; Slika 2).

Na zbranih podatkih temelji ACDM, nadgradnja slovarja podatkovnega kataloga (ang. *Data Catalog Vocabulary*, dalje DCAT). V ARIADNE smo DCAT slovar izbrali zaradi zasnove, saj je primeren za opis vladnih podatkovnih katalogov, kot na primer Data.gov and data.gov.uk (Maali in Ericson, 2014). Dodatna razloga sta, ker DCAT nudi možnosti ponovne uporabe in predvsem ker je to priporočeno orodje za objavljanje odprtih podatkov (ang. *Open Data*). S tem ima ARIADNE odlično izhodišče za objavljanje odprtih podatkov. V ta namen sledimo tudi priporočilom DCAT-AP (Evropska skupnost, 2015) glede uporabe DCAT ontologije o tem, kateri atributi ali razredi so obvezni. Vendar, ker DCAT v ARIADNE zaenkrat uporabljamo kot interni standard, priporočilom ne sledimo v popolnosti. Tako v ACDM uporabljamo naslednja imenska mesta (ang. *namespaces*):

- dcat: http://www.w3.org/ns/dcat#
- dct: http://purl.org/dc/terms/
- dctype: http://purl.org/dc/dcmitype/
- foaf: http://xmlns.com/foaf/0.1/
- rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
- rdfs: http://www.w3.org/2000/01/rdf-schema#
- skos: http://www.w3.org/2004/02/skos/core#
- xsd: http://www.w3.org/2001/XMLSchema#.

Osrednji element modela je razred *ArchaeologicalResource* katerega vrednosti so podatkovni viri, opisani v katalogu. Zelo pomemben razred modela je tudi metapodatkovni format, *MetadataFormat*, ki vrednosti - formati različnih virov – prav tako črpa iz omenjenega kataloga (Papatheodorou et al. 2013; Gavrilis in Papatheodorou, 2014).



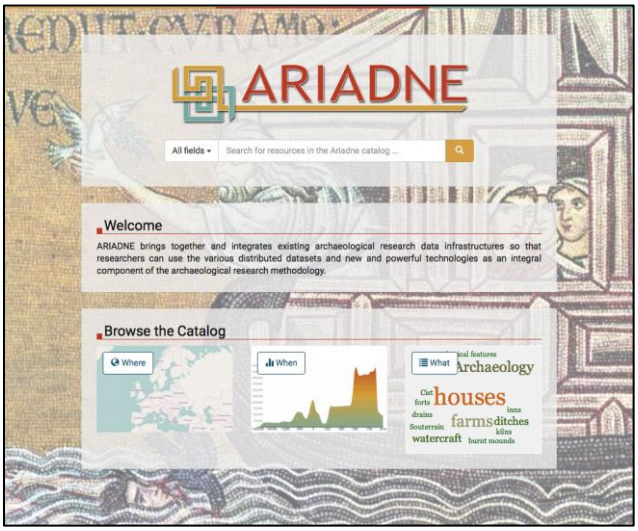
Slika 2: ARIADNE metapodatkovni register, pogled na začetni meni za izbor podatkovnega tipa.

3.2 Portal ARIADNE

Osrednja storitev je portal ARIADNE (http://portal.ariadne-infrastructure.eu). Pri načrtovanju smo se naslonili predvsem na izsledke poglobljene raziskave potreb uporabnikov (Selhofer in Geser, 2015). Sledeč izraženim potrebam je bil naš cilj razviti enotno, globalno dostopno točko, ki bo imela vlogo posredovanja med ponudniki in odjemalci podatkovnih virov. Na eni strani ponudniki lahko registrirajo svoje podatkovne vire, oziroma zelo natančne opise le-teh, sledeč ontologiji ACDM. S tem močno povečajo možnost, da bodo njihovi podatki ob relevantnem poizvedovanju odkriti (ang. *discoverability*). Na drugi strani odjemalci lahko brskajo ali opravljajo strukturirane in nestrukturirane poizvedbe po raznorodnih podatkih v enotnem ARIADNE okolju. Portal omogoča splošno iskanje, iskanje po času oziroma arheoloških obdobjih, po prostoru in iskanje po temah (Slika 3).

Namen portala ni ustvariti vseobsegajoč centraliziran repozitorij podatkov o arheološki dediščini, saj to pogosto ni združljivo s politikami posameznih institucij ali z zakonodajami posameznih držav. Namen je raziskovalcem olajšati iskanje in dostop do podatkov. Po izvedenem iskanju tako uporabnik dostopa do rezultatov poizvedb - podatkov ali storitev - na način, kot ga omogoča posamezni ponudnik (Meghini, 2014; prim. Hollander in Hoogerwerf, 2014). Najpogostejša sta dva načina: neposredna povezava do izbranega podatka ali povezava na vstopno točko partnerja.

Portal že sedaj omogoča dostop do podatkov vseh štiriindvajsetih projektnih partnerjev (Ronzino et al., 2013a) in večine izmed štirinajstih pridruženih partnerjev. Med partnerji projekta so mnogi izmed največjih evropskih arheoloških institucij, zato portal že v času pisanja prispevka vsebuje več kot 1,8 milijona enot.



Slika 3: ARIADNE portal, slika začetne strani.

3.3 Storitve za objavljanje vizualnih medijev

Izčrpna raziskava potreb uporabnikov (Selhofer in Geser, 2015) in delo skupine ekspertov (Scopigno, Dellepiane 2013) sta že na začetku projekta zaznala potrebo po spletni storitvi za objavljanje kompleksnih vizualnih medijev. Seveda so dejanske storitve, ki smo jih v projektu razvili, presek želja uporabnikov in razpoložljive tehnologije. Na tem mestu predstavljamo naslednje storitve: trirazsežni (dalje 3R) modeli, dvoinpolrazsežni modeli in fotografije visoke ločljivosti (http://visual.ariadne-infrastructure.eu).

Storitev objavljanja vizualnih medijev temelji na tehnologiji 3R spletne predstavitve dediščine (ang. *3D Heritage Online Presenter*, dalje 3DHOP). 3DHOP je okvir (ang. *framework*) za napredno spletno vizualizacijo 3R vsebin v visoki ločljivosti, ki je prilagojen potrebam kulturne dediščine. Razvoj okvira je temeljil predvsem na naslednjih treh zahtevah: enostavnost uporabe, učljivost in učinkovitost. Zato je 3DHOP zasnovan modularno, kar omogoča prilagodljivost različno zahtevnim uporabnikom in/ali načinom uporabe, odvisno od predhodnega znanja posameznika. 3DHOP je napisan v programskem jeziku JavaScript in temelji na SpiderGL knjižnicah, ki uporabljajo WebGL podskupine HTML5 jezika. Vse to omogoča prikazovanje 3R in drugih vsebin na večini modernih spletnih brskalnikov brez dodatnih vtičnikov (Potenziani idr. 2014; Potenziani idr. 2015).

3.3.1 3R modeli

Storitev 3R modeli (ang. *3D models*) je nastala iz potrebe, ki smo jo prepoznali raziskovalci (Scopigno in Dellepiane 2013). Že nekaj let število 3R modelov v arheologiji strmo narašča, uporaba v raziskovalne namene (npr. Aspöck in Fera, 2015) pa je komajda zaznavna. Ena temeljnih preprek je nezmožnost objavljanja 3R modelov v visoki ločljivosti na način, ki bi arheologu - nespecialistu za 3R tehnologije, ki nima namenske programske in strojne opreme - omogočil delo. Večidel neuspešni poizkusi z različnimi platformami (npr. Štular, 2012; Štular et al., 2013; Štular in Štuhec, 2015) so pokazali, da je edina dolgoročna rešitev spletna storitev, ki deluje brez vtičnikov.

V okviru projekta ARIADNE smo leta 2014 takšno storitev predstavili med prvimi na svetu. Naša storitev 3R modeli do danes ostaja edina tovrstna storitev, ki je popolnoma brezplačna, odprta in pri uporabi katere vse avtorske pravice ostanejo pri kreatorju (Slika 4).

Slika 4: ARIADNE Visual Media Services, 3R model kipa "Orator" iz 2. st. pr. n. št. (http://visual.ariadne-infrastructure.eu/3d/arringatore).

3.3.2 Dvoinpolrazsežni modeli

Storitev *RTI images* omogoča ogled modelov izdelanih s tehnologijo imenovano računalniško upodabljanje pretvarjanja odbojnosti na podlagi slikovnega gradiva (ang. *reflectance transformation imaging*, krajše RTI) s pomočjo polinomskih teksturnih preslikav (ang. *polynomial texture mapping*, krajše PTM). Gre za skupek postopkov računalniške grafike in obdelave fotografij, s pomočjo katerih izdelamo posebne vrste sliko z informacijami o svetilnosti predmeta glede na položaj vira svetlobe. Nova podoba je sestavljena iz množice fotografij obravnavanega predmeta, ki je bil vsakič osvetljen iz drugega položaja (Štuhec, 2012). Skupno ime za izdelke te in podobnih tehnologij je dvoinpolrazsežni modeli.

Podpora projekta ARIADNE tej tehnologiji je še posebej pomembna zato, ker se tehnologija komercialno ni uveljavila; ostaja nišna tehnologija, ki je zelo pomembna na primer na področju epigrafike in numizmatike (Slika 5).

Slika 5: ARIADNE Visual Media Services, dvoinpolrazsežni model avera bizantinskega novca (http://visual.ariadne-infrastructure.eu/rti/testnomprojetmj30052016_02).

3.3.3 Visokoločljive slike

Storitev visokoločljive slike (ang. *High-resolution images*) je, kot pove ime, namenjena objavljanju slik visoke ločljivosti. V prvi vrsti je storitev namenjena objavljanju visoko ločljivih fotografij, ki so posnete s posebnimi fotoaparati in se v kulturni dediščini uporabljajo za dokumentiranje. Ker tovrstna fotografija lahko presega velikost sto megapikslov so objavljene skoraj vedno v izsekih ali v zmanjšani ločljivosti.

Hkrati se je storitev izkazala kot izhodišče za inovativne načine uporabe. Tako smo jo na primer uspešno uporabili za objavo vizualizacije digitalnega modela reliefa, ki je bil izdelan iz lidarskih podatkov (Slika 6).

4 Zaključek

Projekt ARIADNE arheološkim raziskovalcem prinaša več spletnih raziskovalnih orodij, o katerih so ob začetku projekta leta 2011 razmišljali le vizionarji. Zagotovo je najvidnejši rezultat portal ARIADNE, ki bo sprva zagotovo tudi najbolj obiskan. Dolgoročno pa je morda še pomembnejši dosežek ACDM, orodje, ki v trenutno poplavo podatkovnih in metapodatkovnih standardov prinaša praktično rešitev, uporabno takoj in za vsakogar. Je odgovor na klic arheološke skupnosti "manj standardov in več standardizacije"! ACDM je zagotovilo za rast portala ARIADNE tudi po izteku projekta.

Zagotovo bodo v razvoju arheologije imele pomembno vlogo tudi storitve ARIADNE za objavljanje vizualnih medijev. Med tem ko je storitev 3R modeli, ki ima komercialne alternative, namenjena predvsem raziskovalcem senzibilnim za avtorske pravice, sta storitvi dvoinpolrasežni modeli in visokoločljive slike edinstveni.

Slika 6: ARIADNE Visual Media Services, visokoločljiva slika; prikazan je izsek iz vizualizacije lidarskih podatkov območja prazgodovinske utrjene naselbine Gradišče nad Knežakom, Slovenija (Edisa Lozić; http://visual.ariadne-infrastructure.eu/img/lidar_a).

5 Literatura

Nicola Aloia, Christos Papatheodorou, Dimitris Gavrilis, Franca Debole in Carlo Meghini. 2014. Describing Research Data: A Case Study for Archaeology. 13th International Conference on Ontologies, Data Bases, and Applications of Semantics (ODBASE 2014), Amantea, Italy, October 2014. V: R. Meersman et al. (ur.), *On the Move to Meaningful Internet Systems: OTM 2014 Conferences, Lecture Notes in Computer Science (LNCS)* No. 8841: Springer-Verlag, str. 768-775.

Nicola Aloia, Carlo Meghini, Dimitris Gavrilis, Christos Papatheodorou, Luca Versienti, Franca Debole in Nicola Makri. 2015. *Specification of the ARIADNE Catalogue Data Model v. 2. 5. 5*. http://ariadne-support.dcu.gr/files/ACDM_Version_2.5.5.pdf.

Nicola Amico, Paola Ronzino, Achille Felicetti in Franco Niccolucci. 2013. Quality management of 3D cultural heritage replicas with CIDOC-CRM. V: Vladimir Alexiev, Vladimir Ivanov, Maurice Grinberg (ur.): *Practical Experiences with CIDOC CRM and its Extensions (CRMEX 2013) Workshop, 17th International Conference on Theory and Practice of*

Digital Libraries (TPDL 2013), Valetta, Malta, September 26, 2013, str. 61-69, CEUR-WS.org/Vol-1117.

Edeltraud Aspöck in Martin Fera. 2015. 3D-GIS für die taphonomische Auswertung eines wiedergeöffneten Körpergrabes. *AGIT – Journal für Angewandte Geoinformatik* 1, str. 2-8.

Edeltraud Aspöck in Gunthram Geser. 2014. What is an archaeological research infrastructure and why do we need it? Aims and challenges of ARIADNE. *CHNT 18, 2013 – Proceedings*, http://www.chnt.at/wp-content/uploads/Aspoeck_Geser_2014.pdf.

Patrick Le Boeuf, Martin Doerr, Christian Emil Ore in Stephen Stead (ur.). 2015. *Definition of the CIDOC Conceptual Reference Model, Version 6.2*. http://83.212.168.219/CIDOC-CRM/sites/default/files/cidoc_crm_version_6.2.pdf.

Martin Doerr. 2014. Tailoring the Conceptual Model to Archaeological Requirements V: *ARIADNE. The Way Forward to Digital Archaeology in Europe*, str. 65-74, Rim, Italija. ARIADNE, http://www.ariadne-infrastructure.eu/content/download/4569/26666/version/2/file/Ariadne+Booklet.pdf.

Martin Doerr in Keith Schaller. 2008. The Dream of a Global Knowledge Network - A new Approach. *ACM Journal on Computers and Cultural Heritage* 1(1).

Evropska skupnost. 2015. *DCAT Application Profile for data portals in Europe Version 1.1*. https://joinup.ec.europa.eu/node/137964/.

Achille Felicetti, Tiziana Scarselli, M.L Mancinelli in Franco Niccolucci. 2013. Mapping ICCD Archaeological Data to CIDOC-CRM: the RA Schema. V: Vladimir Alexiev, Vladimir Ivanov, Maurice Grinberg (ur.): *Practical Experiences with CIDOC CRM and its Extensions (CRMEX 2013) Workshop, 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013), Valetta, Malta, September 26, 2013*, str. 11-22, CEUR-WS.org/Vol-1117.

Achille Felicetti. 2014. Moving Ahead: the Integration Process. V: *ARIADNE. The Way Forward to Digital Archaeology in Europe*, str. 53-59, Rim, Italija. ARIADNE, http://www.ariadne-infrastructure.eu/content/download/4569/26666/version/2/file/Ariadne+Booklet.pdf.

Kate Fernie. 2014. *ARIADNE Report D3.3: Report on data sharing policies*. http://www.ariadne-infrastructure.eu/index.php/Resources/D3.3-Report-on-data-sharing-policies.

Dimitris Gavrilis in Christos Papatheodorou. 2014. Towards interoperability: the ARIADNE Registry. V: *ARIADNE. The Way Forward to Digital Archaeology in Europe*, str. 45-52, Rim, Italija. ARIADNE, http://www.ariadne-infrastructure.eu/content/download/4569/26666/version/2/file/Ariadne+Booklet.pdf.

Guntram Geser in Franco Niccolucci. 2012. Virtual museums, digital reference collections and e-science environments. *Uncommon Culture* 3(5/6), 12-37, http://uncommonculture.org/ojs/index.php/UC/article/view/4714/3677

Hella Hollander in Maarten Hoogerwerf. 2014. *ARIADNE Report D13.1: Service Design*. http://www.ariadne-infrastructure.eu/Resources/D13.1-Service-Design.

Fadi Maali in John Erickson. 2014. *Data Catalog Vocabulary (DCAT). W3C Recommendation 16 January 2014*. http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/

Anja Masur, Edeltraud Aspöck, Gerald Hiebel in Keith May. 2014. Comparing and mapping archaeological excavation data from different recording systems for integration using ontologies. V: *Proceedings of the 18th International Conference on Cultural Heritage and New Technologies*, Vienna, Austria, November 2013, http://www.chnt.at/wp-content/uploads/Masur_etal_2014.pdf.

Carlo Meghini. 2014. Providing services: search and beyond. V: *ARIADNE. The Way Forward to Digital Archaeology in Europe*, str. 59-64, Rim, Italija. ARIADNE, http://www.ariadne-infrastructure.eu/content/download/4569/26666/version/2/file/Ariadne+Booklet.pdf.

Franco Niccolucci in Julian D. Richards. 2013a. ARIADNE: Advanced Research Infrastructure For Archaeological Dataset Networking in Europe, International. *Journal of Humanities and Arts Computing* 7.1-2, pp 70–88, Edinburgh University Press, DOI: 10.3366/ijhac.2013.0082.

Franco Niccolucci in Julian D. Richards. 2013b. ARIADNE: Advanced Research Infrastructures for Archaeological Dataset Networking in Europe. A new project to foster and support archaeological data sharing. V: *The European Archaeologist* 39, Summer 2013, http://e-a-a.org/TEA/TEA39.pdf.

Kieron Niven in Holly Wright. 2014. *ARIADNE Report D4.4: Initial Report on Good Practices*. http://www.ariadne-infrastructure.eu/Resources/D4.4-Initial-Report-on-Good-Practices.

Christos Papatheodorou, Dimitris Gavrilis, Holly Wright, Paola Ronzino in Carlo Meghini. 2013. *ARIADNE Report D3.1. Initial report on standards and on the project registry*. http://www.ariadne-infrastructure.eu/Resources/D3.1-Initial-Report-on-the-project-registry.

Marco Potenziani, Marco Callieri, Massimiliano Corsini, Marco Di Benedetto, Federico Ponchio, Matteo Dellepiane, in Roberto Scopigno. 2014. An advanced Solution for Publishing 3D Content on the Web. V: *Internationa Conference on Museum and the Web Florence*, Firence (Italija), Feb 2014. http://mwf2014.museumsandtheweb.com/paper/an-advanced-solution-for-publishing-3d-contents-on-the-web/.

Marco Potenziani, Marco Callieri, Matteo Dellepiane, Massimiliano Corsini, Federico Ponchio in Roberto Scopigno. 2015. 3DHOP: 3D Heritage Online Presenter. *Computers & Graphics* 52, str. 129–141.

Julian D. Richards. 2012. Digital Infrastructures for Archaeological Research: A European Perspective. *CSA Newsletter* XXV (2), September 2012. http://csanet.org/newsletter/fall12/nlf1202.html

Paola Ronzino, Kate Fernie, Christos Papatheodorou, Holly Wright in Julian Richards. 2013a. *ARIADNE Report D3.2 Report on project standards*. http://www.ariadne-infrastructure.eu/Resources/D3.2-Report-on-project-standards.

Paola Ronzino, Nicola Amico, Achille Felicetti in Franco Niccolucci. 2013b. European standards for the documentation of historic buildings and their

relationship with CIDOC CRM. V: Vladimir Alexiev, Vladimir Ivanov, Maurice Grinberg (ur.), *Practical Experiences with CIDOC CRM and its Extensions (CRMEX 2013) Workshop, 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013), Valetta, Malta, September 26, 2013*, str. 70-79, CEUR-WS.org/Vol-1117.

Paola Ronzino, Franco Niccolucci, Achille Felicetti in Martin Doerr. 2016. CRMba a CRM extension for the documentation of standing buildings. *International Journal on Digital Libraries*, Focussed issue on Networked Knowledge Organization Systems 17(1), str. 71-78, DOI: http://dx.doi.org/10.1007/s00799-015-0160-4.

Roberto Scopino in Matteo Dellepiane. 2013. *ARIADNE Report of the First Meeting of the Ariadne SIG "3D & Visualization"*. http://www.ariadne-infrastructure.eu/content/download/1933/10854/version/1/file/ARIADNE_Report_WS+_MMdata_v2_c.pdf.

Hannes Selhofer in Guntram Geser. 2015. *ARIADNE Report D2.2. Second Report on Users' Needs, v2.1*. http://www.ariadne-infrastructure.eu/content/view/full/1188.

Spletni vir 1. ARIADNE Reference Model http://www.ariadne-infrastructure.eu/Resources/Ariadne-Reference-Model (dostop 10.6.2016).

Seta Štuhec. 2012. Dvoinpoldimenzionalno in tridimenzionalno upodabljanje artefaktov (2.5D and 3D Visualizations of Artefacts). *Arheo* 29, 87-98.

Benjamin Štular. 2012. iKnjiga - novi medij? = iBook - a new medium? V: Ines Vodopivec (ur.), *Ljubljana v BiTiH - BiTi v Ljubljani: prispevki iz prvega ljubljanskega kongresa digitalizacije kulturne dediščine = papers from the first Slovenian congress for digitisation of cultural heritage*. Zveza bibliotekarskih društev Slovenije, Narodna in univerzitetna knjižnica, Ljubljana, str. 223-231.

Benjamin Štular, Ana Ornik Turk in Andrej Pleterski. 2013. *Dotik dediščine. Trirazsežni prikaz zgodnjesrednjeveškega naglavnega nakita iz najdišča župna cerkev v Kranju*. Založba ZRC, Ljubljana. https://itunes.apple.com/si/book/dotik-dediscine/id789166886?mt=11&ign-mpt=uo%3D4.

Benjamin Štular in Seta Štuhec. 2015. *3D Archaeology. Early Medieval Earrings from Kranj*. Založba ZRC, Ljubljana. https://itunes.apple.com/si/book/3d-archaeology/id972355479?mt=11.

Andreas Vlachidis, Doug Tudhope, Milco Wansleeben, Katie Green, Lei Xia, Michael Charno in Holly Wright. 2015. *ARIADNE Report D16.2: First Report on Natural Language Processing*. http://www.ariadne-infrastructure.eu/index.php/Resources/D16.1-First-Report-on-Data-Mining.

Xander W. Wilcke. 2015. *ARIADNE Report D16.1: First Report on Data Mining*. http://www.ariadne-infrastructure.eu/index.php/Resources/D16.1-First-Report-on-Data-Mining.

Holly Wright. 2014. *ARIADNE Report D12.1 "User Requirements"*. http://www.ariadne-infrastructure.eu/Resources/D12.1-Use-Requirements.

Digital Video in Digital Humanities Methodology: A Case Study

Aleš Vaupotič,[*] Marco Buziol,[†] Narvika Bovcon[‡]

[*] Research Centre for Humanities, University of Nova Gorica
Vipavska cesta 13, SI-5000 Nova Gorica
ales.vaupotic@ung.si
[†] Academy of Fine Arts, Venice
Dorsoduro, 423, 30123 Venezia, Italy
accaquattro@libero.it
[‡] Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, SI-1000 Ljubljana
narvika.bovcon@fri.uni-lj.si

Abstract

The paper presents a digital reconstruction of the building that hosts the new headquarters of the Academy of Fine Arts in Venice as a digital humanities methodology case study. The building has been appropriated for different functions throughout history, from being a hospital for incurables to an orphanage. Recently the space was re-functioned for the studios of the Art Academy. In the project we have integrated multifaceted information about the different stages of the building's use in the medium of digital animation. Finally, the digital 3-D model was used for exhibiting contemporary art works by students of the Academy and videos by artists from the Society for Connecting Art and Science ArtNetLab.

1. Introduction

The programmatic collective monograph *Digital_Humanities* (2011) starts from the idea that it is not the digital humanities as a new field that is coming to life and evolving, but that the humanities in general have changed irreversibly: "The digitization of the world's knowledge and its movement across global networks [...] have transformed what we understand by and how we approach the humanities in the 21st century" (p. 26). The so-called second-generation digital humanities reflects on the unavoidable intertwinings of (new) communication media with existing humanities research practices. One of the main foci of the book are the "transmedia modes of argumentation" (p. 4), i.e. the non-linear and non-verbal modes of argument, which must, of course, be effective, clear, and rigorous in the use of evidence. The monograph argues for a new "form of scholarly practice; multimedia modes of argumentation that are object-based rather than discursive" (p. 33). Here "object-based" refers to the creation of new artefacts that convey meanings.

The consequence of the medial shift in scholarly practice is the necessity for combining the "perspectives of humanists, designers, and technologists" (p. 10). From the vocabulary of graphic and multimedia design many new humanities terms and formulations, sometimes refused as fashionable and jargon-laden, have been taken. However, it is clear that in the case of a multimedia argument the language of video is the encompassing semiotic framework: it extends the possibilities of film language into the pre-montage cinema and into the motion graphics paradigm which, as a rule, combines the moving pictures with graphic elements, usually animated—in all of these cases the film space emerging from the montage of shots is missing, the surface in motion, in combination with sound and other communication channels, is the medium.

2. The goal of the project

The task of this paper is to present a concrete example of argumentation not tied to the verbal language. We maintain that it is necessary to realise pilot projects that embody concrete solutions—thereby facilitating critical reflection—and that, at least in the current moment, it would be extremely difficult and risky to develop the grammar of transmedia modes of argumentation in isolation from practical prototyping. The design process and its result offer deep insight into the practices of humanities that are useful in a digital environment. The paper presents in detail the visual argument for presenting a historical diachrony in a geolocation in the medium of digital animation.

The aim of our experimental project was a digital reconstruction of the new (since 2004) location of the Venice Fine Arts Academy at Dorsoduro and a presentation of its diverse functions through centuries in a digital animation. Originally, the building was the Hospital for Incurables, from about 1550 till 1807, and the courtyard, now a vast open space, was the site of a pilgrimage church. In the next period, from 1807 till 1819, the building was a regular hospital of Venice, and the ground floor was used for housing patients, the first floor for orphans. The church was famous for concerts of classical music that attracted numerous visitors and were the source of financing for the operation of the hospital. In the years 1819-1934 the building was transformed into military barracks, and in this context the church was demolished in 1831. Afterwards the building was transformed and used as a Centre for the Rehabilitation of Minors (1938-1977).

3. Gathering of historical information, and reference photos and videos

The first step of the project was gathering the historical information and visual reference materials about the building that we planned to recreate digitally. A visit to the CIRCE Cartography Lab of the University IUAV of

Venice was essential, since they provided us with architectural plans of the building as it exists today after the renovation works. We used the blue-prints in 3-D modeling. The visit to the library of the Academy of Fine Art of Venice disclosed further interesting historical facts and additional floor plans. The book *La Nuova Accademia di Belle Arti di Venezia* edited by Renata Codello (2011) documents the period when the building was transformed and used as a Center for the Rehabilitation of Minors. The inscription on a building wall documented on a photograph pointed us in the direction of looking for more inscriptions about the historical facts and dates, such as the inscription board marking the old pharmacy or a date carved in stone in roman numbers on one of the four wells in the courtyard. We recreated these inscriptions in the digital model and used them as corner-stones of our visual narrative on the building's history.

The next step was the visit to the Academy premises in order to collect contemporary reference photos and videos. For modelling of a digital object it is necessary to collect many high quality reference photographs and videos to understand its shape, the proportions of the object and its position in space. The photographs are used also for achieving the correct visual result when we work on textures, materials, lighting and rendering algorithms, and as image sources that are enhanced in Photoshop software and applied to the model as textures.

The movement of the camera was animated with great care, since its function was also to evoke the atmosphere of the ambient, the memory of e.g. riding a vaporetto or walking under colonnades, to build suspense and to communicate emotions. The visit to the actual site gave us the experience of the place, which involves the spatial relations, the light and the pace of the inhabitants' activities. The physical experience of spatial orientation, proportions and volumes, as well as light and rhythm, is recursively needed as a reference to evaluate the digital animation—as a criterion about whether the animation gives you the right “feel” of the space.

4. The structure of the visual narrative

The selection of most interesting architectural and urbanistic features guides the visual narrative, the montage, and the movements of the virtual camera. Four main parts of the building were identified—the waterfront facade, the courtyard, the church and the library (former pharmacy). Each of the architectural parts was then presented in a separate take (sometimes two takes) of the virtual camera in the digital animation. However, a fluid experience of the connected parts of the building complex had to be achieved, therefore some parts of the building reappeared throughout the animation. The angles of the camera view were carefully considered for each take.

Figure 1: Digital reconstruction of the new location of the Academy of Fine Arts in Venice at Dorsoduro. The waterfront facade from the beginning of the digital video.

The video begins with a wide shot of the waterfront that shows the facade of the Academy and the neighboring buildings, lined-up along the Canale della Giudecca, so that the viewer understands immediately that we are in Venice. The movement of the camera is animated carefully and synchronized with the waves of the canal waters—as though the viewer was approaching the Academy on a vaporetto, the water bus. The movement is relatively slow, it allows the viewer to explore the set-up (Figure 1), enjoy the details of the facade, and remember the rhythm of riding on a vaporetto; at the same time the animated take builds suspense. The camera enters the Academy through the main portal and halts as it enters the vast open space of the courtyard, surrounded by beautiful colonnades. Here is the first montage cut, in the next take the camera is animated from the point of view of a visitor who walks under the colonnades and admires the play of light and shadow cast by the columns around the sun-bathed courtyard (Figure 2). This take ends as the camera steps into the central area—after walking approximately a 90 degrees angle around the courtyard—and looks towards the entrance, through which the waters of the canal are again visible in the distance.

Figure 2: The courtyard.

The next take, in contrast to the previous ones, builds on the main characteristic of the virtual camera, i.e. it being free of any physical constraints: it jumps over the building and rotates to show the rectangular courtyard from a bird's-eye perspective view, in a very bright light under the sky. This in reality impossible camera movement is used in its first part to show fine detail such as lace-like railings, the shutters and stones on the facade (Figure 3), while towards the end of the take the now non-existent church appears in the courtyard. The camera shows it from above (Figure 4), thus revealing the bell tower and the partitioning of the remaining space of the courtyard by the corridors that connected the church to the main building at each of the four sides. There are four wells, one in each corner, they were used in case of fire, one for each partition of the courtyard.

Figure 3: Details on the facade.

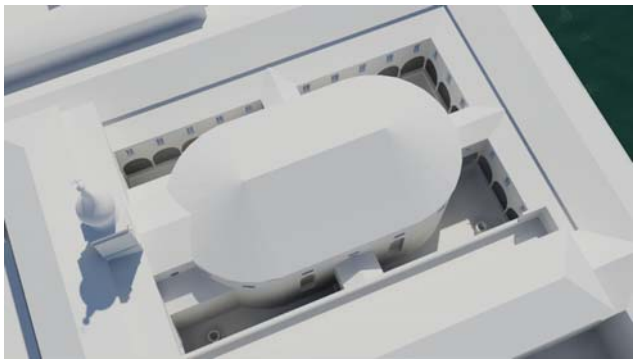


Figure 4: The church of the Hospital of Incurables (1550–1807).

The next take is darker, the camera moves again as if following a virtual person walking under the colonnades, however, this time the central space is filled with the walls of the church and the feeling is entirely different, claustrophobic, dark and even sad (Figure 5). The viewer senses the heavy atmosphere of the period when the building was a hospital for incurable patients. Two inscriptions appear in this part, the date in the well reporting when it was built and the dates over the entrance to the church denoting the period of its existence.

Figure 5: The colonnades and the church in the courtyard.

Following the rule of repetition, which pin-points the beginning and the end point of an era, the next take starts again from above, showing the church dissolving by crumbling to pieces (Figure 6).

Figure 6: The church is demolished—the removal of the church is represented in a stylized fashion.

As soon as the church disappears we notice the inscription on the facade in the courtyard that says “Centro di operazione maternita infanzia” (Center for the Rehabilitation of Minors). The camera flies past the inscription denoting the orphanage into an elaborate room which was, as the inscription on a stone plate reveals, a pharmacy, but now it is used for the library (Figure 7). In this room the viewer experiences the interior of the building for the first time.

Figure 7: The former pharmacy, now the library.

The room has many details, stucco and marble on the walls. The digital animation as it was realized concludes here by showing a very different space, by its luminescent floor maybe reminding the viewer of the concluding part of Stanley Kubrick's famous film *2001: Space Odyssey* (1968).

While still working on the project we also planned to make a 3-D scan of the frieze from the ex-pharmacy/library, however in the end it was not possible to include this activity in the project. The long frieze of stucco under the ceiling was constructed from the high resolution photograph simply by making a displacement map for it. It is shown in a close up, in a traveling of the camera in a new landscape of decoration, contrasting the previous plain white walls (Figure 8). Thereby a completely new domain of representations are added, which differs from the phenomenal world and opens up towards the universe of imagination consisting of floral and anthropomorphic patterns.

Figure 8: The close-up of frieze of stucco decoration in the library.

5. Use of software

The architectural blue-prints were referenced and reconstructed digitally first in AutoCAD 2D, creating the shapes of the floors, roofs and the vertical shape of the columns. Then they were imported into 3D Studio Max to give them the third dimension and every single part was moved in the right position. At last the details were added to the main volumes: the windows, the railings, the shutters, the bases and capitals of the columns, the flooring, the doors, the wells. Images were used as bump and displacement maps to create planar relief forms when possible: the stones on the facade, lace-like railings, some of the stucco decorations.

The virtual camera was animated using key-frames along a path. The accelerations, non-uniform movements, rhythms of walking and of vaporetto—as indicated by the virtual camera shot at the beginning of the video—were defined carefully. Some parts were animated with the help of a live-footage reference video.

The sea was created from a plane by applying to it a water material, to which a modifier noise was added for animation. We spent a lot of time to understand and recreate the right movement and appearance of the waves.

The crumbling of the church to pieces was done with a "warp bomb" algorithm, where it was possible to control the parameters of the explosion's strength: speed, gravity, the "chaos" parameter, rotation of the pieces while they fall down.

The video contains just two materials (shaders). The main focus was the material for the architecture, which is white to enhance the conceptual reading of the forms. However, this material has a soft feeling of an uneven wall, on which the light interferes with the surface in a more tactile and elaborate manner. The illumination of the exterior was done using a virtual daylight, that simulates the light of the sun, the color of the sky and the shadows in a particular location at a particular point in time, in this case: Venice in May at 10 a.m.

It was important also to fine-tune the indirect illumination of the sections, where the light can't reach directly, such as in the corridors around the courtyard behind the columns. Those parts, of course, aren't completely dark, which is achieved by setting the parameters that control the rays and the bouncing of the light.

The second material was the one used for the water of the laguna. The same material was applied also to the windows that overlook the laguna. The reduction of the number of different materials added to the clearness of the presentation.

The interior of the library was illuminated with additional lights that simulate artificial light.

The scenes were rendered with Mental ray rendering engine. The rendered TIFF images were composited in Adobe After Effects. All the cuts were sharp. We used the transition between a series of key-frames to blend-in the church as it first appears in the courtyard.

6. Exhibitions

The digital reconstruction of a building, and of its historical transformations, does not provide us with a simple result. Apart from the video-animation the digital object has to be reintegrated into the existing communication practices, which is by definition not limited to the medium of scientific journal article.

This project was subsequently integrated in a video presentation of students' works from the Academy of Fine Arts of Venice—the footage of contemporary life in the building was combined with parts of the digital animation. It was shown at the meeting on the Cultural Property of the Academies of Fine Arts of Italy, organised by the Academy of Fine Arts of Naples and the Ministry of Education, University and Research of Italy in Naples in June 2013.

Figures 9 and 10: Exhibition of the videos of artists from ArtNetLab, Society for Connecting Art and Science

On another occasion, the digital model was used also as a virtual exhibition space: videos of artists from the ArtNetLab Society for Connecting Art and Science from Ljubljana were placed on virtual panels under the colonnades in the courtyard in the 3-D model of the

Venice Academy (Figure 9 and 10). For this reason, several additional digital animation clips were generated that showed each video in the virtual setting with an animated camera, which evoked the undercurrent of a history-laden space at the same time as it conveyed the most recent video works.

An important distinction emerges, which is crucial for reading the video image. The temporality of these artistic videos becomes related to the temporality of the video-animation, and to the historical diachrony of the reuse of the building in its location. The video installation of this virtual-reality exhibition was presented at the gallery Dimenzija napredka (Dimension of Progress) in Solkan, a town on the border between Slovenia and Italy, in October 2013. The virtual reconstruction of a building, and the videos in it, were presented in a real space—one architecturally very dynamic, consisting of glass surfaces that multiplied and reflected different objects, images, and screens—, which presented simultaneously other art objects, sometimes related to the videos in the digital animation.

7. Inspiration in similar projects

The task to digitally reconstruct the location of the Academy in Venice was a straightforward assignment in the phase of modeling the virtual building according to the architectural floor-plans. Afterwards the main question arose as how to present the virtual 3-D model. The digital video is only one possible solution, we could have created also an interactive virtual-reality space which is displayed interactively in a real-time video. A project of this kind was done in the Laboratory for Computer Vision at the Faculty of Computer and Information Science already in 1998, when Srečo Dragan, an established Slovenian new media artist, initiated (in collaboration with ZDSLU, the Slovenian Association of Fine Artists Societies) the project of *Jakopič Virtual Gallery*, which was a virtual reconstruction of Jakopič's Pavilion, a venue for Impressionist art exhibitions since 1908, however, it was demolished in 1962. The virtual gallery reconstructed in VRML on-line 3-D engine hosted exhibitions of several Slovenian artists that could be viewed by walking through the model of the gallery (Solina, 2000).

The next decision needed was about the layering of information with the use of image-textures and compositing, i.e. as a new flat layer over the rendered virtual reality. An interesting example of such an approach is the reconstruction of the Parthenon in Athens done by Paul Debevec (2004), where the virtual model is shown in a fluid traveling of the camera around the model and between different modes of presentation: from the rendering mode in the wireframe to the polished surface of the same model after the textures, materials and illumination were added to make the image indiscernible from the camera footage of the real building. However, nowadays the hyper-realistic renderings of whole sets of ancient cities can be observed in numerous hollywood movies that use special effects at the highest level of verisimilitude. We needed a more conceptual approach, due also to the working hours and production constraints. Nevertheless, with a correct and very focused approach on the image of the final presentation, these were no real limitations.

A canonical reference for 3-D virtual spaces is a series of projects *The Legible City* by Jeffrey Shaw. The environment is similar to a—sometimes particular—city, but the buildings are replaced by large 3-D letters. Therefore the direct link to the verisimilitude is broken, and the space becomes a spatial constellation of semiotic entities. Such an approach was realized in projects *VideoSpace* by Narvika Bovcon and Aleš Vaupotič (2003), and *Data Dune* by the same authors in collaboration with Barak Reiser (2005).

It is important to note that all these project are related to computer games, however at the same time they explicitly do not use the gaming element. E.g. here is no problem solving and rewarding system. The 3-D computer generated space—either a real-time one or a photo-realistic high-quality rendered one—provides "non-verbal" modes of argument, which can efficiently mediate and interpret the historical records.

8. Conclusion

A reconstruction of a historical building that is linked to different social functions in its environment can never be reduced to its material 3-D shape, such as can be captured potentially by 3-D scanners. The key task in fact is to reduce the overwhelming masses of information available in the existing state of the architecture and in cultural records linked to the past appearances and uses of the site. Nevertheless, a mere reduction of the data resolution, e.g. the level of detail, is not the solution. What is needed is to convey the human experience of the place reconstructed and represented, either based on personal experience or one based on studying historical documents.

In the case of the building of the Ex Ospedale degli Incurabili which now houses the Accedemy of Fine Arts Venice the most striking insight that we have learned during the project is, apart from its different uses in history, the contrast between the existence of a large church in its centre and the, later, empty courtyard. This is an opportunity for the language of 3-D computer animation, which can simulate the atmosphere emanated from the enclosed spaces and looming volumes, and also the brilliance of an open space. The suggestive transformation—in fact a dramatic one, considering the past residents—virtually by itself invites new contents, liberated by artistic creativity. In the project we have embraced these possibilities and enriched the past with the artworks stemming from the context of presentation, in analogous way that the building itself has become the new home of Venetian art students.

The video-animation realized in the project is accessible on-line at the URL: http://black.fri.uni-lj.si/atlas_benetke_buziolvaupoticbovcon.

9. Acknowledgment

Our project was developed in the frameworks of the lifelong learning programme Leonardo Da Vinci (a European project) in collaboration between the Academy of Fine Arts of Venice and the Faculty of Computer and Information Science of the University of Ljubljana. The internship of the author of the digital animation Marco

Buziol, a former student of the Venice Academy, took place at the University of Ljubljana from October 2012 till March 2013 under the supervision of Assoc. Prof. Narvika Bovcon. Special thanks go to Prof. Laura Safred and Prof. Gloria Vallese from Venice Academy, and to Prof. Franc Solina from the University of Ljubljana. Thanks to the artists from ArtNetLab Society for Connecting Art and Science, who alongside the authors of this text contributed their videos for the virtual exhibition in the project: Jure Fingušt Prebil, Eva Lucija Kozak, Gorazd Krnc, Dominik Manič, Vanja Mervič, Tilen Žbona.

10. References

Narvika Bovcon and Aleš Vaupotič. 2003. *VideoSpace*. Ljubljana, ArtNetLab.

Narvika Bovcon and Barak Reiser and Aleš Vaupotič. 2005. *If you look back, it wont' be there anymore*. Ljubljana, ArtNetLab.

Anne Burdick and Johanna Drucker and Peter Lunenfeld and Todd Presner and Jeffrey Schnapp. 2012. *Digital_Humanities.* Cambridge, Mass.: MIT Press.

Renata Codello. 2001. *La Nuova Accademia di Belle Arti di Venezia*. Marsilio, Venice.

Paul E. Debevec. 2004. "Making The Parthenon," SIGGRAPH 2004: Electronic Theatre. http://gl.ict.usc.edu/Films/Parthenon

Jakopičeva virtualna galerija. 1998. http://black.fri.uni-lj.si/jakopic

Jeffrey Shaw. 1998. *The Distributed Legible City*. http://www.jeffrey-shaw.net/html_main/show_work.php?record_id=102

Jeffrey Shaw, Dirk Groeneveld. 1989. *The Legible City*. http://www.jeffrey-shaw.net/html_main/show_work.php?record_id=83

Franc Solina. 2000. Virtual technology and remote observation over the Internet for art applications. In: *Konferenzband EVA 2000 Berlin: Elektronische Bildverarbeitung & Kunst, Kultur, Historie*, pages 171-177. Berlin. Gesselschaft zur Förderung Angewandter Informatik e.V.

Alvise Zorzi. 1972. *Venezia Scomparsa*. Electa, University of Michigan.

Digitalna humanistika v šoli

Maja Vičič Krabonja

Srednja ekonomska šola Maribor
Trg Borisa Kidriča 3, 2000 Maribor
maja.vicic1@guest.arnes.si

Povzetek

Prispevek prinaša pregled sprememb, ki jih v pouk (na konkretnem primeru zgodovine) prinaša uporaba sodobne tehnologije, ter jih podpira z nekaterimi empiričnimi podatki. Analizira uporabo različnih virov, nove oblike vizualizacije in razlage ter sodelovanja in komunikacije na daljavo. Opisane so nekatere nove možnosti zbiranja podatkov o poučevanju in učenju, ki so temelj za hitro podajanje in sprejemanje povratnih informacij. ter prilagajanje učnih aktivnosti potrebam učencev in doseganju načrtovanih ciljev. V povezavah z novimi pedagoškimi paradigmami, ki poudarjajo na učenca osredinjen pouk, navaja mnoga orodja, ki jih v procesu poučevanja in učenja uporabljata tako učitelj kot učenec. Ugotavlja, da uporaba informacijsko komunikacijske tehnologije omogoča večjo aktivnost in ustvarjalnost učencev ter s tem spreminja tudi vlogo učitelja.

Digital Humanities in School

The article includes a review of changes, which school lessons (a specific example of the school subject History) include in using modern technology and support them with some empiric data. It analyses using of different sources, new forms of visualization, explanation, co-operation, and long-distance communication. There are some new possibilities of gathering data about teaching and learning described, which are foundations for fast giving and getting of feedback information and adaptation of teaching/learning activities to students and reaching the planned goals. In connection to new pedagogical paradigms, which focus on student oriented lessons, it lists many tools, which are used in the process of teaching and learning by both the teacher and the student. It establishes that using of information technology enables more activity and creativity of students. Therefore, it changes the role of the teacher as well.

1 Uvod

Če velja, da je digitalna humanistika presek digitalnih tehnologij in humanistike, potem s šolskim okoljem dodajamo še tretje področje, kjer se v središču prekrivanja vse treh področij znajdeta učitelj in učenec, ki izbirata med mnogimi digitalnimi (in drugimi) tehnologijami, s katerimi naj postane vsebina učencu[1] bolj dostopna in ki podpirajo izbran pedagoški pristop. Z vpeljevanjem (digitalnih) tehnologij v šolski prostor se je klasični pedagoški trikotnik (vsebina-učitelj-učenec) v zadnjih dveh desetletjih razširil in preoblikoval, tako da učitelj vse bolj postaja mentor, ki ustvarja učne priložnosti, pri vsebini se je povečala predvsem njena dostopnost in pestrost, učenec pa postaja aktiven dejavnik in ne več le pasivni prejemnik, ki reproducira v učnem načrtu zahtevane vsebine. Učiteljeva vloga zaradi uporabe digitalnih tehnologij pri pouku ni manjša ali manj pomembna, zaradi njihove uporabe in dostopnosti v vsakdanjem življenju postaja zahtevnejša in odgovornejša; učence mora, tudi s svojim zgledom, torej uporabo pri pouku, pripraviti na njihovo smotrno in odgovorno rabo.

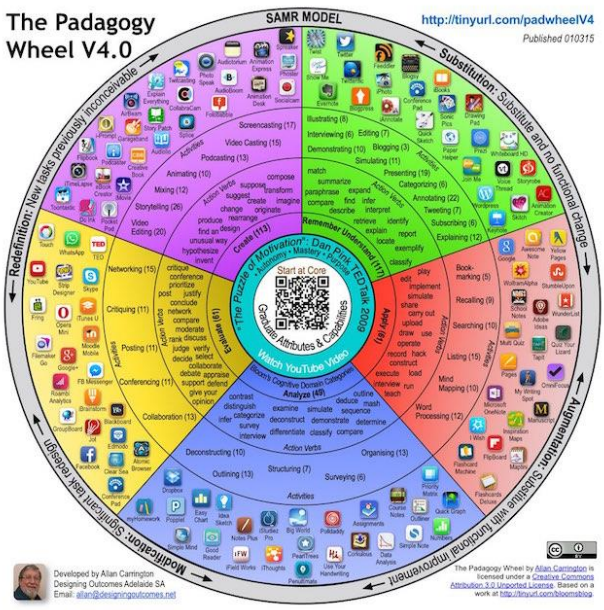
Učitelji moramo ponovno premisliti svoje znanje in kompetence ter reflektirati svoje izkušnje na vseh treh področjih TPACK (Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge, 2006) modela, torej tehnologije, pedagogike in vsebine, katerih presek je osnova za kakovostno poučevanje s tehnologijo, ki služi doseganju ciljev, razvijanju kompetenc in v ospredje postavlja na učenca usmerjen pouk.

2 Namen članka

V prispevku želimo na primeru zgodovine (humanistični šolski predmet) pokazati, kako uporaba digitalne tehnologije spreminja način usvajanja znanja (učenja v šoli in izven nje) in poučevanja. Prikazani bodo različni načini uporabe digitalne tehnologije glede na stopnjo aktivnosti učitelja in učenca: pasivna raba (učenci opazujejo vizualizirane pojme, e-kompetence ne razvijajo), kvazi aktivna raba (ob pripravljenih vsebinah učenci razvijajo tudi veščine dela s tehnologijo) in aktivna raba (podani so odprti problemi, učenci sami izbirajo orodja za reševanje problemov).

Poleg tehničnih naprav, kjer tablice in telefoni vedno bolj nadomeščajo računalnike, je seveda pomembna tudi izbira orodij. Vsa orodja, ki bodo omenjena ali opisana, so prosto dostopna (vsaj v osnovni verziji). Pri izbiri orodij se učitelji znajdejo vsak po svoje, vedno pa morajo imeti v mislih, da izberemo tisto, ki omogoča doseganje zastavljenih ciljev in pričakovanih dosežkov. Na spletu so je moč najti mnoge sezname in ideje (npr. http://c4lpt.co.uk/top100tools/), mnogi med njimi so usklajeni z revidirano Bloomovo taksonomijo (gl. Slika 1), kar učiteljem dodatno olajša delo.

[1] Izraz učenec uporabljamo v smislu učečega se, kdor se uči, ne glede na raven izobraževanja, kot je razložen v SSKJ (http://bos.zrc-sazu.si/cgi/a03.exe?name=sskj_testa&expression=učenec&hs=1, dostop 1. 5. 2016).



Slika 1: Predlog izbora digitalnih orodij glede na Bloomovo taksonomijo (Carrington, 2015).

Med slovenskimi portali bi izpostavila SIO portal, kjer se v zavihku Podpora (http://podpora.sio.si/) zbirajo opisi in povezave do spletnih orodij, hkrati pa je omogočeno iskanje in filtriranje po kategorijah in oznakah.

2.1 Digitalne tehnologije in pouk zgodovine

V učnem načrtu (*Zgodovina: gimnazija: strokovna gimnazija*, 2008: 8). je zapisano, da pouk zgodovine v šoli izhaja iz zgodovine kot znanstvene discipline, katere glavni koncepti so: koncept časa in prostora, koncept sprememb, koncept kontinuitete, koncept vzročnosti in posledičnosti. Poudarek pri pouku je na kritični analizi in interpretaciji podatkov ... in na oblikovanju samostojnih zaključkov, mnenj in stališč o pojavih in procesih ... Iz tega izhaja, da cilji pouka zgodovine niso več le znanje in razumevanje, pač pa tudi cilji, ki se nanašajo na razvijanje spretnosti in veščin ter razvijanje odnosov, ravnanja, naravnanosti in stališč.

Če smo učitelji doseganje ciljev znanja in razumevanja še zmogli z zgodbo, pa pri razvijanju spretnosti stopijo v ospredje aktivnosti učencev, pri razvijanju odnosov in stališč pa so učenci z razvojem digitalne tehnologije in njeno dostopnostjo vedno bolj pod vplivom različnih elektronskih medijev in družabnih (družbenih) omrežij.

Z izkušnjami iz prakse sama ocenjujem (*Kombinacija različnih IKT-orodij v srednji šoli na primeru predmeta zgodovina*, 2010: 368), da digitalne tehnologije omogočajo največje spremembe v načinu poučevanja in učenja ravno na področju razvijanja spretnosti in veščin, povezanih s predmetom, ter ključnih kompetenc za vseživljenjsko učenje (Evropska komisija, 2007), med katerimi je tudi digitalna pismenost. Z uporabo digitalnih tehnologij se približamo današnjim generacijam digitalnih domorodcev[2], ki pričakujejo drugačen model poučevanja, ki naj bo osredotočeno na učenca, izkustveno, raziskovalno, ustvarjalno, sodelovalno, vizualno bogato in vključujoče (Flogie et al., 2014: 72–73). Zbliževanje koncepta učenja, kot avtonomne vseživljenjske aktivnosti in sodobne tehnologije, je povzeto v Tabeli 1.

Mayer (2013: 165) navaja 10 kategorij učnih okolij, podprtih s tehnologijo, med katerimi jih je večina značilna tudi za humanistične predmete.

Sodobno učenje	**Sodobne tehnologije**
personalizirano	osebne
na učenca usmerjeno	usmerjene k uporabniku
vezano na lokacijo	mobilne
sodelovalno	omrežene
vseprisotno	vseprisotne
vseživljenjsko	trajne

Tabela 1: Zbliževanje učenja in tehnologije (Sharples et al., 2006: 3).

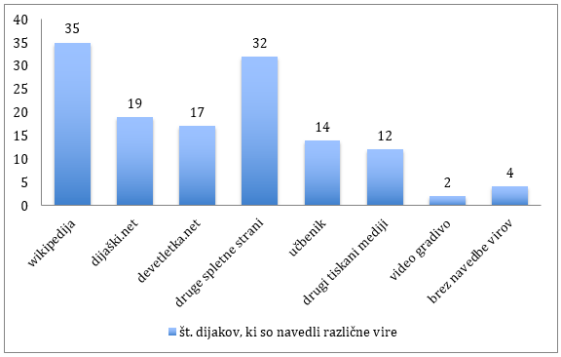
2.2 Novi viri podatkov in informacij

Vilma Brodnik (2015: 16) v *Smernicah za uporabo IKT pri predmetu zgodovina* navaja 17 spletnih mest, kjer lahko učitelji in učenci najdemo digitalizirana gradiva za pouk zgodovine (npr. dLib, KAMRA, DEDI in mnoge arhive), torej tudi vire prve roke, ki so bili do sedaj dostopni (zbrani in že izbrani) le v zgodovinskih čitankah (tudi novejših učbenikih) in kot taki ponujeni učencu. Slednji lahko sedaj vire samostojno zbira in izbira, jih presoja glede na relevantnost in primerja med seboj, sklepa, izvaja zaključke. Prav presojanje relevantnosti in iskanje uporabnih ter verodostojnih informacij je v 21. stoletju pomembna veščina, saj večina učencev dobiva informacije ne le z bolj ali manj specializiranih televizijskih kanalov, pač pa predvsem iz različnih družabnih omrežij (npr. YouTube).

Kljub temu velja, da je samostojno zbiranje informacij značilno predvsem za motivirane učence; žal se jih večina zadovolji s prvimi zadetki, to pa so običajno Wikipedija, Dijaški.net (ki se deklarira kot najbolj obiskana izobraževalna stran v Sloveniji) ter Devetletka.net (kjer učenci prepisujejo informacije iz predstavitev, ki so namenjene za učiteljevo uporabo v razredu). Graf 1 prikazuje pogostost uporabe različnih virov, ki jih je 59 dijakov 3. letnika programa ekonomske gimnazije[3] uporabilo pri pisanju dnevnika svojega sovrstnika iz 19. stoletja:

[2] Izraz digital native je prvi uporabil Prensky leta 2001, ki v svojem delu dokazuje, da današnji učenci razmišljajo in procesirajo drugače kot starejše generacije, da je način njihovega odraščanja z uporabo digitalne tehnologije vplival na fizično strukturo njihovih možganov in da so se spremenili vzorci njihovega razmišljanja (Prensky, 2001).

[3] Na Srednji ekonomski šoli Maribor, šolsko leto 2015/16.



Graf 1: Viri, ki jih dijaki najpogosteje uporabljajo.

Le dva učenca sta uporabila samo tiskane vire. V te podatke niso zajeti viri slik, kjer se kaže precej bolj pestra podoba (prevladujejo slike s tujih spletnih strani, medtem ko je med slovenskimi največ gradiva, ki ga je objavil Etnografski muzej), hkrati pa učenci niso uporabili niti ene slike iz tiskanih medijev. V razgovoru z učenci sem ugotovila, da je pestrost virov pri iskanju slik večja zato, ker se niso omejevali le na slovenske strani, torej ni bilo jezikovne ovire.

Iz prikazanih podatkov je razvidno, da je treba Wikipediji pri pouku posvetiti posebno pozornost z vidika verodostojnosti in uporabnosti informacij, ne glede na to, da so preverjanje, primerjanje virov ter multiperspektivnost eno izmed temeljnih načel pouka zgodovine in ne le posebnost uporabe Wikipedije. Res pa je ravno področje zgodovine (Moram, 2011) eno izmed tistih, ki so najbolj izpostavljena pristranskosti in različnim interpretacijam, zato je treba pri učencih dosledno razvijati zavedanje,[4] da je oblika javne enciklopedije, da lahko vanjo prispeva kdorkoli in da so tudi uredniki lahko bolj ali manj naklonjeni izbrani interpretaciji.

Ob dejstvu, da je svetovni splet glavni vir informacij za naše učence, jih je treba pri uporabi brskalnikov opozoriti na nekritično sprejemanje tako imenovanega internetnega mehurčka, saj jih večina[5] pregleda le zadetke na prvi strani, pa še to običajno samo prve tri. Dražič (2015) predlaga, da v razredu izvedemo preizkus z vpisom enakega niza besed v brskalnik, nato pa učenci dobljene rezultate primerjajo med seboj in z rezultati, ki jih dobijo z uporabo anonimnega brskalnika StartPage.

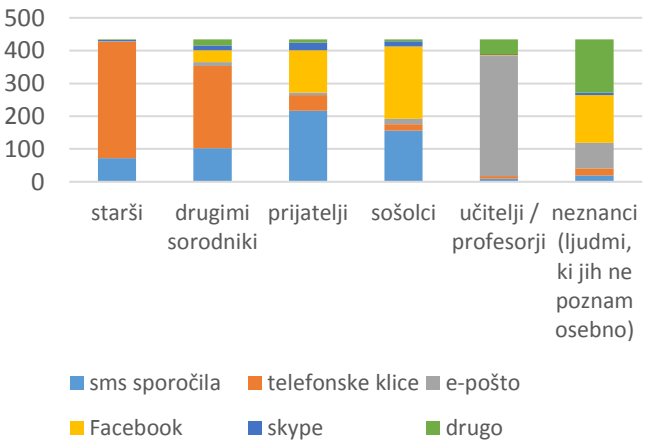
2.3 Nove oblike vizualizacije, predstavljanja in razlage

Nove oblike vizualizacije, predstavljanja in razlage niso nove le po imenu in uporabi digitalnih pristopov (npr. PPT ali Prezi predstavitve namesto grafoskopa ali episkopa), pač pa omogočajo uporabo interaktivnih slik (npr. Thinglink), zemljevidov (npr. Animaps), videoposnetkov (npr. Zaption, TEDed), časovnih trakov (npr. Timetoast), ki povezujejo besedilo, zemljevide, slike in zbirke podatkov ter tako omogočajo drugačno vizualizacijo časa in prostora, ki je ustreznejša za nelinearno, mrežno razmišljujoče učence, navajene na branje spletnih strani z mnogimi hiperpovezavami.[6] Vse navedene predstavitve lahko pripravi učitelj ali pa učenci sami, kar je še posebej želeno. S pojmovnimi mrežami (npr. xMind), ki jih pripravijo učenci, spodbujamo učenje z razumevanjem, izgled pojmovne mreže, ki je neposredna odslikava učenčevih kognitivnih struktur, pa nam omogoča razbiranje pravilnih in napačnih/pomanjkljivih povezav: ob povezovanju že znanega z novimi informacijami je pojmovno mrežo v e-obliki enostavno nenehno dopolnjevati ter spreminjati povezave, odnose med pojmi v skladu z novimi spoznanji (*Pojmovne mreže– ključ do učenja z razumevanjem*, 2015: 67).

Multimedijo, interaktivne simulacije, hipertekst in hipermedijo vsebujejo tudi i-učbeniki, ki so v slovenščini na voljo za naravoslovje, matematiko, pa tudi jezike in umetnost (https://eucbeniki.sio.si/). Interaktivnega učbenika za zgodovino še ni, so pa v z učnim načrtom za geografijo in zgodovino v osnovni in srednji šoli usklajena bogata e-gradiva *Kartografija v učni snovi* (http://egradiva.gis.si/web/guest).

2.4 Nove možnosti sodelovanja in timskega dela

Nove možnosti sodelovanja in timskega dela so učencem blizu, vsaj kar se tiče komunikacije na daljavo. Anketa[7] je pokazala, da sredstvo za komunikacijo na daljavo mladi izbirajo glede na to, s kom komunicirajo. Graf 2 prikazuje, kako pri komunikaciji s sovrstniki največ uporabljajo sms sporočila in Facebook, e-pošto pa le za učitelje oziroma profesorje. Učitelji so tako postavljeni pred izziv, kako pritegniti učence v svoj komunikacijski kanal (npr. forum v spletni učilnici), oz. dilemo, ali se jim pridružiti na družabnih omrežjih.



Graf 2: Kako se mladi sporazumevajo na daljavo z različnimi skupinami?

[4] Verjetno bi bilo smiselno tej temi ciljano nameniti več časa v okviru Knjižnično informacijskih znanj, seveda v povezavi s konkretnimi šolskimi predmeti in projekti.

[5] 98 % vprašanih učencev 3. letnika programa ekonomske gimnazije na SEŠ Maribor, oktober 2015.

[6] Nekatere študije ugotavljajo, da hiperpovezave upočasnijo potek branja in zmanjšajo koncentracijo bralca, saj se mora le-ta odločiti, ali bo povezavo odprl ali ne (Tanner 2014, 5).

[7] Anketo smo izvedli na vzorcu 451 mladih (13 % osnovnošolcev, starih nad 13 let, 75 % srednješolcev in 12 % študentov) jeseni 2014 za potrebe konference EDID.

Manj vešči[8] so sodelovanja na daljavo, npr. skupnega urejanja dokumentov v oblaku (Googlovi dokumenti, Padlet, Prezi ...). Pri zgodovini tako skupaj rešujejo problemsko zastavljene naloge, digitalna tehnologija pa učitelju omogoča, da spremlja in ovrednoti prispevek vsakega posameznika v skupini.

Eno izmed okolij, v katerem lahko poteka sodelovalno delo, je tudi Wiki. Z ustvarjanjem svojih wikijev (npr. v spletni učilnici Moodle) učenci dobijo izkušnjo ustvarjanja in deljenja znanja ter prevzemanja odgovornosti zanj. Če jih v tem primeru navajamo na preverjanje zapisanega in na sklicevanje na vire, bodo tudi Wikipedijo uporabljali z večjo mero kritičnosti.

Slika 2: Primer wikija v spletni učilnici.

2.5 Spremenjeno razmerje med šolskim in domačim delom

Spremenjeno razmerje med šolskim in domačim delom kot posledica uporabe digitalnih tehnologij najbolje odseva v pedagoškem modelu obrnjenega učenja (Flipped Learning), pri katerem učitelj razlago posname, učenci jo pogledajo doma (kadarkoli, kjerkoli, na katerikoli napravi ... v čemer zasledimo elemente personalizacije[9]), naloge pa rešujejo v šoli, pogosto v parih ali manjših skupinah, ob prisotnosti (pomoči) učitelja. Pri pouku zgodovine je taka oblika dela primerna za krajšo video razlago konceptov in reševanje avtentičnih nalog, vezanih na (multiperspektivne) vire. Zaradi jezikovnih ovir (večina posnetkov na spletu je v tujih jezikih), razen če gre za medpredmetno povezovanje, se mora učitelj potruditi in videoposnetek pripraviti sam (pogosto s snemanjem zaslona, npr. Screencast-O-matic), ga objaviti (npr. YouTube) in deliti z učenci (npr. v spletni učilnici). Pri tem se mora zavedati, da bistvo obrnjenega učenja ni v posnetku, ampak v nalogah, izzivih, ki jih učenci rešujejo s pomočjo znanja, pridobljenega ob gledanju posnetka.

Kot prednosti obrnjenega učenja učenci[10] navajajo kratko, strnjeno razlago, ki si jo lahko večkrat ponovijo (tudi pri reševanju nalog) in po potrebi ustavijo. 89,3 % vprašanih učencev je menilo, da jim posnetek učitelja pomaga pri razumevanju snovi, 92,8 % pa, da so jim koristila vprašanja, ki so jih usmerjala ob posnetku ter jih spodbujala k sprotni pripravi zapiskov. Dijakinja je v refleksiji strnjeno zapisala: "Ta način učenja je veliko boljši, kot pa da profesor v šoli razloži snov, nato pa moramo sami doma reševati učne liste, saj tako nimamo možnosti vprašati za nasvet. V redu se mi zdi, da lahko video ustavimo, si zapišemo informacije, slabost pa bi mogoče bila, da naša koncentracija morda ni enaka kot pri pouku v razredu." Zanimivo je, da se tudi mnenja o koncentraciji pri gledanju posnetkov razlikujejo, saj na drugem mestu zasledimo: "… tak način dela mi je všeč, saj ni razgrajačev, ki bi motili pouk, lahko si ponovno zavrtim del videa, ki ga nisem razumela in si lahko poiščem dodatne informacije s spleta/učbenika."

2.6 Nove možnosti zbiranja podatkov

Nove možnosti zbiranja podatkov se ne nanašajo le na uporabo spletnih brskalnikov, pač pa tudi na možnost izdelave in uporabe spletnih anket in vprašalnikov (npr. Google Obrazci), ki jih učenci uporabijo pri seminarskih in raziskovalnih nalogah. Z njihovo uporabo je proces pridobivanja in analiziranja grafičnega prikazovanja podatkov hitrejši, vzorci pa večji.

2.7 Zbiranje in uporaba podatkov o učenju in poučevanju

Zbiranje in uporaba podatkov o učenju in poučevanju za učitelja pomeni informacijo o napredku in dosežkih učencev. Pri tem se mu ni treba zanašati le na različne vrste preverjanja znanj (npr. Kahoot, kvizi v spletni učilnici Moodle ...), ampak tudi večina drugih aplikacij (npr. Zaption – interaktivni video posnetki, Actively Learn – delo z viri) omogoča dve stvari:

- takojšnjo povratno informacijo učencu, na osnovi katere lahko le-ta v postopku samorefleksije uravnava in načrtuje svoje nadaljnje učenje;
- hkrati pa tudi t. i. learning analytics, le-ta pa učiteljevo kakovostno povratno informacijo in uravnavanje pouka v smeri prilagajanja posameznim učencem oziroma skupinam učencev (*Vodenje pouka z uporabo podatkov o napredku učencev*, 2012: 93).

Takojšnjo povratno informacijo (učitelju in učencu) omogočajo tudi odzivni sistemi (npr. glasovalne naprave ali aplikacije, kot so Socrative ali Kliker, do katerih učenci dostopajo na svojih mobilnih napravah), s tem pa tudi prilagajanje pouka sproti, med samo izvedbo.

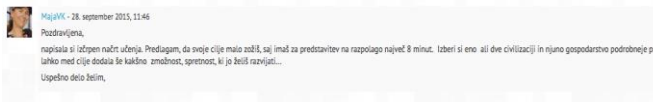
Kakovostna povratna informacija je (poleg drugega) tudi hitra (čimprejšnji odziv) in v smislu dolgoročnega razvoja tudi trajna. Takojšnji, verbalni odziv je seveda pomemben, toda dolgoročni napredek učenja je viden šele ob zbirki učenčevih dosežkov, samorefleksij in refleksij učitelja in sošolcev. Digitalna tehnologija (v tem primeru e-listovnik, npr. Mahara) tudi pri pouku zgodovine

[8] Med 82 učenci 3. letnika ekonomske gimnazije SEŠ (šolsko leto 2014/15) jih le 9,4 % redno uporablja skupno urejanje dokumentov, 22,4 % te možnosti ne pozna, naslednjih 22,4 % pa jo sicer pozna, a ne uporablja.

[9] Personalizacija pouka v središče procesa postavlja učenca kot posameznika, ki identificira svoje učne cilje, odloča, kdaj, kaj in kako se bo učil, ter zato analizira in ozavešča svoj učni stil, sledi lastni učni poti, sam izbira, katero tehnologijo in kako jo bo uporabil. Učitelj je v funkciji mentorja, za posameznike zagotovi različna navodila, pri učencih spodbuja metaučenje in samoregulacijo (*The Book of Trends 2.0*, 2015: 11).

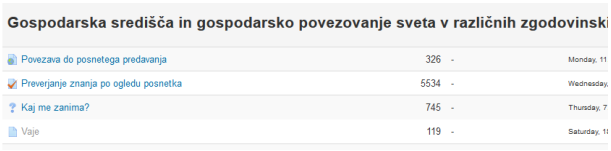
[10] Vprašalnik je izpolnjevalo 28 učencev 3. letnika ekonomske gimnazije na SEŠ Maribor v januarju 2014, po pol leta izkušenj z obrnjenim učenjem.

omogoča razvoj in spremljanje zmožnosti razumevanja konceptov in razvijanje veščin.



Slika 3: Primer učiteljeve povratne informacije v Mahari.

V nasprotju z listovnikom, kjer je učenec tisti, ki z učiteljem deli svoje izdelke in refleksije, so spletna učna okolja (sistemi za upravljanje izobraževanja in učnih vsebin) tisti prostor, kjer se srečujejo vsebine, različne možnosti komunikacije na daljavo in orodja za upravljanje in organiziranje izobraževalnega procesa (Bregar et al., 2010: 154). V slovenskem prostoru se med spletnimi učnimi okolji največ uporabljajo prostodostopni Moodle (v zadnjem času tudi Edmodo), med tržnimi pa Blackboard in WebCT. Učna okolja omogočajo dajanje sprotnih povratnih informacij, spremljanje dejavnosti učitelja in učencev ter prilagajanje poteka poučevanja (in učenja).



Slika 4: Eden izmed možnih prikazov aktivnosti v spletni učilnici Moodle.

Nekateri učitelji se želijo tudi v virtualnem okolju približati učencem z uporabo družabnih omrežjih kot učnih okolij (npr. Facebook), vendar v tem primeru zbiranje podatkov o učenju in poučevanju ni možno, oziroma je zelo oteženo.

3 Sklep

Uporaba digitalne tehnologije omogoča aktivno sodelovanje učencev v procesu učenja, z orodji, ki so jim blizu. V sodobni šoli 21. stoletja mora učitelj pazljivo uravnotežiti vsebino in uporabljeno tehnologijo z učnimi cilji ter pričakovanimi dosežki ter pri tem ohraniti v učenca in učenje – ne v tehnologijo – usmerjen pristop. Uporaba tehnologije pri pouku ne more in ne sme postati ”edutainment”, pač pa sredstvo za doseganje izobraževalnih ciljev, kadar je to mogoče in smiselno, hkrati pa tista pot, ki bo generacijam milenijcev pokazala, da digitalna tehnologija ni le sredstvo za zabavo, druženje in samopromocijo, pač pa (tudi) orodje za delo, pridobivanje in širjenje informacij ...

Zavedati se moramo, da je izobraževanje izjemno pomemben, a le manjši del učenja in pridobivanja znanja ter izkušenj. Virtualni svet in digitalne tehnologije postavljajo pred šolo nove izzive, še posebej pri humanističnih predmetih, ki pomagajo razvijati refleksijo in samorefleksijo ter s tem osebno zavest in družbeno odgovornost, kritično distanco do sodobnih družbenih pojavov, med katerimi je (eden od prevladujočih in izjemno vplivnih) tudi digitalna tehnologija.

4 Literatura

Lea Bregar, Margerita Zagmeister in Marko Radovan. 2010. *Osnove e-izobraževanja*. Andragoški center Slovenije, Ljubljana.

Vilma Brodnik. 2015. *Smernice za uporabo IKT pri predmetu zgodovina.* http://www.inovativna-sola.si/images/inovativna/Smernice/ZGODOVINA_smernice_IKT.pdf.

Allan Carrington. 2015. The Pedagogy Wheel: It's about transformation and integration. V: *Support of Excellence.* http://designingoutcomes.com/allansportfolio/edublog/?p=836.

Sabina Čadež. 2012. *Vodenje pouka z uporabo podatkov o napredku učencev.* V: Vodenje v vzgoji in izobraževanju. Šola za ravnatelje.

Simon Dražič. 2015. Internetni mehurček. *SIRikt 2015: Arnes; Gradiva za obrnjeno učenje in varn šolske ure.* [video]. Kranjska Gora: Arnes, 27. maj 2015.

Evropska komisija. 2007. Ključne kompetence za vseživljensko učenje. Evropski referenčni okvir. http://bookshop.europa.eu.

Andrej Flogie et al. 2014. *Informacijska tehnologija kot temelj vseživljenskega izobraževanja človeka 21. stoletja.* [ured.] Domen Kovačič. Maribor: Zavod Antona Martina Slomška.

Vojko Kunaver et al. 2008. *Učni načrt. Zgodovina: gimnazija: strokovna gimnazija*. http://eportal.mss.edus.si/msswww/programi2016/programi/media/pdf/un_gimnazija/un_zgo_210_ur_strok_gimn.pdf.

Punya Mishra in Matthew P. Koehler. 2006. *Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge.* Teachers College, Columbia University, 2006, Teachers College Record, Izv. 108, str. 1017–1054.

Richard E. Mayer. 2013. Učenje s tehnologijo. [ured.] Hana Dumont, Dvid Istance in Francisco Benavides. *O naravi učenja: uporaba raziskav za navdih prakse.* ZRSŠ, 2013.

Mark E. Moram. 2011. The Top 10 Reasons Students Cannot Cite or Rely On Wikipedija. *FindingDulcinea.* http://www.findingdulcinea.com/news/education/2010/march/The-Top-10-Reasons-Students-Cannot-Cite-or-Rely-on-Wikipedia.html.

Lea Nemec. 2015. *Pojmovne mreže- ključ do učenja z razumevanjem.* Ljubljana. EDUvision, 2015. Sodobni pristopi poučevanja sodobnih generacij. str. 66–74.

Marc Prensky. 2001. *Digital Natives, Digital Immigrants.* NCB University Press. Izv. 9.

Mike Sharples, Josie Taylor in Giasemi Vavuola. 2006. *A theory of learning for the Mobile Age.* Sage publications,

Julee M. Tanner. 2014. Digital vs. Print: Reading comprehension and the Future of Book. *SJSU School of Information Student Research Journal.* http://scholarworks.sjsu.edu/slissrj/vol4/iss2/6.

The Book of Trends 2.0. 2015. A Young Digital Planet, str. 13–16.

Maja Vičič Krabonja. 2010. *Kombinacija različnih IKT-orodij v srednji šoli na primeru predmeta zgodovina.* V: Splet izobraževanja in raziskovanja z IKT - Sirikt 2010 (zbornik).

Generiranje kritičnih prepisov s strojnim prevajanjem na ravni znakov

Katja Zupan, Tomaž Erjavec

Odsek za tehnologije znanja, Institut »Jožef Stefan« in
Mednarodna podiplomska šola Jožefa Stefana
Jamova cesta 39, 1000 Ljubljana
katja.zupan@ijs.si
tomaz.erjavec@ijs.si

Povzetek

Pomembnejši rokopisi so pogosto predstavljeni v dveh prepisih, diplomatičnem in kritičnem, kjer prvi sledi izvirniku, drugi pa ga interpretira, pri čemer tudi delno posodobi njegovo besedilo. Izdelava kritičnega prepisa je zamuden postopek, ki bi ga lahko olajšali z avtomatskimi metodami. V prispevku predstavimo orodje, ki temelji na statističnem strojnem prevajanju znakov in prevaja posamezne vrstice diplomatičnega prepisa v kritičnega. Metodo smo preizkusili na dveh pridigah (1825, 1829) A. M. Slomška in jo primerjali z več drugimi pristopi. Preizkusi pokažejo, da program, naučen na prvi pridigi in preizkušen na drugi, zmanjša delež razlik na ravni znakov za skoraj dve tretjini in deluje bolje kot preostale metode.

Generating Critical Transcriptions with Character-based SMT

Historic manuscripts are often represented in the form of two transcriptions, a diplomatic and a critical one. The former follows closely the original while the latter interprets it, also by partly modernising its text. Generating a critical transcription is a time-consuming process, which could be improved through automatic methods. The paper presents a tool that uses character-based statistical machine translation, translating each line of the diplomatic transcription into the critical one. The method has been tested on two sermons (1825, 1829) by A. M. Slomšek, and compared to several other approaches to text modernisation. The experiments show that training a program on the first sermon and testing it on the second one reduces the character error rate by nearly two thirds, performing better than other methods.

1 Uvod

Napredek jezikovno usmerjenih informacijskih tehnologij vpliva na razvoj številnih področij humanistike. Na področju besedilne kritike je digitalizacija besedil odprla nove možnosti zapisa, raziskovanja in prikaza in s tem spodbudila spremembo paradigme, kako besedilo predstaviti in posredovati uporabnikom.

Ponoven razcvet je tako doživela tudi metodologija znanstvenih oz. znanstvenokritičnih izdaj, kot se v literarnih vedah imenujejo tiste edicije – najpogosteje pomembnejših starejših rokopisov –, v katerih so izvirna besedila (oz. njihovi faksimili) pregledana, prepisana, rekonstruirana, komentirana in naposled objavljena po načelih tekstne kritike (Erjavec in Ogrin, 2004). Takšna besedila so običajno predstavljena v obliki dveh vrst prepisov, diplomatičnega in kritičnega.

Diplomatični prepis je v tiskani/digitalni obliki čim bolj natančen tipografski dvojnik rokopisa z vsemi avtorjevimi nedoslednostmi, napakami, avtorjevimi popravki, vrivki ipd. vred. Njegova funkcija je, da predstavlja izvirnik in tako olajša branje težje berljivih mest rokopisa. Kratice, nejasna in poškodovana mesta rokopisa pušča nerazrešena in nedopolnjena.

Kritični prepis je že tekstnokritična interpretacija besedila, ki izvirno besedilo redigira in nekoliko modificira na črkopisni in oblikoslovni ravni, vendar se pri tem ravna po ekspliciranih načelih, pridobljenih z indukcijo iz analize samega besedila. Vsebuje tudi ustrezen t. i. tekstnokritični aparat z opombami. (Dović, 2006, str. 210). Diplomatični prepis poskuša torej čim zvesteje slediti izvirniku, kritični pa z modificiranjem jezikovnega izraza skuša doseči ravnotežje med ohranjanjem avtentičnosti jezika in razumljivostjo sodobnemu bralcu oz. uporabniku. Modificirani jezikovni izraz torej ni nujno sodobni standardni jezik, temveč mu je samo približan. Stremi k ohranjanju arhaičnega vtisa diplomatičnega prepisa, na primer z (delnim) ohranjanjem starinskih besed, hkrati pa jih poda v sodobni pisni podobi, tj. v sodobnem črkopisu (npr. gajici).

Uredniško delo pri ustvarjanju kritičnega prepisa tako zajema tudi razmeroma preproste in sistematične posege, kot je posodabljanje črkopisa, oz. sistematične posege na oblikoslovno-leksikalni ravni. V pričujočem prispevku raziskujemo, kako bi bilo mogoče tovrstne posege izvajati strojno, s pomočjo prilagojene računalniške metode strojnega prevajanja znakov, ki bi na nezahteven način samodejno predlagala ustrezno (delno) posodabljanje besed ter s tem poenostavila in pospešila uredniško delo.

2 Strojna normalizacija besed

Za posodabljanje (ali, širše, normalizacijo) starejših besedil je bilo razvitih že več računalniških metod (Piotrowski, 2012), od uporabe ročno napisanih pravil za pretvorbo starejših oblik v sodobni oz. normalizirani zapis (Erjavec, 2015), prek avtomatske izpeljave tovrstnih pravil (Bollmann, 2012) do uporabe statističnega strojnega prevajanja na ravni znakov, ki »prevaja« arhaične zapise besed v sodobne oz. normalizirane (Tiedemann, 2009; Scherrer in Erjavec, 2015).

Vsem tem metodam je skupno, da je posodabljanje izvedeno na ravni posamezne pojavnice (besedne oblike) in da se zanaša na obstoj velikega leksikona sodobnega jezika, da izloči hipoteze, ki jih sistem generira, ker so sicer sistemsko možne, a niso del besedja določenega jezika.

Obe domnevi sta problematični, če ju uporabimo pri izdelavi kritičnega prepisa starejših besedil. Pogosta razlika med arhaičnim in posodobljenim zapisom je namreč zapis skupaj oz. narazen, tj. kot dve besedi na ortografski ravni. Vsak sistem, ki besedilo najprej razdeli na pojavnice in nato

izvede njihovo pretvorbo, bo v teh primerih zatajil. Druga težava, ki še posebej velja za kritične izdaje, pa je neuporabnost leksikona sodobnega jezika, saj v njem ne bo arhaično obarvanih besed, ki jih običajno najdemo v kritičnih prepisih.

3 Metodologija

3.1 Strojno prevajanje na ravni znakov

Statistično strojno prevajanje (SSP oz. ang. Statistical Machine Translation, SMT) (Brown et al. 1993, Koehn, 2010) je sklop metod, ki temelji na učenju (kombinacije) modela prevajanja in modela ciljnega jezika. Prvega gradimo na vzporednih korpusih, ki so tokenizirani in večinoma poravnani po povedih, drugega pa na enojezičnem korpusu jezika, v katerega prevajamo. Naučeni sistem je nato sposoben prevajati povedi izvornega jezika v ciljnega. Na tem principu temelji Googlov prevajalnik, v okviru več (tudi zelo velikih) projektov pa je bil izdelan odprtokodni sistem Moses, ki ga je razmeroma preprosto namestiti na računalnik in uporabiti tako za učenje kot tudi izvajanje prevajanja. Moses je kompleksno orodje, ki omogoča prilagojene nastavitve veliko parametrov in izbiro različnih metod pri prevajanju, kjer je od lastnosti obeh jezikov in razmerja med njima odvisno, katere dajo najboljše rezultate. Tudi za slovenščino so bile metode SSP že uporabljene (Vičič, 2002; Sepesy Maučec et al., 2006; Brest, 2009; Verdonik, 2013; Sepesy Maučec et al., 2013; Dugonik, 2013).

Predlagamo preprosto metodo, ki temelji na SSP, a s spremenjenim načinom prevajanja odpravlja zgoraj opisane pomanjkljivosti. Pri eksperimentih smo uporabili Moses (Koehn et al., 2007), najpogosteje uporabljani odprtokodni sistem za strojno prevajanje.

Medtem ko je v klasičnem pristopu k strojnemu prevajanju osnovna enota beseda (pojavnica) oz. besedna zveza (SSP-B), pa statistično prevajanje na ravni znakov (SSP-Z) deluje tako, da znake obravnava, kot bi bili besede, potrebne prevoda. Na praktični ravni to pomeni, da besedilo prilagodimo za obdelavo v Mosesu enostavno tako, da med posamezne črke vstavimo presledke, nekdanje presledke (meje med besedami) pa označimo s posebnim znakom, npr. podčrtajem. Pomanjkljivost SSP-B je tudi ta, da zna prevajati samo besede oz. besedne zveze, ki so bile vključene v učno množico. SSP-Z sicer prevaja na enak način, a ker je nabor znakov veliko manjša končna množica kot nabor besed posameznega jezika, bodo v veliki večini že zastopani v učnem modelu in jih bo tako sistem znal prevesti. Kot rečeno, je tovrstni znakovni pristop že uveljavljen način za prevajanje zelo sorodnih jezikov oz. jezikovnih različic, kjer so razlike večinoma omejene na ortografsko raven (Vilar, 2007; Nakov in Tiedemann, 2012; Scherrer in Erjavec, 2013; Pettersson et al., 2013; Sánchez-Martínez et al., 2013).

Kar loči našo metodo od obstoječih pristopov, je način prevajanja znakov. Običajni način prevaja posamezne črke (znake), iz katerih je sestavljena beseda (pojavnica), ne upošteva pa ko(n)teksta. Beseda je namreč zanj zaključena enota, kot bi bil v klasičnem prevajanju stavek, zato ne sega prek besedne meje. Naša metoda poskuša preseči to omejitev, zato kot »stavek« ne vzame besede, temveč širšo besedilno enoto. V našem primeru smo se odločili, da bodo to posamezne vrstice, saj so te praviloma označene (in s tem poravnane) tako v diplomatičnem kot v kritičnem prepisu.

Metoda se opira na predpostavko, da ima del diplomatičnega prepisa že izdelan svoj kritični prepis oz. »prevod«, ta vzporedni korpus pa je nato uporabljen kot učna množica, na kateri se bo učil prevajalski model. Ciljni jezikovni model je naučen na že izdelanem delu kritičnega prepisa, s predpostavko, da bo deloval bolje brez opiranja na zunanje vire, kot sta korpus in/ali leksikon sodobnega jezika.

3.2 Slomškovi pridigi kot učna in testna množica

Da bi preizkusili delovanje metode, smo izvedli niz eksperimentov na podatkovni množici, vzeti iz digitalne znanstvenokritične izdaje »Treh pridig o jeziku« (Faganel et al., 2004), delo Antona Martina Slomška (1800–1862), znanega slovenskega škofa, pedagoga in pisatelja, pa tudi pomembnega reformatorja slovenskega kulturnega, narodnostnega in verskega življenja – zlasti v vzhodni Sloveniji in na Koroškem –, ki je opozarjal na vpliv agresivne germanizacije in skušal s svojo dejavnostjo zmanjšati ta vpliv (Erjavec in Ogrin, 2004).

Digitalna izdaja je del širše zbirke, imenovane *Elektronske znanstvenokritične izdaje slovenskega slovstva* (eZISS, http://nl.ijs.si/e-zrc/), ki se je kot skupni projekt začela graditi leta 2001 pod vodstvom Matije Ogrina z Inštituta za slovensko literaturo in literarne vede ZRC SAZU ter Tomaža Erjavca z Odseka za tehnologije znanja na IJS, avtor obeh vrst prepisov Slomškovih pridig pa je Jože Faganel.

Vse digitalne izdaje eZISS so zapisane v skladu s smernicami za kodiranje besedil TEI (Konzorcij TEI), torej v zapisu XML, usklajenem s shemo, ki je parametrizacija smernic TEI za namene projekta eZISS.

Izdaja Slomškovih del je v zbirki zastopana s tremi pridigami kot primeri retorske proze v prvi polovici 19. stoletja, ki še posebno simbolizirajo njegovo delo in prizadevanja za širšo rabo slovenskega jezika. Gre za naslednje pridige:

1) »Za krščansko govorjenje« (1825);
2) »Jezik je vir dobrega in zla« (1829);
3) »Svoj jezik je treba spoštovati« (1838).

Prvi dve pridigi sta ohranjeni v avtorjevem rokopisu, rokopis tretje, ki je najbolj znana, pa je izgubljen, a je bilo besedilo pridige kmalu po nastanku dvakrat natisnjeno. Elektronska izdaja vsebuje predgovor, faksimile, diplomatični prepis (razen za 3. pridigo), kritični prepis in urednikove opombe. Vsi prepisi so povezani s faksimili, medsebojno pa so povezani po vrsticah. To omogoča vzporedni prikaz faksimilov s prepisi, kakor tudi vzporedni prikaz faksimila z obema prepisoma.

Ker za tretjo pridigo ni na voljo diplomatičnega prepisa, smo jo izločili iz eksperimenta. Prvi dve pridigi sta nastali v približno istem časovnem obdobju, ko se v slovenskem jeziku še ni oblikoval standardni zapis besed, niti se še ni začel ključni proces poenotenja zapisa, saj se je to zgodilo šele sredi stoletja s t. i. novimi oblikami.

Za ponazoritev jezika, ki ga je uporabljal Slomšek, in njegovega kritičnega prepisa poda slika 1 tri vrstice iz prve pridige.

Diplomatični prepis	Kritični prepis
3.She enkrat bel ſe' veſeli duſha kerſhanſka! Kader ſvete	3.Še enkrat belj se vesêli duša krščanska! Kader svete
godove ſvetnikov obhajaſh, s'akaj ravno oni ſo tebi taji‸	godove svetnikov obhajaš, zakaj ravno oni so tebi taji-
ſte ſvetle s'ves'de, ki ſe is'ſ. nebeſ na tebe os'irajo, tebi	ste svetle zvezde, ki se iz sv. Nebes na tebe ozirajo, tebi

Slika 1: Primer diplomatičnega in kritičnega prepisa pridige A. M. Slomška v izdaji eZISS.

Primer pokaže, da razlike med prepisoma po eni strani niso omejene samo na prečrkovanje bohoričice v gajico, pač pa zajemajo tudi druge premene v besedah (npr. *kerſhanſka (→ keršanska) → krščanska*) kot tudi spremembe ločil (npr. ‸ → -), po drugi strani pa ponazarja, da kritični prepis ne uporablja sodobnega standarda, pač pa v mnogo primerih še vedno ohrani starinsko obarvani zapis besed (*Kader, tajiste*). Dodatno je pomembno, da kritični prepis ohrani deljene besede na koncu vrstic, kar je tudi problematično za klasične metode, ki se opirajo na posodabljanje posameznih besed.

Za preizkus predlagane metode smo izdelali učni in testni korpus. Da bi bila evalvacija bolj realistična, nismo premešali vrstic obeh pridig, pač pa smo kot učni korpus vzeli prvo pridigo, kot testnega pa drugo.

Iz izvornega zapisa TEI smo odstranili vse oznake XML (npr. poudarjeno besedilo) ter upoštevali avtorjeve popravke besedila. S tem smo dobili dva vzporedna korpusa, ki sta poravnana po vrsticah in vsebujeta samo golo besedilo, točno tako, kot je predstavljeno v sliki 1.

Velikost učnega in testnega korpusa je podana v tabeli 1, ki pokaže, da je testna pridiga več kot za tretjino večja od učne, pri čemer prva vsebuje približno 2.000 besed, druga pa 3.300.

	Učna pridiga (1825)		*Testna pridiga (1829)*	
Prepis:	*Dipl.*	*Krit.*	*Dipl.*	*Krit.*
Vrstic:	252	252	362	362
Pojavnic:	2.036	2.081	3.393	3.346
Znakov:	13.958	12.809	23.353	20.732
DR-Z	20,68		22,26	

Tabela 1: Velikost uporabljene podatkovne množice in delež razlike na ravni znakov med dipl. in krit. prepisom.

Za evalvacijo bomo uporabili delež razlik na ravni znakov (DR-Z), ki meri povprečno Levenshteinovo razdaljo (razliko) med avtomatsko generiranim kritičnim prepisom in ročno izdelanim kritičnim prepisom pridige, uporabljene kot testne množice. Razdalja se meri v številu t. i. posegov – kamor spadajo vstavljanje, brisanje ali zamenjava znaka (Levenshtein, 1966) – ki je potrebno, da izhodiščno besedilo »popravimo« v ciljnega. Razmerje med številom posegov in številom znakov v referenčnem (ciljnem) besedilu izrazimo v odstotnih deležih, ki bodo naša evalvacijska mera učinkovitosti delovanja posamezne metode.

Kot izhodiščno mero evalvacije, ki jo bomo skušali izboljšati, smo vzeli razdaljo med diplomatičnim prepisom, če v njega ne posegamo z nobeno izmed metod (in ga tako obravnavamo kar kot neke vrste strojno generirani kritični prepis), ter med ročno izdelanim kritičnim prepisom. Kot pokaže tabela 1, delež razlik v znakih med njima v testnem korpusu znaša 22,62 odstotka, kar je nekaj več kot v učni množici, kjer je 20,68 odstotka.

3.3 Opis eksperimenta

Preizkusili smo tri metode:

- *statistično strojno prevajanje na ravni znakov (SSP-Z)*, pri čemer je prevodna enota ena vrstica;
- *statistično strojno prevajanje na ravni besed* (SSP-B) in
- *kombinacijo več normalizacijskih metod v Normi.*

Normo (Bollmann et al., 2012) smo se odločili uporabiti za primerjavo s SSP, ker gre za eno bolj znanih odprtokodnih orodij za normalizacijo. Orodje združuje več metod normalizacije, vendar pa je sestavljeno iz različnih zunanjih modulov, imenovanih »normalizatorji«, ki predstavljajo različne metode normalizacije (mdr. avtomatsko naučena pravila, Levenshtein, historični leksikon) in jih je mogoče uporabiti ali izključiti iz procesne verige. Privzete nastavitve za svoje delovanje potrebujejo dvojezični historični leksikon (diplomatična in njena kritičnoprepisna besedna oblika) ter enojezični leksikon sodobnega jezika.

Pri obeh metodah strojnega prevajanja smo preizkusili tri različice modelov ciljnega jezika, in sicer smo kot učne modele vključili:

A. zgolj besedilo pridige, uporabljene za učno množico;
B. poleg A tudi kritične prepise avtorjevih drugih del (tj. zbrana dela A. M. Slomška);
C. poleg A tudi jos100k (Erjavec et al., 2010), referenčni korpus za jezikoslovno označevanje slovenskega jezika.

Poleg različic jezikovnega modela smo pri strojnem prevajanju preizkusili tudi različne stopnje jezikovnega modela, t. i. n-grame. Ti zaznamujemo dolžino enot, ki jih sistem obravnava kot potencialne besedne zveze, kar v našem primeru pomeni dolžino niza zaporednih znakov (črk, cifer, ločil ipd.). Dolžino smo omejili v razponu od dveh do petih zaporednih znakov (tj. 2-, 3-, 4- in 5-grami), saj so preizkusi modelov z enim znakom in več kot petimi pokazali, da ti delujejo slabše kot modeli v izbranem razponu.

4 Rezultati

Rezultati, podani v tabeli 2, pokažejo, da nam izhodiščni delež razlik uspe zmanjšati z vsemi tremi metodami, tako z obema vrstama strojnega prevajanja kot z Normo. Najboljši rezultat (DR-Z = 7,59 %) je dosegla predlagana metoda, tj. SSP-Z z osnovnimi nastavitvami sistema Moses, najpreprostejšim modelom ciljnega jezika (zgolj besedilo učne pridige) in najnižjo stopnjo modela (2-grami). V obravnavanem primeru to pomeni, da če bi si prepisovalec pomagal s SSP-Z, bi si prihranil dve tretjini dela, kar zadeva popravke na ravni posameznih znakov, v primerjavi s tem, da bi za izhodišče pri izdelavi kritičnega prepisa druge pridige vzel kar diplomatični prepis in ga nato popravljal povsem ročno.

Medtem ko z najosnovnejšim in najkompleksnejšim jezikovnim modelom (A in C) najbolje delujejo 2-grami kot osnovna enota prevajanja, pa se pri učnem modelu s širšim naborom Slomškovih del kot najučinkovitejši kažejo nizi treh zaporednih znakov, daljši nizi od treh znakov pa v vseh primerih dosegajo slabše rezultate, kar nakazuje na to, da je večina sprememb med diplomatičnim in kritičnim prepisom omejena na razpon treh znakov. Če upoštevamo, da je najbolj tipični uredniški poseg posodobitev črkopisa, je rezultat mogoče pojasniti s tem, da je tipična sprememba bohoričice v gajico omejena na zamenjavo največ dveh grafemov z enim (npr. *zhaft* > *čast*). Če primerjamo posamezne jezikovne modele, ugotovimo, da je pri vseh tipih n-gramov najslabši model C, kar kaže na to, da sodobni jezik kot model ciljnega jezika ni primeren za kritični prepis, saj ta ne stremi k izrazni podobi sodobnega jezika.

Da bi določili, ali je za generiranje strojnega prepisa zares bolj smiselno uporabiti prilagojeno metodo strojnega prevajanja, tj. prevajanje znakov, smo preizkusili tudi, kakšne rezultate daje strojno prevajanje na ravni besed (SSP-B). Generirani prepis smo zaradi primerljivosti evalvirali na enak način, tj. glede na delež razlik v znakih, in ne kot odstotek napačno prevedenih besed, kar je običajna metoda evalvacije klasičnih strojnoprevajalskih sistemov. Eksperiment je pokazal, da klasična metoda za obravnavani primer ni ustrezna, saj deluje za polovico slabše kot SSP-Z: medtem ko prevajanje znakov prihrani do dve tretjini dela, ga prevajanje besed samo tretjino. Spreminjanje stopnje jezikovnega modela je imelo pri SSP-B zanemarljiv vpliv na rezultat, saj je ta pri vseh n-gramih praktično enak; izjema je nekoliko boljši rezultat 2-gramov, kar potrjujejo že izsledki pri SSP-Z, le da gre tu za niz dveh besed in ne znakov. Poleg tega osnovna enota prevajanja ni bila več celotna vrstica, temveč smo besednim oblikam iz diplomatičnega prepisa samodejno določili njihove kritičnoprepisne ustreznice s pomočjo Gize++, orodja za poravnavo vzporednih prevodov.

S pomočjo poravnavnih tabel Gize (t. i. datoteka e2f) smo na ta način tudi strojno izdelali dvojezični leksikon in ga skupaj z leksikonom sodobnih besednih oblik Sloleks (Dobrovoljc et al., 2015) dodali v Normo, sicer pa uporabili privzete nastavitve. Orodje deluje za skoraj tri odstotne točke bolje od SSP-B, a še vedno skoraj šest odstotnih točk slabše kot SSP-Z.

	2-grami	*3-grami*	*4-grami*	*5-grami*
SSP-Z, model A	**7,59**	8,58	8,71	9,18
SSP-Z, model B	8,05	**8,14**	8,21	8,80
SSP-Z, model C	**8,26**	8,45	8,48	9,50
SSP-B	**16,84**	16,87	16,86	16,87
Norma:	14,19			
Izhodišče:	22,26			

Tabela 2: Odstotni delež razlik v znakih glede na različico in stopnjo modela ciljnega jezika na ravni znakov oz. besed ter prim. z Normo in izhodiščem.

5 Sklepne misli in prihodnje delo

Preizkus strojnega generiranja kritičnega prepisa je pokazal, da je mogoče že s privzetimi nastavitvami odprtokodnega sistema za strojno prevajanje Moses ter preprostim prevajalskim in jezikovnim modelom, delujočim na ravni znakovnih nizov, uredniku izdaje prihraniti do dve tretjini dela, če ga ovrednotimo kot število potrebnih posegov v besedilo.

Metodo je mogoče uporabiti tudi v primerih, ko ima sorazmerno majhen del diplomatičnega prepisa že izdelan svoj kritični prepis, saj metoda SSP-Z za svoje delovanje ne potrebuje velike učne množice, kar dokazuje tudi obravnavani primer relativno kratkega pridižnega besedila, kjer SSP-B deluje za polovico slabše, le nekoliko boljša je kombinacija »neprevajalskih« metod v Normi.

V širšem smislu bi izhodno besedilo metode SSP-Z lahko uporabili tudi za preučevanje sistemskih in nesistemskih posegov v diplomatični prepis pri »posodabljanju« za kritični prepis, torej postopkov na ortografski ravni, uporabljenih za približanje razumevanja besedila sodobnemu bralcu. Vpogled v sistemskost bi bil uporaben tudi za izboljšanje kakovosti procesiranja starejših besedil z avtomatskimi orodji za jeziko(slov)no označevanje. Pri nadaljnjem razvoju metode nameravamo preizkusiti metodo na novih primerih uporabe in dodatnih možnostih združevanja učnih modelov v okviru orodja Moses, nato pa uporabiti še nekoliko zahtevnejše nastavitve parametrov, ki jih omogoča Moses, kot je npr. uglaševanje parametrov namesto privzetih nastavitev.

6 Zahvala

Avtorja se zahvaljujeta anonimnim recenzentom za koristne pripombe. Raziskava, opisana v prispevku, je bila opravljena v okviru raziskovalnega programa Tehnologije znanja P2-0103 in programa Mladi raziskovalci (št. 37487), ki ju financira ARRS.

7 Literatura

Marcel Bollmann. 2012. (Semi-)Automatic Normalization of Historical Texts using Distance Measures and the Norma tool. V: *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*, str. 3–14, Lizbona, Portugalska. https://www.linguistics.ruhr-uni-bochum.de/comphist/pub/acrh12.pdf.

Janez Brest. 2009. Statistično strojno prevajanje iz slovenščine v angleščino. V: *Zbornik povzetkov delavnic "Algoritmi po vzorih iz narave" v študijskem letu 2008/2009*, str. 15, Ljubljana, Slovenija.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer. (1993) The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics,* 19(2): 263–311. MIT Press.

Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec in Miro Romih. 2015. *Morfološki leksikon Sloleks 1.2*. Slovenski repozitorij jezikovnih virov CLARIN.SI. http://hdl.handle.net/11356/1039.

Marijan Dović. 2006. Tekstualna tradicija in elektronski medij: od digitalne slikovne reprodukcije do znanstvenokritične izdaje. V: *Sborník prací Filozofické fakulty brněnské univerzity, Studia minora Facultatis philosophicae Universitatis Brunensis*. 9: 208–215.

https://digilib.phil.muni.cz/bitstream/handle/11222.digilib/102980/X_SlavicaLitteraria_09-2006-1_25.pdf?sequence=1.

Jani Dugonik. 2013. *Uglaševanje parametrov pri statističnem strojnem prevajanju*. Magistrsko delo, Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru. http://dkum.uni-mb.si/IzpisGradiva.php?id=40979.

Tomaž Erjavec in Matija Ogrin. 2004. E-Slomšek: elektronska znanstvenokritična izdaja retorske proze 19. stoletja po standardu XML TEI. V: *Jezikovne tehnologije: zbornik B*, str. 87–93. http://nl.ijs.si/e-zrc/bib/sdjt04-16erjavec.pdf.

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

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. http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf.

Jože Faganel, Matija Ogrin in Tomaž Erjavec. 2004. Anton Martin Slomšek: Tri pridige o jeziku: elektronska znanstvenokritična izdaja. Inštitut za slovensko literaturo in literarne vede, ZRC SAZU, Ljubljana. http://nl.ijs.si/e-zrc/slomsek/.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Praga, Češka.

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

Konzorcij TEI (ur.). *Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/P5/.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8): 707–710.

Preslav Nakov in Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. V: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, str. 301–305, Stroudsburg, Pennsylvania, ZDA. Association for Computational Linguistics. http://anthology.aclweb.org/P/P12/P12-2.pdf#page=329.

Eva Pettersson, Beáta Megyesi in Jörg Tiedemann. 2013. An SMT approach to automatic annotation of historical text. V: *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013, NEALT Proceedings Series*, 18: 54–69. http://www.ep.liu.se/ecp/087/005/ecp1387005.pdf.

Michael Piotrowski. 2012. *Natural language processing for historical texts* (Synthesis Lectures on Human Language Technologies, ur. Graeme Hirst, vol. 17). Morgan & Claypool Publishers.

Felipe Sánchez-Martínez, Isabel Martínez-Sempere, Xavier Ivars-Ribes in Rafael C. Carrasco. 2013. An open diachronic corpus of historical Spanish: annotation criteria and automatic modernisation of spelling. ArXiv:1306.3692. http://www.dlsi.ua.es/~fsanchez/pub/pdf/sanchez-martinez13a.pdf.

Yves Scherrer in Tomaž Erjavec. 2013. Modernizing historical Slovene words with character-based SMT. V: *BSNLP 2013-4th Biennial Workshop on Balto-Slavic Natural Language Processing*, Sofija, Bolgarija. https://halshs.archives-ouvertes.fr/hal-00838575/document.

Yves Scherrer in Tomaž Erjavec. 2015. Modernising historical Slovene words. *Natural Language Engineering*, doi: 10.1017/S1351324915000236.

Mirjam Sepesy Maučec, Janez Brest in Zdravko Kačič. 2006. Statistical alignment models in machine translation from Slovenian to English. *Elektrotehniški vestnik*, 73(5): 273–78. http://www.dlib.si/details/URN:NBN:SI:DOC-GDCH7YIE.

Mirjam Sepesy Maučec, Gregor Donaj in Zdravko Kačič. 2013. Improving statistical machine translation with additional language models. V: *Human language technologies as a challenge for computer science and linguistics: proceedings / 6th Language & Technology Conference*, Poznanj, Poljska.

Jorg Tiedemann. 2009. Character-based PSMT for closely related languages. V: *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT)*, str. 12–19, Barcelona, Španija. http://stp.lingfil.uu.se/~joerg/published/eamt09_related.pdf.

Darinka Verdonik. 2013. Strojno prevajanje z Mosesom. *Življenje in tehnika*, 64(7/8): 48–64.

Jernej Vičič in Tomaž Erjavec. 2002. Statistično strojno prevajanje na osnovi vzporednih korpusov. V: *Zbornik enajste mednarodne Elektrotehniške in računalniške konference ERK 2002*, str. 217–220, Portorož, Slovenija.

David Vilar, Jan-T. Peter in Hermann Ney. 2007. Can we translate letters? V: *Proceedings of the Second Workshop on Statistical Machine Translation,* str. 33–39. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1626360.

Sintetizator govora za slovenščino eBralec

Jerneja Žganec Gros,* Boštjan Vesnicer,* Simon Rozman,† Peter Holozan,† Tomaž Šef‡

* Alpineon razvoj in raziskave, d. o. o.
Ulica Iga Grudna 15, 100 Ljubljana
jerneja.gros@alpineon.si
bostjan.vesnicer@alpineon.si
† Amebis, d. o. o.
Bakovnik 3, 1241 Kamnik
simon.rozman@amebis.si
peter.holozan@amebis.si
‡ Institut Jožef Stefan
Jamova 39, 1000 Ljubljana
tomaz.sef@ijs.si

Povzetek

V članku predstavljamo novi sintetizator govora za slovenski jezik, *eBralec*, ki je prvenstveno namenjen slepim in slabovidnim uporabnikom ter osebam z motnjami branja. Poglavitna prednost *eBralca* v primerjavi s predhodnimi sintetizatorji govora za slovenski jezik je v občutno višji stopnji naravnosti rezultirajočega sintetičnega govora. Ženski glas *eBralec Maja* predstavlja prvi ženski slovenski sintetični glas, ki je bil že dolgo na spisku želja slepih in slabovidnih uporabnikov. V članku predstavljamo zgradbo novega sintetizatorja govora, njegove module ter jezikovne vire, ki so bili uporabljeni pri njegovem razvoju.

The eBralec Speech Synthesis System for Slovenian

A new text-to-speech synthesis system for the Slovenian language, *eBralec*, is presented in the paper. *eBralec* has been developed based on a thorough analysis of user requirements provided by the primary end user group representing blind and visually impaired users. *eBralec* outperforms existing solutions for Slovenian speech synthesis in the output speech quality as it yields close-to-natural sounding output speech. *eBralec* also includes a female voice, which represents the first Slovenian synthetic female voice, which has topped the end user wish lists for a considerable time. In the paper we present the structure of the new speech synthesiser and provide a description of its modules and the underlying language resources.

1 Uvod

V članku predstavljamo novi sintetizator govora za slovenski jezik, *eBralec*. *eBralec* je bil razvit v okviru projekta Knjižnica slepih in slabovidnih in je prvenstveno namenjen slepim in slabovidnim uporabnikom ter osebam z motnjami branja.

Poglavitna prednost *eBralca* v primerjavi s predhodnimi sintetizatorji govora za slovenski jezik, kot so denimo *S-5* (Gros et al., 1997), *Govorec* (Šef, 2002), *Proteus TTS* (Žganec Gros in Žganec, 2008) ter *eSpeak*, je v občutno višji stopnji naravnosti rezultirajočega sintetičnega govora.

Na željo končnih uporabnikov je bila velikost pomnilniškega prostora, potrebnega za namestitev ter delovanje sintetizatorja govora, ohranjena na ravni predhodnih Govorčeve oz. Proteusove. To je narekovalo tudi izvedbeno različico končnega sintetizatorja govora, ki temelji na parametrični predstavitvi zakonitosti govora v slovenskem jeziku. Teh zakonitosti se sintetizator govora nauči samodejno na podlagi obsežnega učnega govornega korpusa, ki je bil posebej posnet v te namene, in ki vključuje relevantne akustične ter prozodijske fenomene, ki so značilni za govorjeno slovenščino.

Nova glasova *eBralca* sta moški glas, *eBralec Renato*, ter ženski glas, *eBralec Maja*. Ženski glas *eBralec Maja* predstavlja prvi ženski sintetični glas, ki je na voljo za slovenščino, in je bil že vrsto let na spisku želja slepih in slabovidnih uporabnikov. Za uporabnike, ki so bili vajeni predhodnikov *eBralca*, so razvijalci v *eBralca* vključili tudi glasove iz *Govorca*, razmišljajo pa tudi o vključitvi Proteusovih glasov.

eBralec bo slepim in slabovidnim uporabnikom znatno olajšal delo z računalnikom, dostop do novic in informacij ter tako omogočil njihovo boljšo e-vključenost v sodobno informacijsko družbo.

V članku predstavljamo zgradbo novega sintetizatorja govora, njegove module ter jezikovne vire, ki so bili uporabljeni pri njegovem razvoju.

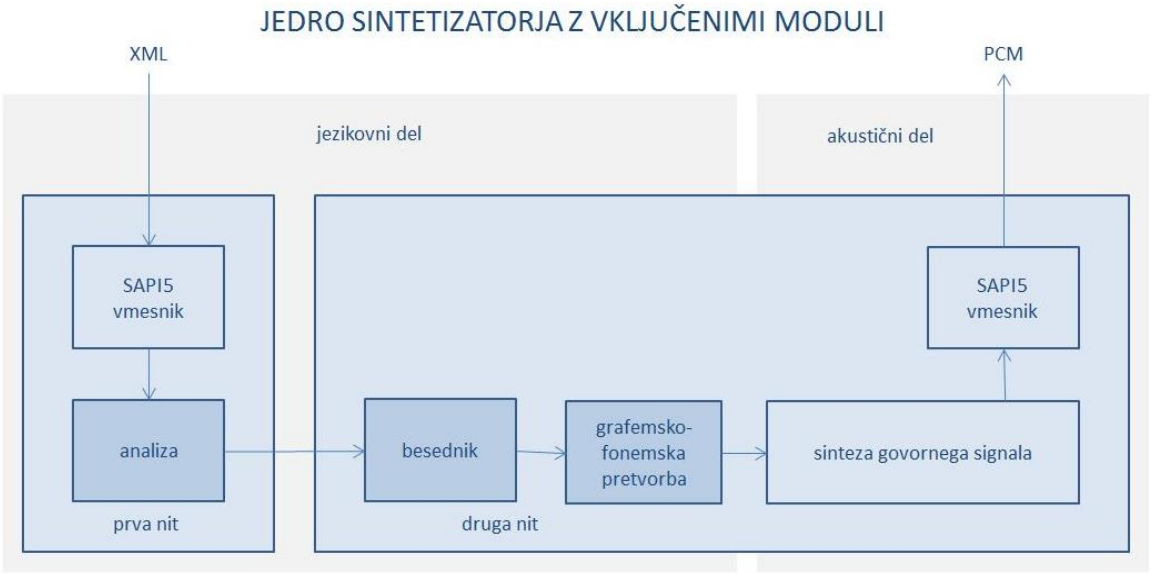
2 Zgradba sintetizatorja govora

Naloga *jedra* sintetizatorja govora oziroma povezovalnega cevovoda je povezovanje sestavnih modulov sintetizatorja govora v enoten proces.

Jedro sintetizatorja govora usklajuje delo posameznih delov sintetizatorja tako, da v ustreznem vrstnem redu vključuje oziroma kliče module sintetizatorja govora. Posamezni moduli oz. faze pretvorbe zaradi pohitritve in večje paralelizacije procesov lahko hkrati delujejo v ločenih nitih (procesorjih ali računalnikih). Zaradi enakomerne porazdelitve procesne obdelave se v prvi niti v trenutni izvedbi izvajajo vsi moduli, potrebni za analizo besedila, v drugi pa ostali moduli, ki so potrebni za nemoteno delovanje sintetizatorje ob izklopu te analize.

Zasnova jedra sintetizatorja govora *eBralec* je prikazana na sliki 1. Moduli, ki jih vključuje jedro *eBralca*, so: jezikovni analizator, besednik, modul za grafemsko-fonemsko pretvorbo in modul za sintezo govornega signala.

Na vhodu in izhodu se jedro sintetizatorja govora lahko poveže z ustreznim vmesnikom, npr. SAPI 5, s pomočjo katerega vhodno besedilo z morebitnimi dodatnimi ukazi spreminja v ustrezen govorni signal.



Slika 1: Shema *jedra* sintetizatorja govora *eBralec*.

Vhodno besedilo sprva obdela *jezikovni analizator*, ki poskrbi za ustrezno predobdelavo vhodnega besedila ter razdvoumljanje izgovornih različic. Rezultat modula za jezikovno analizo je zapis, v katerem so vsebovane vse potrebne informacije o izgovarjavi besed glede na njihovo pozicijo in pomen v vhodnem stavku oziroma povedi.

Modul *besednik* v odvisnosti od vhodnih nastavitev poskrbi za pretvorbo simbolov in števil v besede. Ti elementi so namreč zelo pogost sestavni del besedil, zato je njihovo pravilno izgovarjanje pomembno za razumljivost govora.

Modul »*grafemsko-fonemska pretvorba*« poskrbi za pretvorbo v fonemski zapis (Gros et al., 1997).

Modul za »*sintezo govornega signala*« je zadolžen za oblikovanje prozodije in tvorjenje izhodnega govornega signala.

3 Jezikovna analiza vhodnega besedila

Jezikovna analiza uporablja podatke iz Amebisove *jezikovne baze Ases* (Arhar in Holozan, 2009). Ta za slovenščino v tem trenutku vsebuje več kot 257.000 lem, ki vsebujejo 8,1 milijona oblik, od katerih je 5,7 milijona oblik dodatno opremljenih s podatki o izgovarjavi. Dodatno je za slovenščino v bazi še 36.000 zvez in 8.000 glagolskih predlog. Glagolske predloge podajajo informacije o vezljivosti glagola (Holozan, 2004).

Jezikovni analizator mora narediti razrez besedila na povedi, stavke in besede, potem pa za vsako besedo določiti še ustrezno *lemo* in *oblikoskladenjsko oznako*. *Ases* ločuje leme, ki se različno izgovarjajo, npr. »téma« in »temà« predstavljata dve ločeni lemi.

Že sam razrez vhodnega besedila je lahko težaven. Tako je npr. v primeru »*Videl sem ga. Micka ga je tudi videla*.« treba narediti dve povedi, v primeru »*Videl sem in ga. Micka ga je tudi videla*.« pa le eno. Poleg krajšav so glede tega težavni tudi vrstilni števniki, npr. v primeru »*Bil je na 28. Mednarodnem festivalu*.« Razrez vpliva tako na stavčno intonacijo, kot tudi na branje same kritične besede (*osemindvajset* proti *osemindvajsetem*).

Za branje so predvsem pomembni primeri besed (lahko bi jih imenovali *raznoglasnice*), ki se različno izgovarjajo glede na *pomen v stavku*, in v slovenščini je takih besed zelo veliko.

Take besede so denimo »*je*« (*biti* ali *jesti* ali *osebni zaimek*), »*pol*« (»*Ob pol je pol ljudi šlo na severni pol*.«), »*samo*«, »*celo*«, »*tema*«, »*tako*«, »*mora*«, »*svet*«, »*leti*« (»*Dve leti že leti in leti so vedno daljši*.«), »gori« (»*Gori na gori gori*.«), »hotel« (»Hotel sem hotel.«), tudi »*me*«. Tak primer je npr. tudi »*Vršič*«, kjer je izgovarjava odvisna od tega, ali gre za prelaz ali priimek.

Pri nekaterih glagolih se pri zapisu prekrivata tudi povedni sedanjik in velelnik, ki pa se različno izgovarjata: »*Mati božja prosi za nas*!« proti »*Mati božja, prosi za nas*!«. V slednjem primeru je dodatna težava lahko še to, ali je pisec prav postavil vejico.

Jezikovni analizator, vgrajen v *eBralca*, deluje na podlagi pravil in podatkov iz jezikovne baze *Ases*, pri čemer so osnova glagolske predloge. Analizator je uporabljen tudi v strojnem prevajalniku *Presis* in slovničnem preverjevalniku *Besana*.

Slaba stran zapletenega analizatorja, ki je potreben za razdvoumljanje izgovornih različic besed, je njegova časovna zahtevnost, saj za tekoče branje zahteva razmeroma hiter računalnik. Pokazalo se je, da je analizator v povprečju sicer dovolj hiter, pojavljajo pa se izolirani kritični primeri povedi, v katerih se jezikovna analiza ekstremno podaljša. Najbolj težavne so povedi z velikim številom vejic (npr. dolga naštevanja) in pa daljša zaporedja predložnih zvez, pri katerih se analizator odloča med tem, ali gre za povedkova določila ali za desne prilastke, pri čemer lahko najde zelo veliko število možnih kombinacij.

V *eBralcu* to težavo rešujemo s predčasno prekinitvijo analize, ko se preseže določeno število poskusov, vendar je čas analize v nekaterih primerih še vedno predolg. Slaba stran prekinitve je tudi potencialno nepravilno analizirana poved. Dolgoročna rešitev bo predelava jezikovnega analizatorja, smiselno tudi v smeri izrabe več

jeder procesorja pri analizi, saj trenutno analizator uporablja le eno jedro procesorja, težava pa je, da se v zadnjih letih hitrost računalnikov povečuje bolj z dodajanjem novih jeder kot s samo hitrostjo delovanja jeder (Mattsson, 2014).

4 Tvorjenje govornega signala

Kot smo že omenili v uvodnem poglavju, smo pri razvoju novega sintetizatorja govora upoštevali željo končnih uporabnikov po kompaktni namestitvi. To je narekovalo izvedbeno različico končnega sintetizatorja govora, ki temelji na parametrični predstavitvi zakonitosti govora v slovenskem jeziku. Teh zakonitosti se sintetizator govora nauči samodejno na podlagi obsežne učne govorne zbirke, ki je bila posebej posneta v te namene, in ki vključuje relevantne *akustične* ter *prozodijske* fenomene, ki so značilni za govorjeno slovenščino.

V tem poglavju predstavljamo govorno zbirko, ki smo jo uporabili za učenje parametričnih modelov govora ter postopke modeliranja prozodije in tvorjenja govora s pomočjo prikritih Markovovih modelov.

4.1. Govorna zbirka *eBralca*

Najpomembnejša dejavnika pri snovanju govorne zbirke za potrebe visokokakovostne sinteze govora sta izbira njene vsebine in označevanje posnetkov. Izbira velikosti govorne zbirke je posledica kompromisa med želenim številom variacij glasov oz. njihovim pokritjem na eni strani ter časom in stroški, vezanimi na razvoj na drugi strani. Upoštevati je potrebno tudi čas za kasnejše preiskovanje govorne zbirke in potreben prostor za njeno hranjenje (Amdal in Svendsen, 2005; Hunt in Black, 1996).

Kakovostna sinteza govora zahteva, da ima govorna zbirka pravilno označeno tako identiteto posameznih govornih segmentov kot tudi njihov natančen položaj znotraj zbirke. Običajno samodejnim postopkom za označevanje govorne zbirke sledi »ročno« popravljanje oznak, ki je ne glede na hiter razvoj tehnologije še vedno časovno potratno.

Postopek *zasnove govorne zbirke eBralca* je obsegal naslednje korake (Šef in Romih, 2011):

- ustvari se obsežna tekstovna zbirka besedil, ki pokriva različne zvrsti (dnevni časopis, revije, leposlovje ipd.),
- iz zbirke besedil se odstranijo vse oznake, vezane na oblikovno podobo (glava besedila, tabele ipd.),
- okrajšave, števila ipd. se pretvorijo v polno besedno obliko (normalizacija besedil),
- besedila se pretvorijo v predvideni fonetični prepis (grafemsko-fonemska pretvorba),
- da bi dosegli statistično ustrezno vzorčenje izbranega področja govorjenega jezika, se optimizira obseg zbirke glede na vnaprej pripravljene kriterije z uporabo postopkov požrešnega iskanja,
- izbrane povedi se posnamejo,
- posneto govorno gradivo se fonetično in prozodično označi (samodejno grobo označevanje, fino ročno popravljanje oznak).

4.1.1. Postopek za izbiro povedi

V nadaljevanju bolj podrobno opisujemo postopek za izbiro povedi za snemanje, ki smo mu posvetili veliko pozornosti.

1. *Statistična obdelava besedil*

Statistično obdelamo celoten besedilni korpus in določimo *pogostost pojavljanja posameznih glasov* in *glasovnih nizov* v besedilu. Pri tem dodatno razlikujemo med naglašenimi in nenaglašenimi glasovi ter glasovi, ki se pojavljajo na koncu stavka oz. na mestih zajema zraka ob ločilih (Mihelič et al., 2006). Presledke na drugih mestih lahko prezremo oz. odstranimo, ker je končno besedilo ob branju izgovorjeno povezano.

Vključimo vse stavke (povedne, velelne, vprašalne itd.) in izdelamo statistiko posameznih vrst povedi oz. stavkov.

2. *Izdelava spiska glasovnih nizov* z oceno zaželenosti posameznega niza

V spisek vključimo nabor vseh teoretično možnih kombinacij difonov. Zaradi robustnosti sintetizatorja govora vključimo tudi difone, na katere pri statistični obdelavi nismo naleteli.

V spisek želenih glasovnih nizov vključimo vse trifone, štirifone ter ostale zaželene najpogostejše polifone, na katere smo naleteli pri statistični obdelavi besedil.

Utež oz. ocena zaželenosti niza je odvisna od pogostosti njegovega pojavljanja v besedilu.

3. *Postopek izbire povedi*

Ocenimo doprinos glasovnih nizov za vsako poved iz besedilnega korpusa.

Doprinos povedi je enak vsoti vseh ocen zaželenosti nizov glasov iz spiska želenih glasovnih nizov, ki se v povedi pojavijo.

Doprinos posamezne povedi normiramo z dolžino povedi, izraženo s številom besed v povedi oziroma številom fonemov v povedi.

Določimo takšno utež, da bodo dolžine izbranih stavkov čim bolj ustrezale statistični porazdelitvi dolžin stavkov iz korpusa.

Izberemo poved z najvišjim normiranim doprinosom.

Iz spiska želenih glasovnih nizov odstranimo vse glasovne nize, ki jih izbrana poved vsebuje.

Ponovno ocenimo vsako poved in izberemo najboljšo, glede na novi popravljeni spisek želenih glasovnih nizov.

Postopek ponavljamo, dokler ne dosežemo vnaprej izbranega želenega števila povedi za posamezni sklop obdelave.

4. *Vmesno vrednotenje rezultatov*

Ko obdelamo sklop vnaprej izbranega števila povedi, izdelamo statistiko difonov, trifonov, štirifonov in drugih polifonov, ki jih že pokrivamo: gre za glasovne nize, ki smo jih do takrat že izločili iz zgoraj omenjenega spiska.

5. *Dodatne izboljšave algoritma*

Ker mora zbirka vsebovati vse možne kombinacije difonov, algoritem popravimo tako, da difonom priredimo dodatno težo glede na ostale polifone. Na takšen način bo algoritem na začetku dajal prednost povedim, ki bodo pokrile čim več novih difonov. Predvidoma se vsi difoni pokrijejo že po okoli 100 začetnih dodanih povedih.

Pri trifonih in štirifonih upoštevamo pri robnih glasovih tudi podatek o glasovni skupini, ki ji pripadajo Na primer, štirifon "krak" ne bo doprinesel prav dosti novega v našo zbirko, če ta že vsebuje štirifon "krat", zato oceno koristnosti takega štirifona popravimo navzdol. To lahko naredimo preprosto tako, da v spisek vnesemo

dodatne nize, skupaj z njihovimi frekvencami pojavljanja v korpusu: primer takega štirifona: "k"+"r"+"a"+ "pripornik".

Algoritem za izbiro povedi z različnim uteževanjem izboljšamo tako, da končni nabor vsebuje različne vrste naklonov in raznovrstne povedi: povedne, vprašalne, velelne, enostavne, sestavljene, naštevanje, in podobno. Tako lahko isti korpus učinkovito uporabimo tudi za generiranje prozodijskih parametrov pri sintezi govora.

4.1.2. Snemanje govorne zbirke

Snemanje govorne zbirke je potekalo v studiu RTV Slovenija ob prisotnosti izkušenega tonskega tehnika. Med desetimi profesionalnimi govorci smo izbrali najustreznejši moški in ženski glas. Med branjem besedila so govorci imeli nameščene elektrode laringografa, s katerimi smo spremljali nihanje glasilk za lažje kasnejše označevanje osnovnih period govornega signala.

Samo snemanje je zaradi obsežnosti besedila, ki ga je bilo treba prebrati, trajalo več mesecev. Pri tem so nastavitve opreme ves čas ostale nespremenjene. Pred vsakim snemanjem je govorec poslušal svoje predhodne posnetke, s čimer se je skušalo zagotoviti čim bolj enoten način govora med posameznimi snemalnimi sejami.

Uporabljena govorna zbirka pokriva skoraj vse možne kombinacije difonov in trifonov, ki smo jih identificirali pri analizi dela besedilnega korpusa FidaPLUS[1], ki je obsegal več kot 7 milijonov povedi (približno 30 % povedi v korpusu). Za vsak glas je bilo prebranih 4.000 povedi povprečne dolžine 11 besed.

Sledil je ročni pregled posnetega gradiva, grobo samodejno označevanje mej med glasovi ter časovno zahtevno ročno popravljanje napak.

V primerjavi z obstoječimi govornimi zbirkami, namenjenimi sintezi slovenskega govora (Rojc in Kačič, 2000; Šef, 2002; Mihelič et al., 2006), predstavlja govorna zbirka *eBralec* najbolj obsežno izdelano govorno zbirko za slovensko sintezo govora. V tabeli 1 povzemamo pregledne podatke o novi govorni zbirki.

Velikost besednega korpusa	7.145.345 povedi 77 milijonov besed
Obseg govorne zbirke	*4.000 povedi* 46.785 besed 6 ur 3 min posnetkov za ženski glas 5 ur 33 min posnetkov za moški glas
Število različnih difonov v zbirki	1.883
Število različnih trifonov v zbirki (št. kombinacij v korpusu)	21.369 (24.702)

Tabela 1: Podatki o govorni zbirki *eBralca*. Vsak govorec je posnel 4.000 povedi.

4.2. Modeliranje prozodije in tvorjenje govora z uporabo prikritih Markovovih modelov

Prikriti Markovovi modeli (PMM) so bili vse od prvih poskusov uporabe, ki segajo nekje v sedemdeseta leta prejšnjega stoletja (Jelinek, Baker), pa vse do nedavnega, nepogrešljiva tehnologija na področju samodejne prepoznave govora.

Za razliko od *prepoznave* govora, so se prvi obetavni poskusi *sinteze* govora z uporabo PMM-jev pričeli pojavljati šele v zadnjih letih prejšnjega stoletja. Pionirsko delo na tem področju je opravil Tokuda s sodelavci (Tokuda, 1995). Razlog za razmeroma pozen začetek uporabe lahko pripišemo temu, da se zdijo PMM-ji zaradi svoje statistične narave na prvi pogled neprimerni za nalogo, kot je tvorjenje oziroma sinteza govora. Bistvena razlika med tvorjenjem in prepoznavo govora je namreč ta, da želimo pri prepoznavi iz govora izluščiti le bistvene značilnosti govora, medtem ko želimo pri tvorjenju govora doseči ravno nasprotno, v govoru želimo ohraniti čim več značilnosti, ki so prisotne v naravnem govoru.

Sinteza govora z uporabo PMM-jev ima v primerjavi z bolj klasičnimi postopki tvorbe govora, pri katerih govor tvorimo z »lepljenjem« krajših ali daljših govornih izsekov, nekaj privlačnih prednosti:

- za zadovoljivo kakovost govora potrebujemo razmeroma majhno govorno zbirko (zadošča že ura ali več posnetega govora),
- govorne zbirke ni treba zelo natančno označiti,
- omogoča enovito in sočasno modeliranje akustičnih in prozodičnih lastnosti govora,
- omogoča zgoščen zapis modela govora; za tvorbo govora ni treba hraniti celotne izvorne govorne zbirke,
- omogoča visoko naravnost prozodije tvorjenega govora.

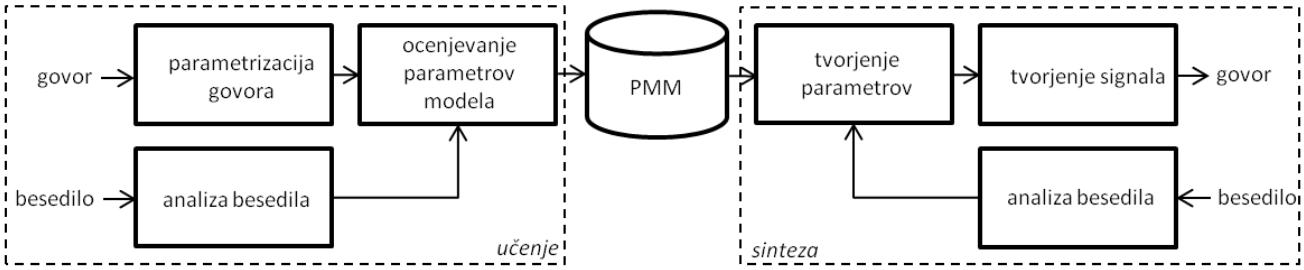
Po drugi strani pa imajo sistemi PMM tudi nekatere slabosti. Govor je lahko na trenutke nekoliko manj razumljiv. Govor ima lahko ponekod značilen »robotski« prizvok, ki je posledica parametrizacije govornega signala.

Od prvih poskusov pa do danes je bilo predlaganih veliko izboljšav in nadgradenj, ki so pripomogle, da je tvorba govora z uporabo PMM-jev postala dobra alternativa bolj uveljavljenim postopkom tvorbe govora, kot je denimo korpusna sinteza govora. Med najpomembnejše tovrstne izboljšave štejemo naslednje:

- postopek generiranja z uporabo dinamičnih parametrov (Tokuda, 10995),
- uporabo kontekstno-odvisnih osnovnih glasovnih enot (Yoshimura et al., 1999),
- združevanje parametrov modela z uporabo fonetičnih pravil (Yoshimura et al., 1999),
- vpeljavo kriterija globalne variance (Toda in Tokuda, 2007),
- uporabo naprednejših metod za parametrizacijo govora, kot je denimo STRAIGHT (Kawahara e tal., 1999),
- uporabo alternativnih statičnih akustičnih modelov (Shannon et al., 2013; Zen in Sak, 2015).

Postopek tvorbe govora z uporabo PMM-jev je sestavljen iz postopka učenja in postopka sinteze. V postopku *učenja* iz večjega števila označenih govornih posnetkov ocenimo parametre modela, v postopku *sinteze* pa naučeni model uporabimo za generiranje govora. Ker surovi zvočni posnetki niso neposredno primerni za gradnjo modela, govor predhodno pretvorimo v zgoščen parametričen zapis. Na sliki 2 je prikazana osnovna shema tvorbe govora z uporabo PMM-jev.

[1] https://sl.wikipedia.org/wiki/FidaPLUS



Slika 2. Shema sistema za tvorjenje govora z uporabo PMM.

V postopku *učenja* želimo izbrati tisto vrednost parametrov statističnega modela, pri kateri bo funkcija verjetja dosegla največjo vrednost. V ta namen po navadi uporabimo učinkovit in teoretično dobro raziskan optimizacijski postopek maksimizacije upanja EM (angl. expectation maximization).

Podoben postopek uporabimo tudi pri *sintezi*, le da v tem primeru na podlagi podanega vhodnega besedila iščemo optimalen niz parametrov govornega signala, medtem ko vrednosti parametrov modela ne spreminjamo. V zadnjem koraku niz parametrov govornega signala pretvorimo v govorni signal.

4.2.1. Tvorjenje govora z uporabo PMM v eBralcu

V nadaljevanju bomo bolj natančno predstavili posamezne korake in različne nastavitve, ki jih uporabljamo za sintezo govora v *eBralcu*. Pri večini korakov smo uporabljali posamezna programska orodja iz sklopa orodij HTS[2], nekatera pomožna orodja pa smo razvili tudi sami.

Govorne zvočne datoteke, katerih frekvenca vzorčenja je znašala 48 kHz, smo najprej pretvorili v nize koeficientov melodičnega kepstra po sledečem postopku. Celoten zvočni signal, ki je ustrezal eni zaključeni stavčni povedi, smo razdelili na 25 ms trajajoče govorne izseke, pri čemer sta se dva sosednja izseka prekrivala v 80-ih odstotkih. Iz vsakega izseka smo izračunali 35 koeficientov melodičnega kepstra, ki smo jim pripeli še logaritem osnovne frekvence, ki smo jo izračunali s postopkom RAPT (Talkin, 1995). Tem 36-razsežnim vektorjem značilk smo dodali še dinamične značilke prvega in drugega reda, tako da smo dobili 108-razsežne vektorje značilk.

Modeli PMM osnovnih govornih enot so imeli pet stanj, topologija pa je bila levo-desna. Trajanje posameznih govornih enot smo modelirali na ekspliciten način – trajanje vsakega stanja modela govorne enote smo opisali z normalno porazdelitvijo. Na podoben način smo modelirali tudi osnovno frekvenco. Posebej smo ocenili tudi globalno varianco značilk na nivoju povedi, za katero se je izkazalo, da lahko pripomore k večji naravnosti govora (Toda in Tokuda, 2007).

Učenje oziroma postopek ocenjevanja parametrov modelov smo izvedli z dobro znanim postopkom Bauma in Welcha, ki predstavlja poseben primer postopka EM.

Pri učenju potrebujemo poleg zvočnih datotek, ki vsebujejo govorne signale (ena datoteka za posamezno stavčno poved), tudi datoteke s fonetičnimi oznakami.

Načeloma je dovolj, če je za vsako poved določen fonetični prepis, še bolje pa je, če so alofoni opremljeni tudi s časovnimi oznakami, ki povedo, kdaj se posamezen alofon v povedi začne in kako dolgo traja. Za dobro kakovost tvorjenega govora je pomembno, da so te oznake čim bolj natančne.

Ker je znano, da so akustične lastnosti posameznega alofona zelo odvisne od okolice, v kateri se alofon nahaja, alofonom pripišemo tudi *kontekst*. Kontekst lahko definiramo poljubno, pomembno pa je, da podaja tiste glasoslovne in jezikoslovne faktorje, ki najbolj vplivajo na akustične lastnosti konkretnega fonema oz. alofona. V našem primeru je kontekst med drugim vseboval naslednje kontekstne faktorje:

- predhodni in sledeči fonem,
- mesto fonema v zlogu,
- mesto zloga v besedi oziroma stavku,
- mesto besede v stavku,
- se fonem nahaja v poudarjenem/nepoudarjenem zlogu,
- razdalja do poudarjenega zloga,
- dolžina prejšnjega/trenutnega/naslednjega stavka,
- stavčni naklon ter
- število zlogov/besed/stavkov v povedi.

Tako definiran kontekst določa osnovno govorno enoto. Idealno bi bilo, če bi za vsako tako govorno enoto lahko naučili lasten PMM, vendar to zaradi kombinatorične eksplozije ni možno. V še tako veliki govorni zbirki bi namreč »videli« le majhen delež vseh takšnih govornih enot.

To težavo rešimo tako, da s pomočjo fonetičnih pravil določimo roje oziroma skupine govornih enot, ki si delijo skupne parametre. Na ta način je mogoče parametre PMM-jev oceniti dovolj robustno. Dodatna prednost deljenja parametrov je tudi ta, da je končni model, ki ga potrebujemo pri sintezi govora, zelo kompakten, četudi je mogoče izvorna govorna zbirka, ki smo jo uporabili za učenje tega modela, zelo obsežna.

5 Zaključek

V prispevku smo predstavili zasnovo in izvedbo novega visokokakovostnega sintetizatorja govora za slovenski jezik. *eBralec* bo slepim in slabovidnim uporabnikom znatno olajšal delo z računalnikom, dostop do novic in informacij ter tako omogočil njihovo boljšo e-vključenost v sodobno informacijsko družbo.

Pri nadaljnjem razvoju *eBralca* imamo še veliko načrtov. Zaradi velike časovne zahtevnosti postopkov jezikovne analize načrtujemo predelavo jezikovnega analizatorja z več stopnjami predelave, od izrazito

[2]HMM-based Speech Synthesis System (HTS), http://hts.sp.nitech.ac.jp.

površinske pa vse do poglobljene jezikovne analize. Razmišljamo v smeri izrabe več jeder procesorja pri jezikovni analizi. Prav tako načrtujemo izboljšave pri sintezi krajših govornih segmentov s kombinacijo korpusne sinteze in PMM.

6 Zahvala

Razvoj *eBralca* je bil delno financiran v okviru projekta Knjižnica slepih in slabovidnih Minke Skaberne. Operacijo je delno financirala Evropska unija iz Evropskega socialnega sklada. Operacija se je izvajala v okviru Operativnega programa razvoja človeških virov, razvojne prioritete "Enake možnosti in spodbujanje socialne vključenosti", prednostne usmeritve "Dvig zaposlenosti ranljivih družbenih skupin na področju kulture in podpora njihovi socialni vključenosti".

7 Literatura

I. Amdal in T. Svendsen. 2005. Unit selection Synthesis Database Development Using Utterance Verification, V: *Zbornik INTERSPEECH 2005*, str. 2553-2556.

Špela Arhar in Peter Holozan. 2009. ASES – leksikalna podatkovna zbirka za razvoj slovenskih jezikovnih tehnologij. V Mikolič (ur.). *Jezikovni korpusi v medkulturni komunikaciji*. Koper: Založba Annales.

Jerneja Gros, Nikola Pavešić in France Mihelič. 1997. Text-to-Speech synthesis: a complete system for the Slovenian language. *CIT*, let. 5, št. 1, str. 11-19.

Peter Holozan. 2004. Uporaba glagolskih predlog pri strojnem prevajanju. V: Zborniku Konference JEZIKOVNE TEHNOLOGIJE 2004, str. 128. Ljubljana.

A. Hunt in A. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. V: *Proceedings of ICASSP 96*, zvezek 1, str. 373-376.

H. Kawahara, I. Masuda-Katsuse in A. de Cheveigné, 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction. *Speech Communication*, zvezek 27, str. 187–207.

P. P. Mattsson. 2014. *Why Haven't CPU Clock Speeds Increased in the Last Few Years?* https://www.comsol.com/blogs/havent-cpu-clock-speeds-increased-last-years/

Aleš Mihelič, Jerneja Žganec Gros, Nikola Pavešić in Mario Žganec. 2006. Efficient subset selection from phonetically transcribed text corpora for concatenation-based embedded text-to-speech synthesis. *Informacije MIDEM*, letn. 36, št. 1, str. 19-24.

Matej Rojc in Zdravko Kačič. 1999. Design of optimal Slovenian speech corpus for use in the concatenative speech synthesis system. V: *Proceedings of the Second international conference on language resources an evaluation*. str. 321-325. Athens. Greece.

M. Shannon, H. Zen in W. Byrne. 2013. Autoregressive models for statistical parametric speech synthesis. *IEEE Trans. Acoust. Speech Lang. Process.*, zvezek 21, št. 3, str. 587–597.

Tomaž Šef. 2002. Sistem GOVOREC za sintezo slovenskega govora. *Elektrotehniški vestnik*, str. 165-170.

Tomaž Šef in Miro Romih. 2011. Zasnova govorne zbirke za sintetizator slovenskega govora Amebis Govorec, V: *Zbornik 14. mednarodne multikonference Informacijska družba*, zvezek A, str. 88-91.

David Talkin. 1995. A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, zvezek 495, str. 518.

T. Toda in K. Tokuda. 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions Inf. Syst.*, zvezek E90-D, št. 5, str. 816–824.

K. Tokuda, T. Kobayashi in S. Imai. 1995. Speech parameter generation from HMM using dynamic features. V: *Proceedings of the ICASSP*, str. 660–663.

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi in T. Kitamura. 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. V: *Proceedings of the Eurospeech*, str. 2347-2350, september 1999.

H. Zen in H. Sak. 2015. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. V: *Proceedings of the ICASSP*, str. 4470–4474.

Jerneja Žganec Gros in Mario Žganec. 2008. An efficient unit-selection method for concatenative text-to-speech synthesis systems. *CIT*, zvezek. 16, št. 1, str. 69-78.

Razpoznavanje tekočega govora v slovenščini z bazo predavanj SI TEDx-UM

Andrej Žgank, Darinka Verdonik, Mirjam Sepesy Maučec

Inštitut za elektroniko in telekomunikacije
Fakulteta za elektrotehniko, računalništvo in informatiko
Univerza v Mariboru
Smetanova ul. 17, 2000 Maribor
andrej.zgank@um.si, darinka.verdonik@um.si, mirjam.sepesy@um.si

Povzetek

V članku bomo predstavili novi slovenski govorni vir, nastal na osnovi posnetkov predavanj TEDx. Govorna baza vsebuje posnetke 242 predavanj, v skupni dolžini 54 ur. Transkribiranje govora v bazi SI TEDx-UM smo izvedli v dveh delih. Učni nabor smo transkribirali avtomatsko, z uporabo razpoznavalnika govora UMB Broadcast News. Razvojni in testni nabor baze, ki obsega 3 ure govornega materiala, pa smo transkribirali ročno, v skladu z nadgrajenimi priporočili za transkribiranje korpusa GOS. Razpoznavalnik govora UMB Broadcast News ASR je prvenstveno namenjen televizijskim oddajam, zato smo v nadaljevanju izvedli analizo, kakšen vpliv ima tip jezikovnega modela na uspešnost razpoznavanja govora v bazi SI TEDx-UM. Primerjali smo privzeti jezikovni model domene televizijskih oddaj s splošnim jezikovnim modelom grajenim na korpusu FidaPLUS. Povprečna napaka razpoznavanja besed na testnem naboru baze SI TEDx-UM je znašala 50,7%. Govorna baza SI TEDx-UM je prosto dostopna.

Slovenian Continuous Speech Recognition with the SI TEDx-UM Talks Database

This paper presents a new Slovenian spoken language resource for automatic speech recognition. The SI TEDx-UM speech database contains 242 different Slovenian TEDx talks in total length of 54 hours. The training set was transcribed automatically using the UMB Broadcast News speech recognition system. The development and evaluation set were manually transcribed, using adapted transcription guidelines for the Slovenian GOS corpus. The influence of language model domain was also analysed. The UMB Broadcast News ASR language model was compared with the general language model build on FidaPLUS corpus. The average word error rate of 50.7% was achieved on the SI TEDx-UM evaluation set. The SI TEDx-UM speech database is freely available.

1. Uvod

Avtomatsko razpoznavanje tekočega govora predstavlja eno izmed pomembnih IKT tehnologij, tako na področju vmesnikov človek-stroj kot tudi na področju digitalnih vsebin. Porast le-teh v zadnjih nekaj letih ima za posledico nujnost avtomatske obdelave velike množice takšnega medijskega materiala. Razpoložljivost ustreznih govornih virov predstavlja še vedno eno izmed kritičnih točk na področju razvoja govornih tehnologij, predvsem avtomatskega razpoznavanja govora. Klasičen pristop izdelave govornih virov je zaradi ročnega transkribiranja dolgotrajen in drag. Posledično vlada za velik delež svetovnih jezikov, med katerimi je tudi slovenščina, pomanjkanje ustreznih govornih virov.

V članku[1] bomo predstavili nov slovenski jezikovni vir, govorno bazo SI TEDx-UM, ki je nastala na osnovi sklopa slovenskih predavanj TEDx. Ta predavanja so v svetu uveljavljena že vrsto let, v zadnjih petih letih so se širše uveljavila tudi v Sloveniji. Predavanja pokrivajo različne aktualne tematike s področja tehnologije, izobraževanja, umetnosti in družbe.

Govorna baza SI TEDx-UM je razdeljena na dva dela. Prvi del predstavlja učni nabor, za katerega smo segmentacijo in transkripcije pripravili avtomatsko, z uporabo razpoznavalnika slovenskega govora. Na takšen način smo uspeli bistveno poenostaviti in pohitriti izdelavo govorne baze, vendar bo potrebno zaradi prisotnih napak v transkripciji v prihodnje uporabljati drugačne pristope razvoja razpoznavalnika govora (Rousseau et al., 2014). Drugi del govorne baze predstavlja razvojni in testni nabor, kjer smo transkripcije predavanj tvorili ročno. Za izdelavo izhodiščne verzije avtomatskih transkripcij smo zaradi njegovega robustnega delovanja uporabili obstoječi avtomatski razpoznavalnik govora UMB Broadcast News (Žgank et al., 2014a), kjer so modeli specifično uravnotežen na domeno televizijskih informativnih oddaj. V prispevku bomo analizirali, kako razlika med splošnim jezikovnim modelom in takšnim, uravnoteženim na televizijske informativne oddaje, vpliva na uspešnost razpoznavanja govora predavanj TEDx.

V nadaljevanju članka bomo najprej predstavili postopek zajema in ročnega transkribiranja govorne baze SI TEDx-UM. V tretjem poglavju bomo predstavili postopek avtomatske segmentacije in transkribiranja zajetega govornega materiala. V četrtem poglavju bo sledila predstavitev dveh tipov jezikovnih modelov, uporabljenih za avtomatsko transkribiranje. Rezultate in analizo vrednotenja razpoznavanja govora bomo predstavili v petem poglavju. Zaključek in smernice za nadaljnje delo bomo podali v šestem poglavju.

2. Slovenska govorna baza SI TEDx-UM

2.1. Zajem materiala

Izvorni material za govorno bazo smo zajeli s spletne strani YouTube, kjer so na voljo posnetki različnih predavanj TEDx v slovenskem jeziku. Ker so predavanja prirejali različni organizatorji, smo na takšen način najlažje zajeli vsa dostopna slovenska predavanja, ki jih je

[1] Raziskovalno delo je bilo delno sofinancirano s strani ARRS po pogodbi št. P2-0069.

bilo več kot 300. Posnetki predavanj so na spletni strani YouTube na voljo v različnih izgubnih kodekih (tipično: zvok: MPEG AAC, video: H.264). Vedno smo uporabili tistega, ki zagotavlja najvišjo možno kakovost govornega materiala. Izvorne govorne posnetke smo pretvorili v format WAV s frekvenco vzorčenja 16 kHz in 16-bitno ločljivostjo, kar je standardni format govorne baze BNSI Broadcast News in kot takšen zadostuje za razvoj razpoznavalnikov govora. Kakovost video materiala je bila pri zajemu drugotnega pomena, saj je video služil zgolj kot pomoč pri transkribiranju govora.

2.2. Transkribiranje

Transkribiranje spontanega govora v jezikovnih virih se je v slovenščini v preteklih letih razvijalo predvsem ob posameznih govornih bazah, kot sta bili BNSI Broadcast News (Žgank et al., 2004) in SiBN Broadcast News (Žibert in Mihelič, 2004), pomembno prelomnico pa predstavlja standard transkribiranja, razvit ob referenčnem govornem korpusu GOS (Verdonik et al., 2013). Novosti v načelih transkribiranja, ki jih je prinesel GOS v primerjavi z bazo BNSI Broadcast News, so v preteklosti že bile podrobno analizirane (Žgank et al., 2014). Pri snovanju specifikacij transkribiranja za slovensko bazo TEDx smo prepoznane razlike v glavnem upoštevali ter pripravili posodobljena načela transkribiranja, ki upoštevajo standarde, vzpostavljene s korpusom GOS in se v primerjavi z bazo BNSI Broadcast News najbolj razlikujejo v načinu segmentiranja izjav na osnovne enote – segmente oz. izjave – in v načinu zapisovanja govora, v primerjavi s korpusom GOS pa vključujejo bolj natančno označevanje akustičnega ozadja in akustičnih dogodkov.

Tako so v SI TEDx-UM osnovne enote segmentirane po načelu, da segmenti približno ustrezajo pojmu izjave, pri čemer izjavo razumemo kot osnovno enoto govora, ki približno ustreza pojmu (kratke) povedi v pisni rabi. Iz tehničnih razlogov smo pri tem upoštevali omejitev, da morajo vsak segment vedno zamejevati vsaj tolikšni premori v govoru, da je mogoče postaviti časovno mejo med segmentoma na način, da ni v zvočnem signalu odrezan noben delček fonema predhodne ali naslednje besede. Prednost imajo vedno krajši segmenti, saj to omogoča uspešnejše učenje akustičnih modelov.

Zapis govora namesto enojnega standardiziranega zapisa vključuje dvotirni sistem zapisovanja, pogovorni in standardizirani, uveden s korpusom GOS. Posebne odločitve so bile potrebne le v zvezi s predhodno ugotovljenimi pomanjkljivostmi zapisovanja v korpusu GOS (Verdonik 2014). Pri ugotovljenih nedoslednostih zapisovanja zvočnika dvoustnični v (ni nosilec zloga), kjer v korpusu GOS najdemo po večkrat tudi pogovorne zapise tipa *mau* (*malo*), *biu* (*bil*), *šou* (*šel*), *dou* (*dol*), *prou* (*prav*), *dau* (*da bo*), *nou* (*ne bo*) itd., ohranimo pravilo, da ga zapisujemo s črko »v« (*prov, nav, navm, odpravt, davn...*) oz. tudi z »l«, če tako izhaja iz knjižne norme (*kosil, mel*), in smo še posebej pozorni na ugotovljene nedoslednosti. Če je u samoglasniški, tj. je nosilec zloga (tudi če gre za predlog *v,* izgovorjen samoglasniško), pa ga enako kot prej pišemo s črko »u« (*pršu, vidu, u tem delu...*). Pri zapisovanju neverbalnih in polverbalnih glasov se držimo seznama, predstavljenega v (Verdonik 2014). Določni člen 'ta' v tipu 'ta rdeči' po novem pišemo skupaj v pogovornem zapisu, narazen pa v standardiziranem zapisu.

Pri označevanju akustičnih dogodkov in ozadja je novost dodatno označevanje akustičnega ozadja, ki traja 3 sek. ali več in se spremeni v primerjavi s tem, kakšno ozadje prevladuje v posnetku večino časa. Za to se uporablja posebna časovna sled. Za seznam akustičnih dogodkov smo uporabljali oznake, podane znotraj orodja Transcriber AG (vključno s funkcijo named entities), je pa bolj natančno in označeno vedno, ko je kak akustični dogodek slišen, ne samo takrat, ko je pragmatično pomemben za komunikacijo.

Transkribiranje je potekalo s pomočjo prenovljene različice programa Transcriber, Transcriber AG. Čeprav orodje ponuja med drugim izredno dobrodošlo novost, da omogoča transkribiranje ob videoposnetku, se je med delom žal pokazalo kot izredno nestabilno ter z nekaj napakami v delovanju zlasti pri označevanju akustičnih dogodkov.

2.3. Podatki o bazi

V končno verzijo govorne baze smo vključili ročno izbranih 242 predavanj, katerih značilnosti so ustrezale kriterijem na področju razpoznavanja govora. Glavna značilnost izločenih predavanj je bila popačena akustična karakteristika kot posledica različnih vzrokov: prekrivanje govora več govorcev, glasbena spremljava oziroma glasbeno ozadje, nizka kakovost posnetkov ipd. Prav tako smo izločili vsa predavanja v tujem jeziku. Posledično vsebuje tipično predavanje govor enega slovenskega govorca v dobrih akustičnih pogojih. V bazi je 66% moških govorcev in 34% ženskih govork. Skupna dolžina vključenih predavanj je 54 ur, izvirajo pa iz časovnega obdobja 6 let. Izmed 242 predavanj smo izbrali 13 predavanj v skupni dolžini 3 ur, ki predstavljajo razvojni in testni nabor govorne baze. Za ta del govorne baze smo v skladu z zastavljenimi pravili izvedli ročno transkribiranje, saj lahko samo na takšen način govorno bazo uporabljamo za eksperimente razpoznavanja tekočega govora. Transkripcije za učni nabor baze v dolžini 51 ur smo pripravili avtomatsko s pomočjo razpoznavalnika slovenskega govora in na takšen način bistveno pohitrili izdelavo nove govorne baze. Transkripcije govorne baze SI TEDx-UM vsebujejo 372k pojavnic, od tega jih je 32k različnih.

3. Avtomatsko transkribiranje

Za pripravo avtomatsko tvorjenih transkripcij predavanj TEDx smo uporabili razpoznavalnik tekočega slovenskega govora UMB Broadcast News (Žgank et al., 2014a). Takšen pristop bistveno poenostavi izdelavo transkripcij, kadar lahko v njih toleriramo večji ali manjši delež napak. Prvi korak avtomatske obdelave govorne baze SI TEDx-UM je predstavljala akustična segmentacija s pomočjo modelov GMM, s katerimi smo tvorili akustično homogene dele govornih posnetkov, primerne za avtomatsko razpoznavanje govora.

Uporabljeni razpoznavalnik govora je bil naučen na kombinaciji govorne baze BNSI Broadcast News in interne govorne baze IETK-TV. Baza BNSI Broadcast News obsega 36 ur obdelanih podatkov, od tega je 30 ur namenjenih učenju akustičnih modelov, 3 ure razvojnemu prilagajanju sistema in 3 ure testiranju razpoznavalnika. Baza je bila zajeta v letih 2003 in 2004 v sodelovanju z RTV Slovenija ter vključuje informativne oddaje tega TV-programa, predvsem TV Dnevnik in Odmeve. Posnetki so bili ročno segmentirani in transkribirani s pomočjo orodja Transcriber, nato pa vključeni v razvoj razpoznavalnika govora. Z uporabo govorne baze IETK-TV smo pridobili dodatnih 29 ur transkribiranega materiala, tako da je končni učni nabor razpoznavalnika govora za transkribiranje baze SI TEDx-UM zajemal 59 ur posnetkov.

Razpoznavalnik govora UMB Broadcast News uporablja za izločanje značilk 12 mel kepstralnih koeficientov z energijo ter prvimi in drugimi odvodi, tako da ima končni vektor 39 elementov. Dodali smo še normalizacijo srednjih vrednosti kepstralnih koeficientov, saj na tak način zmanjšamo razlike med različnimi snemalnimi okolji, kar je izredno pomembno v primeru kombiniranja več govornih baz. Akustični modeli razpoznavalnika govora so bili naučeni z iterativnim postopkom in so v končni obliki vsebovali kontekstno odvisne grafemske medbesedne modele s kombinacijo 32 zveznih Gaussovih porazdelitvenih funkcij verjetnosti na stanje. Več podrobnosti o uporabljenem sistemu UMB Broadcast News je podanih v (Žgank et al., 2014a). V drugem koraku avtomatske obdelave govorne baze SI TEDx-UM smo izvedli razpoznavanje govora na akustično homogenih posnetkih, ki so bili rezultat prvega koraka.

4. Jezikovna modela

Za gradnjo jezikovnih modelov smo uporabili orodje SRI Language Modeling Toolkit (Stolcke, 2002). Izhodiščni jezikovni model, ki smo ga zgradili v okviru razpoznavalnika UMB Broadcast News, je interpolirani 3-gramski model, sestavljen iz štirih komponent. Pri učenju vseh štirih komponent smo uporabili Good-Turingovo glajenje in sestopanje po Katzu. Glede na velikost učnega korpusa in glede na utež v končnem modelu je največja komponenta FidaPLUS. Le-ta je grajena na korpusu FidaPLUS, ki predstavlja referenčno zbirko vsakdanje javne rabe slovenščine v pisnih besedil v obdobju med 1990 do 2006, in vsebuje 621 milijonov besed (Arhar in Gorjanc, 2007). Ostale tri komponente (BNSI-Speech, BNSI-Text in Večer) imajo približno enako utež in torej v enakem deležu prispevajo h končni oceni verjetnosti jezikovnega modela. Tudi komponenta Večer je predstavnik pisanega jezika, ostali dve pa govorjenega. Interpolacijske uteži so bile določene na razvojni množici BNSI-Devel, ki je po strukturi enaka BNSI-Speech in torej predstavlja reprezentativni vzorec ciljne domene za razpoznavalnik UMB Broadcast News. Pričakovano je, da je uspešnost jezikovnega modela BNSI na bazi TEDx-UM, ki predstavlja prehod na novo domeno, manjša.

Slovar razpoznavalnika govora obsega 64.000 besed in je prilagojen domeni BNSI, saj je v pretežni meri sestavljen iz besed korpusov BNSI-Speech in BNSI-Text. Za novo domeno bi bilo sicer smiselno sestaviti nov slovar, a nas v članku zanima, kakšno pokritost nove domene daje obstoječi jezikovni model. Po drugi strani pa nove besede v slovarju razpoznavanja ne izboljšajo, če jezikovni model nima znanja o pojavitvah teh besed. Res je, da na ta način zmanjšamo OOV, ni pa nujno, da izboljšamo razpoznavanje, saj majhne verjetnosti jezikovnega modela zelo verjetno hipoteze, ki vsebujejo nove besede, izločijo iz nabora najverjetnejših hipotez.

Predavanje	Tematika	JM1 PP	JM2 PP	OOV
1	potovanja	409	431	21%
2	tehnologija	390	412	23%
3	družba	440	475	22%
4	tehnologija	379	405	28%
5	umetnost	481	506	26%
6	družba	491	491	26%
7	znanost	323	336	22%
8	znanost	242	234	20%
9	družba	429	451	27%
10	umetnost	400	399	24%
11	družba	428	451	19%
12	znanost	402	412	24%
13	družba	287	260	23%
vsa	različna	390	403	24%
BNSI eval	različna	247	387	4%

Tabela 1: Rezultati jezikovnih modelov UMB Broadcast News in FidaPLUS na testnih vzorcih SI TEDx-UM.

Glede na to, da sta si domeni BNSI in TEDx precej različni, smo preverili tudi uspešnost jezikovnega modela, učenega samo na korpusu FidaPLUS, ki predstavlja bolj splošno domeno. Korpus FidaPLUS je referenčna zbirka pisnih besedil in s tem, ko iz jezikovnega modela odstranimo komponente, ki modelirajo govorjeno rabo, izgubimo modelirane odvisnosti, tipične za govorjeno rabo, ki jih v pisni rabi ni ali pa so zelo redke in se v velikem korpusu zameglijo. Opozoriti velja tudi na značilnosti spontanega govora, ki jih niti jezikovni model BNSI ne modelira, saj gre pretežno za brani govor.

5. Rezultati

Izvedli smo eksperimentalno primerjavo rezultatov razpoznavanja govora z uporabo dveh različnih jezikovnih modelov: interpoliranega na besedilih iz televizijskih informativnih oddaj ter splošnega, grajenega izključno na besedilnem korpusu FidaPLUS. Najprej smo primerjali uspešnost obeh jezikovnih modelov na testnih vzorcih SI TEDx-UM. Rezultati so zbrani v tabeli 1. JM1 PP je perpleksnost jezikovnega modela UMB Broadcast News, JM2 PP pa jezikovnega modela FidaPLUS. Ker oba jezikovna modela uporabljata isti slovar, je delež OOV za oba enak. Perpleksnosti jezikovnega modela FidaPLUS so

boljše le na vzorcih 8 in 13, na vseh ostalih pa slabše. Sklepamo, da je modeliranje govorjene rabe jezika pomembna komponenta jezikovnega modela. Visok delež besed izven slovarja kaže na različnost domen BNSI in TEDx, velike vrednosti perpleksnosti obeh jezikovnih modelov pa na specifičnost domene TEDx.

V drugem koraku vrednotenja obeh jezikovnih modelov smo izvedli eksperimente razpoznavanja govora na ročno transkribiranem testnem naboru baze SI TEDx-UM. Rezultati so podani v tabeli 2 v obliki napake razpoznanih besed (NRB).

Predavanje	JM1 NRB(%)	JM2 NRB(%)
1	50,5	51,3
2	54,7	56,6
3	57,7	58,5
4	39,2	38,5
5	67,1	67,6
6	46,1	45,3
7	52,9	53,3
8	35,5	34,9
9	51,4	52,9
10	35,0	35,5
11	52,4	51,0
12	70,3	69,3
13	38,9	35,1
vsa	50,7	50,7
BNSI eval	26,6	26,6

Tabela 2: Napaka razpoznavanja govora na testnih vzorcih baze SI TEDx-UM za oba jezikovna modela.

Oba jezikovna modela sta dosegla 50,7% napako razpoznavanja besed, kar kaže na to, da imata specifični jezik in tematika predavanj bistveno večjo težo v primerjavi z različnimi besedilnimi viri, ki smo jih vključili v jezikovni model. Analiza na nivoju posameznih predavanj je pokazala, da je NRB podobna, ne glede na uporabljeni jezikovni model. Do edinega odstopanja je prišlo v primeru predavanja številka 13, kjer je JM2 dosegel za 3,8% boljši rezultat. Najboljši rezultat za posamezno predavanje je napaka razpoznavanja besed 34,9%, dosežena pri predavanju številka 8, ki ima tudi najnižjo perpleksnost jezikovnega modela.

Primerjava z bazo BNSI je pokazala da so doseženi rezultati razpoznavanja govora slabši kot tisti, doseženi na govorni bazi BNSI Broadcast News, kar lahko pripišemo bistveno višjemu deležu besed izven slovarja (OOV) ter razlikam v domeni in karakteristikah med obema govornima viroma.

6. Zaključek

Govorna baza SI TEDx-UM predstavlja pomemben nov vir za razvoj govornih tehnologij, saj odpira nove možne tematike raziskovalnega dela za slovenski jezik. Z uporabo takšnih pristopov bo možno v prihodnje bistveno učinkoviteje graditi nove govorne vire, ki so še vedno neobhodno potrebni za nadaljnji razvoj govornih tehnologij, kot je avtomatsko razpoznavanje tekočega govora. Prvi rezultati razpoznavanja govora na bazi SI-TEDx-UM kažejo velik vpliv tematike predavanj, ki pomembno odstopa od trenutno obstoječega sistema za razpoznavanje govora.

V prihodnje bomo skušali rezultat razpoznavanja govora izboljšati s prilagojenimi jezikovnimi modeli, ki jih bomo gradili na novih jezikovnih virih, predvsem na transkribiranem gradivu TEDx in gradivu TED v drugih jezikih, ki je prevedeno v slovenščino.

Govorna baza SI TEDx-UM je v skladu z licenco Creative Commons 3.0 prosto dostopna na spletni strani Inštituta za elektroniko in telekomunikacije UM FERI: http://ietk.feri.um.si/en/portfolio/sitedxumenglish.

7. Literatura

Špela Arhar in Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2, 95--110.

Anthony Rousseau et al.. 2014. Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks. *Proc. of the LREC'14,* Reykjavik, Islandija.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. *International Conference on Speech and Language Processing*, II: 901--904.

Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek in Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language resources and evaluation*, 47(4):1031-1048.

Darinka Verdonik. 2014. Vprašanja zapisovanja govora v govornem korpusu Gos. *Jezikovne tehnologije. IS 2014,* Ljubljana, Slovenija, str. 151-156.

Andrej Žgank, Tomaž Rotovnik, Darinka Verdonik, Zdravko Kačič. 2004. Baza broadcast news za slovenski jezik (BNSI) in sistem za razpoznavanje tekočega govora. *Jezikovne tehnologije. IS 2004,* Ljubljana, Slovenija, str. 98-97.

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. *Proc. of the LREC'14*, Reykjavik, Islandija, str. 2644-2647.

Andrej Žgank, Gregor Donaj, Mirjam S. Maučec. 2014a. Razpoznavalnik tekočega govora UMB Broadcast News 2014 : kakšno vlogo igra velikost učnih virov? *Jezikovne tehnologije. IS 2014,* Ljubljana, Slovenija, str.147-150.

Janez Žibert, France Mihelič. 2004. Development, evaluation and automatic segmentation of Slovenian broadcast news speech database. *Jezikovne tehnologije. IS 2004,* Ljubljana, Slovenija, str. 72-78.

Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino

Kaja Dobrovoljc,* Tomaž Erjavec,† Simon Krek‡

* Zavod za uporabno slovenistiko Trojina
Trg republike 3, 1000 Ljubljana
kaja.dobrovoljc@trojina.si
† Odsek za tehnologije znanja, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si
‡ Laboratorij za umetno inteligenco, Institut »Jožef Stefan«
Jamova cesta 39, 1000 Ljubljana
simon.krek@ijs.si

1 Uvod

Tako kot na drugih področjih procesiranja naravnih jezikov se tudi na področju skladenjskega razčlenjevanja pojavlja vse večja potreba po poenotenju označevalnih sistemov, ki so bili razviti za označevanje posameznih jezikov ali besedilnih zbirk, saj njihova raznolikost onemogoča neposredno primerjavo podatkov in na njih temelječih orodij. Kot protiutež tovrstni razdrobljenosti je bila nedavno vzpostavljena mednarodna pobuda Universal Dependencies (UD),[1] ki si prizadeva za medjezično usklajeno skladenjsko razčlenjevanje besedil z namenom primerjalnih evalvacij, razvoja večjezičnih razčlenjevalnikov, medjezičnega učenja jezikovnih modelov in kontrastivnih jezikoslovnih analiz (Nivre et al., 2016), njeni glavni principi (Nivre 2015) pa v veliki meri temeljijo na drugih predhodnih standardizacijskih projektih (de Marneffe et al., 2014; McDonald et al., 2013; Petrov et al., 2012; Zeman, 2008). Doslej je bilo z označevalno shemo UD označenih že več kot 50 korpusov različnih svetovnih jezikov, med njimi tudi *Univerzalna odvisnostna drevesnica za slovenščino*, skladenjsko razčlenjeni korpus pisne slovenščine.

Univerzalna odvisnostna drevesnica za slovenščino predstavlja tretji samostojni označevalni sistem in četrto prosto dostopno zbirko s formaliziranimi podatki o skladenjskih razmerjih v ročno označenih besedilih v slovenskem jeziku. Ker se je sistem prvega skladenjsko razčlenjenega korpusa, Slovenske odvisnostne drevesnice (SDT, pribl. 30.000 pojavnic) (Džeroski et al., 2006; Erjavec in Ledinek, 2006), ki je izhajal iz modela Praške odvisnostne drevesnice (Hajič et al., 2001), glede na kadrovske in finančne omejitve izkazal za preveč kompleksnega, je bil v okviru projekta Jezikoslovno označevanje slovenščine (JOS) v letih 2007–2009 načrtno razvit robustnejši nabor skladenjskih kategorij (Ledinek in Erjavec 2009). Po tej shemi je bil najprej razčlenjen korpus jos100k (pribl. 100.000 pojavnic), ki je bil nato v okviru projekta Sporazumevanje v slovenskem jeziku (SSJ) v letih 2009–2011 razširjen v korpus ssj500k (Krek et al., 2015), v katerem skladenjsko razčlenjeni del predstavlja nekaj manj kot polovico celotnega korpusa (pribl. 235.000 pojavnic). Ta najobsežnejša zbirka ročno razčlenjenih besedil v slovenščini je bila tako izbrana kot osnova za izdelavo Univerzalne odvisnostne drevesnice za slovenščino, ki jo na kratko predstavljamo v nadaljevanju.

2 Pretvorba iz sistema JOS v sistem UD

Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico je bila zasnovana kot povsem avtomatiziran proces, pri čemer pa je ta glede na številne razlike med obema označevalnima sistemoma, zlasti na ravni skladenjskega opisa, zahteval izdelavo kompleksnega sistema pretvorbenih pravil. Po pretvorbi formata XML TEI, v katerem je izvorno zapisan korpus ssj500k, v tabelarični format CONLL-U, kakršnega predvideva sistem UD, je sledila vsebinska pretvorba na dveh ločenih ravneh: oblikoslovni, ki prinaša informacije o besedni vrsti pojavnic in njihovih oblikoslovnih lastnostih, ter skladenjski ravni, ki vključuje pripis podatka o skladenjski vlogi te pojavnice v stavku. Segmentacijska, tokenizacijska in lematizacijska načela korpusa ssj500k so ostala nespremenjena.

2.1 Oblikoskladenjska raven

Sistem UD razlikuje med 17 univerzalnimi besednimi vrstami, pri čemer se načela besednovrstnega razvrščanja posameznih tradicionalno problematičnih skupin (npr. glagolnikov in deležnikov) večinoma ujemajo z načeli sistema JOS. Med temeljnimi spremembami glede na nabor besednih vrst JOS lahko izpostavimo delitev samostalnikov na občno- in lastnoimenske, delitev veznikov na priredne in podredne,

[1] Spletna stran: http://universaldependencies.org/.

delitev glagolov na pomožne in vse ostale, delitev ločil na pravopisna ločila in simbole, premik kategorije okrajšav v kategorijo 'drugo' in premik nekaterih skupin številnikov med pridevnike. Največja novost glede na sistem JOS in dosedanje slovnične opise slovenskega jezika nasploh pa je kategorija določilnikov (*DET*), v katero se umeščajo besede, ki modificirajo samostalniške zveze in izpostavljajo njihovo referenco v kontekstu. Čeprav se ta kategorija še naprej kaže kot eden izmed preizkusnih kamnov predlaganega standarda, se v slovenski drevesnici vanjo po vzoru drugih slovanskih jezikov umeščajo nekatere skupine zaimkov (npr. *moj, ta, enak*) in prislovov (npr. *nekaj, toliko*), kadar so rabljeni v vlogi določil samostalniških besednih zvez.

Tudi po shemi UD se lahko posameznim besednim oblikam poleg besedne vrste pripisujejo podrobnejše leksikalne in slovnične lastnosti v obliki parov oblikoslovnih lastnosti in njihovih vrednosti. Za razliko od besednovrstnih kategorij njihov nabor ni končen, saj je odvisen od nabora lastnosti, med katerimi razlikujejo posamezni jeziki ali izvorne označevalne sheme, mednarodno usklajena pa so njihova poimenovanja. Seznam 22 oblikoslovnih lastnosti, ki jih vsebuje Univerzalna odvisnostna drevesnica za slovenščino, je skupaj s podrobnejšim opisom in razmerji glede na sistem JOS predstavljen v slovenski dokumentaciji na projektni spletni strani, med večjimi spremembami v primerjavi z naborom lastnosti JOS pa lahko izpostavimo predvsem uvedbo glagolske lastnosti načina (povedni, pogojni, velelni), podrobnejšo členitev številskih tipov (glavni, vrstni, množica, splošno), ki se lahko pripisujejo tudi pridevnikom, ter ukinitev kategorije dvovidskosti.

Na podlagi primerjave podobnosti in razlik v obeh sistemih je bil nato izdelan sistem pravil v obliki medtabeličnih preslikav besednih vrst in lastnosti, pri čemer se ena lastnost sistema JOS lahko prevede v različne lastnosti UD, izbira ustreznega pravila pa je v teh primerih lahko odvisna od leme besede, od drugih oblikoskladenjskih lastnosti ali njene skladenjske vloge.[2] V primeru, da neki vhodni pojavnici ustreza več pravil, imajo specifična pravila prednost pred splošnimi.

2.1 Skladenjska raven

Čeprav oba sistema skladenjskega razčlenjevanja v izhodišču temeljita na teoriji odvisnostne slovnice (Tesniere, 1959; Kübler et al., 2009), se shema UD v nekaterih vidikih od sistema JOS bistveno razlikuje. Najpomembnejša razlika izhaja iz samega obsega skladenjske analize, saj je bil sistem JOS zasnovan predvsem za razčlenjevanje vezljivostnih dopolnil povedka (stavčnih členov) in strukture besednih zvez, medtem ko sistem UD v skladenjsko analizo vključuje tudi vse druge tipe stavčnih struktur, kot so členki (*advmod*), pristavki (*appos*), nagovori (*vocative*), medmeti, pastavki in drugi elementi interakcije (*discourse*), tujejezični elementi (*foreign*), ločila (*punct*), soredja (*parataxis*) itd. Ker je taksonomija UD s 40 univerzalnimi skladenjskimi oznakami[3] bistveno obsežnejša kot nabor 10 oznak sistema JOS, tudi pri razčlenjevanju jedrnih skladenjskih struktur predvideva natančnejše opredelitve skladenjskih razmerij in izdelavo kompleksnejših skladenjskih dreves kot robustni JOS. Tipični primer so denimo besedne zveze, pri katerih sistem UD razlikuje med več različnimi tipi prilastkov (npr. *amod, nmod, nummod, advmod; det; acl*), funkcijskih modifikatorjev (npr. *case, neg, expl, cc*) in razmerij znotraj stalnih besednih zvez (npr. *mwe, name, compound, goeswith*).

Druga temeljna razlika med obema sistemoma izhaja iz deleža vključevanja semantičnih interpretacij v samo skladenjsko analizo, saj sistem JOS pri razčlenjevanju stavčnih členov na določenih mestih upošteva tudi njihovo pomensko vlogo (npr. ločevanje med načinovnimi in drugimi prislovnimi določili), medtem ko sistem UD razlikuje zgolj med t. i. jedrnimi argumenti (osebek in direktni/indirektni predmet) na eni strani ter vsemi ostalimi argumenti povedka na drugi, ne glede na stopnjo njihove vezljivostne ali pomenske obveznosti. Pri tem sistem UD argumente podrobneje razvršča tudi glede na njihovo skladenjsko strukturo, torej ločuje med besednozveznimi in stavčnimi ubeseditvami osebkov (*nsubj* proti *csubj*), predmetov (*dobj/iobj* proti *ccomp*) in drugih dopolnil (*nmod/advmod* proti *advcl*).

Skripto za samodejno pretvorbo skladenjske ravni korpusa ssj500k tako sestavlja niz številnih podrobnih pravil, ki vsaki pojavnici korpusa določijo tip skladenjske povezave in njen nadrejeni element po sistemu UD.[4] Ker zaradi robustnosti sistema JOS vseh neopredeljenih struktur (tj. struktur, vezanih na korenski element) v korpusu ssj500k ni bilo mogoče z dovolj zanesljivo natančnostjo samodejno preslikati v sistem UD, ki obenem dopušča le eno korensko povezavo v povedi, trenutna različica Univerzalne odvisnostne drevesnice za

[2] Primer pravila za pretvorbo na oblikoslovni ravni je denimo ukaz, da se po sistemu UD besedna vrsta pomožnik (*AUX*) pripiše vsem pojavnicam, ki imajo po sistemu JOS na oblikoslovni ravni pripisano kategorijo 'glagol' in vrsto 'pomožni', na skladenjski ravni pa so označeni s povezavo 'del' povezani na drugo glagolsko pojavnico.

[3] V slovenski drevesnici se trenutno pojavlja 31 različnih univerzalnih ali jezikovnospecifičnih skladenjskih oznak, saj ta poleg oznak za strukture, ki se v korpusu niso pojavljale, ne vsebuje tudi povezav, ki jih ni bilo mogoče razdvoumiti ali prepoznati samodejno.

[4] Primer pravila za pretvorbo na skladenjski ravni je denimo ukaz, da se po sistemu UD skladenjska povezava prirednega veznika (*cc*) pripiše vsem pojavnicam, ki so na oblikoslovni ravni UD kategorizirane kot priredni veznik (*CONJ*), podredni veznik (*SCONJ*), členek (*PART*) ali drugo (*SCONJ*), na skladenjski ravni JOS pa so s povezavo 'vez' povezane na pojavnico, ki je sama cilj povezave 'prir'. Za razliko od sistema JOS, kjer je v priredjih veznik podrejen zadnjemu elementu priredja, se po sistemu UD tej pojavnici kot nadrejena pojavnica pripiše prvi element priredja.

slovenščino vsebuje manj povedi kot izhodiščni korpus ssj500k (7.996 proti 11.411), a je tako po obsegu (140.418 pojavnic) kot povprečni dolžini povedi (17,6 pojavnic na poved) primerljiva z univerzalnimi drevesnicami za druge jezike.

3 Dostopnost in nadaljnje raziskave

Najnovejša, druga različica Univerzalne odvisnostne drevesnice za slovenščino je bila skupaj s 54 drevesnicami za 40 drugih svetovnih jezikov, vključno s komplementarno Univerzalno odvisnostno drevesnico govorjene slovenščine (Dobrovoljc in Nivre, 2016), objavljena kot del zbirke Universal Dependencies 1.3 (Nivre et al., 2016), pod licenco CC BY-NC-SA 4.0. Po njej in drugih drevesnicah je mogoče brskati preko dveh spletnih konkordančnikov,[5] poleg številnih večjezičnih razčlenjevalnih sistemov, ki temeljijo na tej ali predhodnih različicah te korpusne zbirke in potrjujejo pomen vpetosti slovenskih jezikovnih virov v širši jezikovnotehnološki prostor, pa je bil nedavno razvit tudi namenski spletni servis[6] za večnivojsko označevanje neoznačenih besedil z jezikovnimi modeli UD, ki tudi za slovenščino dosegajo zelo dobro natančnost (Straka et al., 2016). V prihodnosti nameravamo obstoječo različico slovenske drevesnice nadgraditi v skladu s posodobljenimi smernicami za označevanje, dopolniti njeno spletno dokumentacijo in razširiti nabor pravil za vključitev manjkajočih delov korpusa ssj500k. Poleg podrobnejših jezikoslovnih analiz skladenjskih specifik slovenskega jezika pa bi bilo z jezikovnotehnološkega vidika prioritetno raziskati vpliv spremembe označevalne sheme na natančnost referenčnih označevalnih orodij za slovenščino (Grčar et al., 2012; Dobrovoljc et al., 2012) in na podlagi rezultatov ovrednotiti potrebo po nadaljnjem vzdrževanju samostojnega sistema JOS.

4 Literatura

Kaja Dobrovoljc, Simon Krek in Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. V: *Zbornik Osme konference Jezikovne tehnologije*, str. 42–47.

Kaja Dobrovoljc in Joakim Nivre. 2016. The Universal Dependencies Treebank of Spoken Slovenian. V: *Proceedings of LREC'16*, str. 1566–1573.

Sašo Džeroski et al. 2006. Towards a Slovene Dependency Treebank. V: *Proceedings of LREC'06*, str. 1388–1391.

Tomaž Erjavec in Nina Ledinek. 2006. Slovenska odvisnostna drevesnica: prvi rezultati. V: *Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS 2006*, str. 162–167.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: *Zbornik Osme konference Jezikovne tehnologije*, str. 89–94.

Jan Hajič et al. 2001. The Prague Dependency Treebank: Annotation Structure and Support. V: *Proceedings of the IRCS Workshop on Linguistic Databases*, str. 105–114.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek in Nanika Holz. 2015. Training corpus ssj500k 1.4*, Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1052.

Sandra Kübler, Ryan McDonald in Joakim Nivre. 2009. *Dependency Parsing*. Morgan and Claypool.

Nina Ledinek in Tomaž Erjavec. 2009. Odvisnostno površinskoskladenjsko označevanje slovenščine: specifikacije in označeni korpusi. V: *Infrastruktura slovenščine in slovenistike (Obdobja 28)*, str. 219–224.

Marie-Catherine de Marneffe et al., 2014. Universal Stanford Dependencies: A cross-linguistic typology. V: *Proceedings of LREC'14*, str. 4585–4592.

Ryan McDonald et al., 2013. Universal Dependency Annotation for Multilingual Parsing. V: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, str. 92–97.

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. *Computational Linguistics and Intelligent Text Processing*, (9041):3–16.

Joakim Nivre et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. V: *Proceedings of LREC'16*, str. 1659–1666.

Joakim Nivre et al. 2016. Universal Dependencies 1.3*, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague,* http://hdl.handle.net/11234/1-1699.

Slav Petrov, Dipanjan Das in Ryan McDonald. 2012. A universal part-of-speech tagset. V: *Proceedings of LREC'12*, str. 2089–2096.

Milan Straka, Jan Hajic in Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. V: *Proceedings of LREC'16*, str. 4290–4297.

Lucien Tesnière. 1959. *Eléments de Syntaxe Structural*, Paris: Librairie C. Klincksieck.

Daniel Zeman. 2008. Reusable Tagset Conversion Using Tagset Drivers. V: *Proceedings of Sixth International Conference on Language Resources and Evaluation (LREC'08)*, str. 213–218.

[5] SETS: http://bionlp-www.utu.fi/dep_search; PML Tree Query: http://lindat.mff.cuni.cz/services/pmltq/.

[6] UDPipe: http://lindat.mff.cuni.cz/services/udpipe/.

What is critical in digital humanities?

Mario Hibert

Odsjek za komparativnu književnost i bibliotekarstvo Filozofskog fakulteta Univerziteta u Sarajevu
Franje Račkog 1, 71000 Sarajevo
mario.hibert@gmail.com; mario.hibert@ff.unsa.ba

1. Uvod

Polazište za ovo istraživanje kreće od činjenice da umreženost odnosno "umreženo znanje" u okvirima digitalne humanistike biva dominantno određeno tehnološkim izazovima upravljanja podacima usljed čega izostaje kritički diskurs u recentnim znanstveno-istraživačkim praksama. Podsjećajući na termin "kritičke informacijske studije" nastao prije deset godina u radu "Critical Information Studies: A Bibliographic Manifesto", Sive Vaidhyanathana, profesora medija i prava sa Virginia Univerziteta, kojim se sintetiziranjem transdisciplinarnih presjecišta kulturalnih studija i političke ekonomije naglašava problematiku korporatizacije javne informacijske infrastrukture (Vaidhyanathan, 2006), ovaj rad nastoji naglasiti važnost akademskih rasprava i praksi koje u fokusu imaju javni interes za četri ključna područja koja Vaidhyanathan označava kao fokalne tačke analize i rasprave u kontekstu kritičkih informacijskih studija: a. mogućnosti i slobode korištenja, revidiranja, kritiziranja i rukovanja kulturalnim tekstovima, slikama, idejama i informacijama; b. prava i mogućnosti korisnika (ili potrošača ili građana) da mijenjaju sredstva i tehnike kojima su kulturalni tekstovi i informacije pružene, prikazane i distribuirane; c. odnos između informacijske kontrole, vlasničkih prava, tehnologije i društvenih normi; d. kulturna, politička, društvena i ekonomska ograničenja globalnih tokova informacija (Vaidhyanathan, 2006). Upućujući kako ovaj "derivate kritičke teorije i informacijske teorije" (Ibidem) ne bi smio ostati neprepoznat u i izuzet iz konteksta rasprava o kritičkoj digitalnoj humanistici, posebice danas kada otvorenost i pristup znanju postaju trendovske floskule u novom znanstveno-istraživačkom rječniku, a politički i pravni status digitalnih dokumenata i podatkovne infrastrukture postaje sistemski tj . strukturalni nexus algoritamskog upravljanja, pitanja postavljena u ovome radu imaju za cilj problematiziranje tzv. „podatkovnog fetišizma koji dominira ovim brzorazvijajućim poljem istraživanja" (Hall, 2012).

2. Cilj rada

Premda istraživanja iz područja digitalne humanistike predstavljaju jedan od najrecentnijih trendova u znanstveno-istraživačkim zajednicama, čini se kako njihov najveći udio u svoje obzorje ne uključuje problematiku algoritamskog eksploatiranja, već vizualizacije i rekombinacije, informacijskih resursa. Drugim riječima, interdisciplinarna presjecišta nove kulturalne logike tzv. kibernetike 2.0. nerjetko ostaju izvan uvida kritičkih informacijskih studija koje otvaraju prostor za dijalog, kako sa kompjuterskom znanošću odnosno tako i interpasivniom ulogom "digitalnih humanista" u umreženom društvu. Cilj rada je aktualizirati važnost uvida kritičkih društvenih i humanističkih znanosti, posebice iz perspektiva koje baštine odnos informacijskih znanosti i bibliotekarstva (sistemski vs. korisnički pristup), kako bi se naglasio značaj refokusiranja digitalne humanistike sa tzv. softverskih efekata tehnokulturalizma na digitalnu kolonijalizaciju društvenosti. Namjera je naglasiti značaj kritičkih informacijskih studija za razvoj digitalne humanistike širenjem dijaloškog okvira ka širem spektru znanstvenih predmeta, višestrukim komplementarnim metodologijama ondnosno transdisciplinarnim debtama kojima se propituje kulturalna logika digitalizma.

2.1. Digitalna humanistika u društvu metapodataka

Ispreplетenosti globalne ekonomije sa digitalnim radom, sve intenzivnije organiziranim u otvorenom, umreženom tržišnom režimu, razumijevanje upravljanja mrežnim viškom vrijednosti čini iznimno važnim polazištem za osvještavanje biopolitičke kontrole umrežnog društva. Dogma kibernetičkog totalitarizma, otjelotvorena kroz dominaciju korporativnih platformi, čini se nedovoljno artikuliranom među zagovornicima digitalne humanistike koji se svojim praksama uključuju u informacijski ekosistem ekonomskog modela koja rezultira "infrastrukturalnim imeprijalizmom" (Vaidhyanathan, 2011). Tehno-utopistička ideologija novih medija, aktualizirana proizvodno/distributivnim potencijalima Mreže, prikriva kako se kreiranjem tzv. "viška vrijednosti koda" (*code surplus value*) digitalni rad eksploatira pod okriljem kapitalizma platforme (*platform capitalism*). Naime, novi epistemički poredak generiran "uređajima za akumulaciju valoriziranih informacija, ekstrakciju metapodataka, kalkulaciju mrežnog viška vrijednosti i hranjenje mašinske inteligencije" (Pasquinelli, 2014) nastao je kao posljedica integracije kibernetičkih mašina sa kognitivnim dimenzijama (digitalnog) rada. Novi oblici društvene proizvodnje upravljani informacijskim vektorima odnosno

kontrolirani algoritmima kreiraju tzv. "društvu metapodataka" (Pasquinelli, 2014), zbog čega problematika akumulacije i upravljanja mrežnim viškom vrijednosti sugerira važnost aktualiziranja istraživanja iz domena političke ekonomije Mreže kako bi se kritika kognitivnog kapitalizma mogla značajnije intenzivirati i u istraživanjima iz domena digitalne humanistike. Drugim riječima, taktička refokalizacija akademske pažnje sa tehničkih na društvene, ekomomske i političke konsekvence razvoja i korištenja digitalnih alata, jedan je od preduvjeta kritičkog promišljanja digitalne humanistike, kao i njezinih (h)aktivističkih konotacija (Kirschenbaum, 2012).

2.2. Metapodatkovni punk

Kontroverze u vezi sa hakerskim etosom informacijskog doba, koji se otjelovljuje u otporu spram klase tzv. vektoralista koja upravlja tokovima informacija (Wark, 2015), danas su premjestile svoj fokus sa diskusija o zaštiti autorskog prava i intelektualnog vlasništva na pitanja o mogućnostima „oslobađanja" metapodataka: "važnije je osloboditi informacije o informacijama nego same informacije" (Wark, 2015). Taktički obrat o kojem govori McKenzie Wark proističe iz činjenice da strategije novih oblika eksploatacije pripadaju tzv. strvinarskim industrijama (Wark, 2015) koje su, za razliku od starih kulturnih industrija koje su nastojale zaustaviti slobodan protok informacija, prepoznale kako kontroliranje metapodataka u okruženju obilja neplaćenog digitalnog rada predstavlja novu stepenicu razvoja korporativne hegemonije. Drugim riječima, "ako današnje velike baze podataka, tzv. big data, i kontrola informacijskih tokova iznova čine prirodnima i prihvatljivima devetnaestostoljetne pozitivističke pretpostavke o upravljivosti društva, onda beskonačne mogućnosti rekombinacije odnosa ili veza između kulturnih predmeta prijete da preplave taj iznova oživljeni epistemički okvir barbarizma moderne u njegovom kibernetičkom obličju" (Medak, 2015). Stoga je projekte poput npr. „Javna knjižnica" Multimedijalnog instituta iz Zagreba, koji ukazuju na radikalne prakse otvorenosti (Library Genesis, aaaaarg.org, Monoskop, UbuWeb), te reafirmiraju avangardne bibliotečke prakse kroz primjere kreiranja tzv. „fragilnih infrstruktura znanja...s onu stranu diktata komodifikacije i kontrole" (Ibidem), moguće iščitati i kao evoluciju digitalne humanistike ka njezinim kritičkim formama.

3. Zaključak

U članku „*The Digital Humanities or a Digital Humanism"* Dave Parry ističe kako trendovi digitalne humanistike, upisujući i propisujući veoma konzervativnu formu znanstevnog istraživanja u području humanistike, kulu od slonovače mijenjaju superkompjuterima i farmama servera što vodi zamjeni jednog izolacionizma drugim (Parry, 2012). Za značajniji angažman istraživača u području digitalne humanistike spram ideologema komunikacijskog kapitalizma odnosno kibernetičke (de)regulacije društvenosti je stoga neophodno potaknuti istraživanja žarišnih fenomena umreženog društva koja namjesto digitalne euforije nude „novo epistemičko oko" (Pasquinelli, 2014) čiji pogled može proniknuti do alternativnih vizura globalizacije, suvremenih oblika otpora ideologiji vektoralizma odnosno afirmativne sabotaže (Spivak, 2010) uvjetovane mogućnostima estetskog obrazovanja koje traži snažna humanistička uporišta bez kojih je nemoguće vježbati imaginaciju i njezine epistemološke performanse (Ibidem). U konačnici, prema riječima Geert Lovnika: "*we do not need more tools; what's required are large research programs run by technologically informed theorists that finally put critical theory in the driver's seat. The submissive attitude in the arts and humanities towards the hard sciences and industries needs to come to an end*".

4. Literatura

Dave Parry. 2012. The Digital Humanities or a Digital Humanism. In: Matthew K. Gold, ed., Debates in the Digital Humanities. Minneapolis, MN: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/48

Gary Hall. 2012. There Are No Digital Humanities. In: Matthew K. Gold, ed., Debates in the Digital Humanities. Minneapolis, MN: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/21

Gayatri Chakravorty Spivak. 2010. An Aesthetic Education in the Era of Globalization. https://vimeo.com/23032519

Matthew Kirschenbaum. 2012. Digital Humanisties As/Is a Tactical Term. In: Matthew K. Gold, ed., Debates in the Digital Humanities. Minneapolis, MN: University of Minnesota Press.

Matteo Pasquinelli. 2014. Italian Operaismo and the Information Machine. Theory, Culture, Society, 32(3):49-68.

Matteo Pasquinelli. 2014. The Eye of the Algorithm: Cognitive Antrhophocene and the Making of the World Brain. http://notyouraverage.website/cloudish/pasquinelli%20_%20eye%20of%20the%20algorithm.pdf

McKenzie Wark. 2015. Metapodatkovni punk. U: M. Mars i T. Medak, Javna Knjižnica, str. 41-47. Zagreb: Multimedijalni institut. http://www.whw.hr/download/books/medak_mars_whw_public_library_javna_knjiznica.pdf

Siva Vaidhyanathan. 2006. Critical Information Studies: A Bibliographic Manifesto. Cultural Studies, 20(2-3):292-315.

Siva Vaidhyanathan. 2011. The Googlization of Everything. Berkeley, CA: University of California Press.

Tomislav Medak. 2015. Budućnost iz knjižnice. U: M. Mars i T. Medak, Javna Knjižnica, str. 51-66. Zagreb: Multimedijalni institut.
http://www.whw.hr/download/books/medak_mars_whw_public_library_javna_knjiznica.pdf

Asian Language Teaching and Learning – The Influence of Technology on Students' Skills in SL Classroom

Marijana Janjić,[*] Sara Librenjak,[†] Kristina Kocijan [‡]

[*†‡]Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučica 3, 10 000 Zagreb, Croatia
marijanajanjic@yahoo.com[*], sara.librenjak@gmail.com[†], krkocijan@ffzg.hr[†]

1. Introduction

E-learning/teaching tools have been around for quite some time now. Although reports show many of their positive sides in educational environment (Dovedan, Seljan and Vučković, 2002; Lauc, Matić and Mikelić, 2006; Sendra, Jiménez, Parra and Lloret, 2015) still we have more teachers who do not use these tools for their classes than those who do. Ever since computers have been introduced to the teachers, they have, for different reasons, underused its power as a valid educational tool. Due to the fact that we are living in the 21st century, it is valid to ask if we, as educators, are depriving our students of needed knowledge. Are we failing to educate a 21st century student?

When it comes to teaching a foreign language with e-tools, things seem to be even more complicated. It was in the 18th century that Alexander von Humboldt suggested that the language cannot be taught but that one can only create conditions for learning to take place. Three centuries later, deep into the digital era, it seems that we are still not recognizing the power of technology and its usage for creating such language learning oasis.

2. Goal of the paper

At the Faculty of Humanities and Social Sciences in Zagreb we are conducting a research on the benefits technology can have in second language (SL) classroom taking the students' perspective into account. The goal of the research is to develop e-materials easily embedded into regular classroom lessons, following the outlines of teacher's curriculum and materials he/she wants to cover in the course from the first to the last lesson. The project, funded by EU, is focused on the implementation of such materials in the teaching of several Asian languages (Hindi, Japanese, Korean and Sanskrit) available at the University as well as in several private language schools in Croatia. After the initial survey of students' issues (N= 203 students: Japanese 104, Korean 46, Hindi 32, Sanskrit 21), a range of study materials have been developed to help students with memorising and extracting crucial information about vocabulary, grammar or syntax at the right moment and with a greater speed and precision. In this process we have used existing application developed for memorizing, Memrise, Quizlet and Anki. The reasons for such choice are several: 1. majority of language teachers would choose an already existing app to develop new materials and not develop his/her own; 2. existing application have a good technical team working on technical issues and to test the influence of the technology in the classroom it's good to have a solidly developed application and not a prototype; 3. students were already familiar with the applications as they used them more or less regularly for different purposes.

The paper is interested in following questions: 1. which difficulties do students report in language learning (ex. Figure 1), 2. which factors influence use of technology in language classroom, 3. can the use of technology influence students' interest in the language learning positively, 4.what type of influence do we see in students' approach and knowledge after the use of materials created on the applications. The long-term goal of those questions is to improve the language teaching practice of Asian languages in Croatia, widen students language resources and create an e-environment for learners of Asian languages.

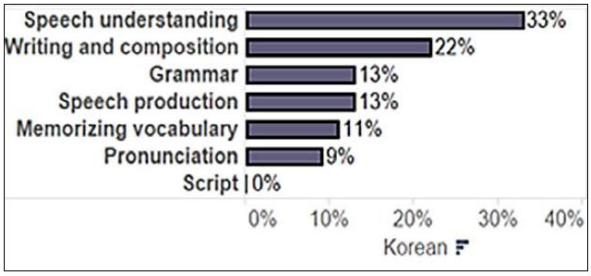


Figure 1: Students' analysis of issues in learning Korean.

Hence, next to the production of materials, the researchers also gathered data on students' usage of materials for a period of several months, roughly from November 2015 to September 2016, in order to relate their usage and the results that students got when they were administered tests. Tests are administered to students in interval of several months in the said period and had tested students' comprehension of grammar and vocabulary. Tests (Figure 2) were created having in mind materials covered by students' curriculum and newly developed materials available on applications.

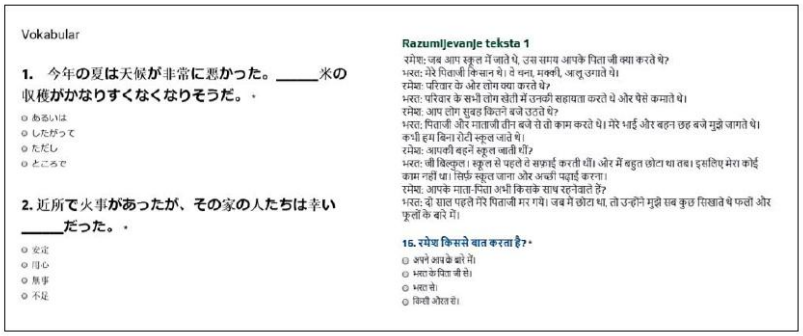


Figure 2: Question sample for Japanese and Hindi.

On the average, students have reacted positively on the creation of e-materials. After only 2 months of using the e-materials, they have obtained better test results which are more than 20% higher than their initial test scores. To understand the issues related to technology in language classroom better, research team had also conducted a short survey among language teachers of various languages at the Faculty, in order to reflect on the perception students and teachers have on the technology as such in the classroom. The results of students' usage show that teacher's involvement and perception of technology can and does leave its trace on students' scores.

3. References

Begoña Bellés-Fortuño and Noemi Ollero Ramírez. 2015. Motivation: A key to success in the foreign language classroom? A case study on vocational training and higher education English courses. 1st International Conference on Higher Education Advances, HEAd'15. http://dx.doi.org/10.4995/HEAd15.2015.431

Zdravko Dovedan, Sanja Seljan and Kristina Vučković. 2002. Multimedia in Foreign Language Learning. In: *Proceedings of the 25th International Convention MIPRO 2002: MEET + MHS*, pages 72–75, Rijeka.

Ela Družijanić Hajdarević, Kristina Vučković and Zdravko Dovedan. 2006. Računalo ili raćunalo uz pomoć računala. In: *Proceedings of the 29th International Convention MIPRO*, pages 283–287, Rijeka.

Tomislava Lauc, Sanja Matić and Nives Mikelić. 2006. Educational multimedia software for English language vocabulary. In: *Proceedings of the 1st International Conference on Multidisciplinary Information Sciences and Technologies: InSciT2006, Vol. I: Current Research in Information Sciences and Technologies Multidisciplinary approaches to global information systems*, pages 117–121, Merida.

Matea Leko, Sanja Kišiček, Josip Knežević, Ivana Martinović, Petar Krešimir Marić and Elvis Vusić. 2013. Interactive Application for Learning the Latin Language. In: *InFuture 1*, Faculty of Humanities and Social Sciences, pages 237–248, Zagreb.

Sara Librenjak, Kristina Vučković and Zdravko Dovedan Han. 2012. Multimedia assisted learning of Japanese kanji characters. In: *Proceedings of the 35th International Convention MIPRO*, pages 1284–1289, Rijeka.

Nadeem Saqlain. 2012. *Technology and Foreign Language Pedagogy: what the literature says.* http://er.educause.edu/articles/2012/6/technology-and-foreign-language-pedagogy-what-the-literature-says

Sandra Sendra, Jose M. Jiménez, Lorena Parra and Jaime Lloret. 2015. *Blended Learning in a Postgraduate ICT course*. 1st International Conference on Higher Education Advances, HEAd'15. http://dx.doi.org/10.4995/HEAd15.2015.491

Eiko Ushida. 2005. The Role of Students' Attitudes and Motivation in Second Language Learning in Online Language Courses. CALICO Journal, 23(1): 49–78. https://journals.equinoxpub.com/index.php/CALICO/article/view/23165/19170

Grace Wiebe and Kaori Kabata. 2010. Students' and instructors' attitudes toward the use of CALL in foreign language teaching and learning. *Computer Assisted Language Learning*, 23(3): 221–234.

Zbiva in EWD, spletni orodji za arheološke raziskave

Bojan Kastelic,* Mateja Belak,† Andrej Pleterski,† Benjamin Štular,† Miran Erič‡

* Samostojni raziskovalec
bojan.kastelic@telemach.net
† Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti, Inštitut za arheologijo,
Novi trg 2, 1000 Ljubljana
mateja.belak@zrc-sazu.si
andrej.pleterski@zrc-sazu.si
benjamin.stular@zrc-sazu.si
‡ Zavod za varstvo kulturne dediščine Slovenije
Cankarjeva 4, 1000 Ljubljana
miran.eric@guest.arnes.si

1 Uvod

V prispevku predstavljamo dve arheološki spletni orodji, ki sta bili narejeni z nadgradnjami platforme Arches, Zbiva in EWD. Obe zbirki bosta na kratko predstavljeni vsebinsko, poudarek pa je na tehnični predstavitvi nadgradenj.

2 Spletni orodji

2.1 Zbiva

Zbiva je arheološka **zbi**rka podatkov za **v**zhodne **A**lpe in obrobje v zgodnjem srednjem veku. Prostorsko pokriva Slovenijo, Avstrijo, Istro in Kvarner na Hrvaškem ter Furlanijo in Julijsko krajino v Italiji, primerjalno pa vsebuje tudi posamična najdišča iz soseščine in predhodnega obdobja. Časovno obsega gradivo in najdišča od naselitve Slovanov do prenehanja dajanja predmetov v grobove, najbolj grobo torej od 7. do 11. stoletja. Zbirka je začela nastajati že leta 1987, na spletu je dostopna od leta 2000.

Osnovo tvori *zbirka najdišč,* ki vsebuje podatke o legi in vrsti najdišč, vrsti najdbe ter o tem, kje jo danes hranijo, časovno opredelitev najdišča ter seznam literature. Drugi sestavni del je *zbirka grobov* z opisom bistvenih sestavin grobišč: zvrsti grobov, obrise in velikosti grobnih jam, globino vkopov, smer grobnih jam in pokojnikov v njih, lego in ohranjenost okostij, odnos okostja do drugega okostja v grobu, ali gre za dvojni ali skupinski grob, odnos groba do sosednjih grobov na grobišču ... Tretji sestavni del je *zbirka predmetov,* kjer so ti opisani številčno in besedno ter prikazani s sliko. Tipsko so opredeljeni lončenina, nakitni predmeti in noži.

Na tem mestu predstavljamo novo različico, v kateri se Zbiva iz podatkovne zbirke prek spletne zbirke razvija v spletno orodje za arheološke raziskave (http://zbiva.zrc-sazu.si/).

2.2 EWD

Globalna pobuda Early Watercraft, ki šteje 27 ambasadorjev z vsega sveta, ima cilj ustvariti sistematično raziskovalno okolje o izvorih navigacije na splošno, s poudarkom na izumih najzgodnejših avtohtonih plovil (debláki, traváki, kožáki, lubjáki, splavi), ki ob ognju, bivališču in orodjih za preživetje sodijo med najpomembnejše človeške izume. Prvi oprijemljivejši rezultat prizadevanja iniciative je spletno orodje **E**arly **W**atercraft **D**atabase (***EWD***). Vsebinsko je EWD namenjen na svetovni ravni združeni evidenci vseh arheološko raziskanih zgodnjih plovil in s tem dokumentiranju vseh pojavov tipološko zgodnjih plovil, ki so še danes pomemben gospodarsko-navigacijski pripomoček v različnih okoljih po vsem svetu. Namenjen pa je tudi dokumentiranju eksperimentalnoturističnim prizadevanjem za trajnostno ohranjanje kulturne dediščine navigacije človeštva, ki je skozi dediščinsko ozaveščenost po vsem svetu v velikem porastu. Taka podatkovna zbirka (http://www.earlywatercraft.org) bo povratno temelj in izhodišče za celovite, polidisciplinarne in poglobljene znanstvene raziskave izuma, ki po nekaterih signalih sega celo v čas *Homo erectusa*.

3 Arches

Poleg arheološke tematike imata Zbiva in EWD skupno še platformo Arches (http://archesproject.org). Začetki projekta Arches segajo v leto 2004, ko sta Inštitut za ohranjanje kulturne dediščine Getty (GCI) in Svetovni spomeniški sklad (WMF) formirala Iniciativo za ohranitev iraške kulturne dediščine (Iraq Cultural Heritage Conservation Initiative). Ker politične in varnostne razmere v Iraku v tistem času niso omogočale

napredka na tem področju, sta se organizaciji povezali z jordanskim oddelkom za antične umetnine (Jordanian Department of Antiquities) in leta 2010 uspešno zaključili projekt MEGA-Jordan, ki je v uporabi še danes. Med izdelavo projekta MEGA so se precej hitro pokazale možnosti, ki jih prinaša takšen sistem, in zanj so takoj pokazale zanimanje tudi druge organizacije, ki se ukvarjajo z zaščito kulturne dediščine. To je bilo glavno vodilo, da sta se obe organizaciji (GCI in WMF) odločili za razvoj splošnega odprtega programskega sistema za vodenje prostorskih podatkov za vse tipe nepremične kulturne dediščine. Projekt so poimenovali Arches.

Na podlagi izkušenj z izgradnjo sistema MEGA in sodelovanja z mnogimi organizacijami z vsega sveta so bila oblikovana naslednja načela, ki so postala temelj za razvoj sistema Arches:

1. Zasnova na podlagi standardov: Sistem mora temeljiti na uveljavljenih mednarodnih standardih s področja kulturne dediščine in informacijske tehnologije, s čimer spodbuja izmenjavo podatkov in njihovo dolgo življenjsko dobo, ki bo neodvisna od tehnološkega napredka.
2. Splošna dostopnost: Za zagotovitev kar najširše dostopnosti mora biti sistem dostopen prek spleta, njegova uporaba pa čim preprostejša.
3. Ekonomičnost: Kot odprtokodni sistem (open source system) mora biti brezplačen in mora omogočati uporabnikom pomoč skupnosti pri prilagoditvah in vzdrževanju sistema.
4. Razširljivost: Sistem mora biti zgrajen modularno in mora omogočati enostavne prilagoditve. Prav tako mora omogočati večjezičnost in prilagoditve za uporabo kjerkoli po svetu.
5. Varnost: Sistem mora omogočati poljubno stopnjo varovanja podatkov – lahko je popolnoma odprt javnosti, popolnoma zaprt ali pa nekje vmes (določen del funkcionalnosti je odprt javnosti, določen del pa dostopen le uporabnikom z ustreznimi pravicami).

Prva verzija sistema Arches (1.0) je bila predstavljena oktobra 2013, od takrat se izvajajo redne nadgradnje. Zbiva in EWD temeljita na tretji verziji (3.0) sistema, izdani aprila 2015.

4 Nadgradnje

Zbiva in EWD intenzivno uporabljata možnosti platforme, ki jih Arches ponuja že v osnovi, predvsem možnosti razširitve podatkov v okviru konceptualnega referenčnega modela (CIDOC CRM), vgrajenih funkcionalnosti iskanja, prikaza in prostorske predstavitve podatkov ter seveda urejanja podatkov. Poudarek prispevka pa je na predstavitvi nadgradenj platforme Arches za potrebe Zbive in EWD.

Nadgradnje za potrebe Zbive so zadevale predvsem:

- podporo večjezičnosti pri uvozu in prikazu podatkov, ki na platformi Arches še ni bila povsem dodelana;
- podporo novim tipom dokumentov (najdišča, grobovi, predmeti) s svojimi podatki;
- napredne možnosti iskanja, kot so ločitev iskanja po tipih dokumentov, večnivojski seznami in omejevanje po poljubnih merah;
- pripravo ponovljivega postopka prenosa podatkov iz osrednje zbirke, ki deluje v okolju MS Access, v Arches.

V okviru EWD smo v platformo vgradili čisto nove funkcionalnosti, ki jih Arches v osnovi ne ponuja:

- podporo procesu objave podatkov, ki zajema celotno obdelavo statusov dokumentov;
- ločevanje lastništva podatkov, ki omogoča urejanje izključno »lastnih« podatkov;
- ločevanje podatkov na zemljevidu z dodatnimi ikonami in barvami glede na tip in material plovil;
- prikazovanje trirazsežnih modelov.

Zbiva in EWD prikazujeta bogate možnosti razširljivosti platforme, saj obe skupini nadgradenj predstavljata dve povsem različni smeri razvoja platforme.

Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres

Simon Krek,[+]* Polona Gantar†, Špela Arhar Holdt[+]†‡, Vojko Gorjanc†

* Laboratorij za umetno inteligenco, Institut »Jožef Stefan«, Jamova 29, 1000 Ljubljana
[+] Center za jezikovne vire in tehnologije, Univerza v Ljubljani, Večna pot 113, 1000 Ljubljana
simon.krek@guest.arnes.si
† Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@ff.uni-ljs.si, vojko.gorjanc@ff.uni-lj.si
‡ Trojina, zavod za uporabno slovenistiko
Partizanska cesta 5, 4220 Škofja Loka
spela.arhar@trojina.si

1 Uvod

Prispevek opisuje projekt[1] nadgradnje korpusa Gigafida ter korpusov Kres, ccGigafida in ccKres. Zadnji trije korpusi bodo iz prvega izpeljani po metodologiji, zastavljeni v projektu Sporazumevanje v slovenskem jeziku (SSJ),[2] v okviru katerega so bili izdelani korpusi prve različice. V drugi različici bo korpus nadgrajen predvsem količinsko, nekatere spremembe pa so predvidene tudi na vsebinski ravni, predvsem s konceptualnim razlikovanjem med deli korpusa, ki po avtorski intenci spadajo v jezikovni standard, ter tistimi, ki segajo izven njega. Na tehnični ravni bo korpus v celoti ponovno jezikoslovno procesiran z nadgrajenimi orodji. V prispevku opišemo izhodiščni projektni načrt in specifikacije, izdelane v prvi fazi projekta.

2 Projektni načrt

Zaradi omejenih sredstev je projekt nadgradnje usmerjen predvsem v tiste segmente gradnje korpusov, ki glede na osnovni namen lahko največ prispevajo h končnemu cilju. Korpusi Gigafida, Kres, ccGigafida in ccKres predstavljajo organsko celoto, zato je o njih smiselno razmišljati kot o seriji korpusov, ki ima izvor v korpusu Gigafida, pri čemer so pri izdelavi izvedenih vzorčenih korpusov (Kres, ccGigafida in ccKres) vedno lahko uporabljeni identični mehanizmi kot v primeru izhodiščne izdelave v okviru projekta SSJ. Ena od pomembnih značilnosti korpusov SSJ so urejena pravna razmerja z besedilodajalci, ki omogočajo javni dostop do vsebine in nadaljnjo distribucijo korpusov pod pogodbeno dogovorjenimi pogoji. V projektu nadgradnje bodo podobne možnosti za javno objavo in nadaljnjo distribucijo korpusov ohranjene, zato se projekt osredotoča na selektivno ciljno zbiranje novih gradiv glede na ugotovljene pomanjkljivosti obstoječih korpusov, ne pa na splošno zbiranje vseh gradiv po načelih, ki so bili uporabljeni pri gradnji korpusa Gigafida in njegovih predhodnikov, korpusov FIDA in FidaPLUS. Drugi segment, ki ga omogoča finančni okvir projekta in lahko prispeva k izboljšanju korpusov, je nova oz. dodatna računalniška obdelava že obstoječih in novih gradiv na podlagi analiz, ki so bile opravljene po izdelavi korpusov. Ker je namen projekta čim širša uporaba nadgrajenih korpusov za različne namene, je vanj vključeno tudi omogočanje javnega dostopa do rezultatov in njihova distribucija po modelu iz projekta SSJ. Projekt nadgradnje korpusov Gigafida, Kres, ccGigafida in ccKres ima torej tri cilje: (a) usmerjeno zbiranje novih gradiv; (b) strojna obdelava novih (in obstoječih) gradiv in (c) javna dostopnost nadgrajenih korpusov, distribucija in diseminacija. V poglavju o specifikacijah opišemo prva dva cilja.

3 Specifikacije

2.1 Zbiranje novih gradiv

Pri izpolnjevanju cilja zbiranja novih gradiv delimo besedila na dva tipa: (a) besedila, za katera je bilo glede na tip, zvrst ali druge kriterije po analizi korpusov Gigafida in Kres ugotovljeno, da so podreprezentirana in (b) besedila izbranih spletnih besedilodajalcev z večjo produkcijo (npr. novičarski portali, dnevni časopisi ipd.), ki zagotavljajo večjo aktualnost korpusnega gradiva.

V prvi kategoriji bodo zbirana predvsem šolska gradiva, tj. (prosto dostopni) učbeniki, delovni zvezki ter sorodna učencem ter dijakom namenjena besedila vseh šolskih predmetov splošnih in poklicnih programov (osnovne šole, gimnazije, poklicne srednje šole). Kot druga večja skupina v to kategorijo spadajo tudi

[1] Projekt financira Ministrstvo za kulturo v letih 2015–2018 v okviru pogodbe št. 33400-15-141007 med ministrstvom in Univerzo v Ljubljani. Izvajalec je Center za jezikovne vire in tehnologije Univerze v Ljubljani (http://www.cjvt.si/).

[2] http://www.slovenscina.eu/.

leposlovna besedila, predvsem tista, ki so glede na podatke o knjižnični izposoji in/ali prodajanosti bolj brana. Med njimi je tudi literatura starejšega izvora, ki pa ima še vedno visoko recepcijo v okviru obveznega šolskega branja. Ta besedila bodo postala integralni del nadgrajenih korpusov Gigafida/Kres/ccGigafida/ccKres 2.0.

V drugi kategoriji bodo zbrana besedila izbranih besedilodajalcev z največjo besedilno produkcijo. Izpostavljeni so predvsem novičarski portali (rtvslo.si, 24ur.com, siol.net, žurnal24.si, sta.si) ter dnevni časopisi (delo.si, dnevnik.si, vecer.si itd.). Iz njih bo sestavljen samostojen podkorpus Novice, ki bo vključen v serijo korpusov Gigafida 2.0, vendar bo zaradi omejene žanrske raznovrstnosti besedil ostal samostojna (pod)enota. Izhodiščno leto objave besedil pri tem podkorpusu je 2010.

Skupna ciljna vsota besed v korpusu Gigafida 2.0 (skupaj s podkorpusom Novice) je 1,5 milijarde besed. Število besed v korpusih Kres, ccGigafida in ccKres po metodologiji iz (Logar in dr., 2012) izhaja iz izhodiščne številke: Kres in ccGigafida 2.0 po 150 milijonov besed, ccKres 15 milijonov besed.

2.2 Strojna obdelava novih in obstoječih gradiv

Jezikoslovno označevanje: od časa označevanja besedil v izvirni seriji korpusov (Grčar in dr. 2012) je bil za slovenščino razvit nov statistični označevalnik. Preliminarni testi so pokazali, da bi natančnejše označevanje lahko dosegli z uporabo metaoznačevalnika, ki upošteva odločitve obeh označevalnikov. Z metaoznačevalnikom, katerega izdelava je predvidena v letu 2016, bo ponovno označen celoten korpus.

Deduplikacija: raba korpusa Gigafida po objavi je pokazala, da bi bilo smiselno proces odstranjevanja dvojnikov izvesti tudi na obstoječih besedilih, saj se v besedilih, ki izhajajo iz tiskanih medijev, pojavljajo ponavljajoči se deli besedil, ki v nekaterih primerih izkrivljajo statistične podatke pri poizvedbah po celotnem korpusu (Logar in dr., 2015). Tipičen primer takih besedil so radijski in televizijski programi, ki so z isto vsebino objavljeni v različnih virih. V procesu priprave serije korpusov Gigafida 2.0 bo deduplikacija izvedena tudi na obstoječih besedilih.

2.3 Jezikovni standard

Trenutno Gigafida in Kres prinašata tako besedila, za katere je glede na okoliščine in medij objave mogoče sklepati, da je bil avtorjev namen pisati v standardnem jeziku, kot besedila, pri katerih takšno sklepanje ni mogoče. Ker je namen nadgradnje oblikovati korpus standardne slovenščine, kot se definira v sociolingvističnih študijah (Cooper 1989; v slovenskem prostoru zlasti Krek 2015; Gorjanc et al. 2015), bo v procesu obdelave obstoječih in novih besedil opravljeno segmentiranje korpusnih dokumentov v tri kategorije. V prvo kategorijo spadajo javno objavljena integralna leposlovna in stvarna besedila, revije, časopisi in podobna besedila (znanstvena besedila, zakonodaja itd.). V drugo kategorijo spadajo predvsem besedila iz prepoznavnih medijev, ki se iz različnih razlogov odločajo za odmik od standarda – najbolj značilni predstavnik kategorije je Novi Matajur, ki je zapisan v regionalni varianti slovenščine. V zadnjo kategorijo uvrščamo predvsem računalniško posredovano komunikacijo, ki je značilna za spletne medije – socialna omrežja, forume ipd. Segmentacija bo opravljena ročno glede na izvor besedila, poleg tega bodo besedila preverjena tudi strojno s prepoznavanjem nestandardnih prvin (Ljubešić et al., 2015).

4 Zaključek

Gigafida, Kres, ccGigafida in ccKres so kot referenčni korpusi za slovenski jezik glavni vir za izvedbo jezikoslovnih raziskav oz. pripravo uporabnojezikoslovnih in jezikovnotehnoloških izdelkov. Na drugi strani sta Gigafida in Kres v vmesnikih SSJ priljubljeno in pogosto uporabljano orodje tudi v širši javnosti, npr. med prevajalci, lektorji, učitelji. Specifični status in vloga teh virov v prostoru zahtevata njihovo kontinuirano nadgrajevanje in opisani projekt odgovarja na nekatere od glavnih identificiranih potreb. S tem predstavlja težko pričakovani korak naprej – čeprav seznam nalog in želja za nadaljnji razvoj ostaja dolg in raznolik.

Literatura

Cooper, Robert L. 1989. Language Planning and Social Change. Cambridge: Cambridge University Press.

Erjavec, Tomaž in Darja Fišer. 2013. Jezik slovenskih tvitov: korpusna raziskava. *Družbena funkcijskost jezika: (vidiki, merila, opredelitve), Obdobja 32*. Ljubljana: Znanstvena založba Filozofske fakultete, 109–116.

Gorjanc, Vojko, Simon Krek in Damjan Popič. 2015. *Med ideologijo knjižnega in standardnega jezika*. Ljubljana: Znanstvena založba Filozofske fakultete. 32–48.

Krek, Simon. 2015. Standardni in knjižni jezik – drugi poskus. Smolej, M. (ur.). *Obdobja 34: Slovnica in slovar – aktualni jezikovni opis*. Ljubljana: Znanstvena založba Filozofske fakultete, 401–407.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. *10th International Conference on Recent Advances in Natural Language Processing*: *Proceedings of RANLP 2015 Conference*, 7–9 September 2015, Hissar, Bulgaria. Hissar: 371–378.

Logar Berginc Nataša, 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, Kaja Dobrovoljc in Špela Arhar Holdt. 2015. Gigafida: interpretacija korpusnih podatkov. V: M. Smolej, ur., *Obdobja 34: Slovnica in slovar - aktualni jezikovni opis*, 2. del, str. 467-477. Ljubljana: Znanstvena založba Filozofske fakultete.

Miha Grčar, Simon Krek in Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. *Zbornik Osme konference Jezikovne tehnologije, 8. do 12. oktober 2012*. Ljubljana: Institut Jožef Stefan.

Digitalizacija in uporaba digitalnega etnomuzikološkega zvočnega gradiva – izkušnje Glasbenonarodopisnega inštituta ZRC SAZU

Drago Kunej

Znanstvenoraziskovalni center SAZU, Glasbenonarodopisni inštitut,
Novi trg 2, 1000 Ljubljana
drago.kunej@zrc-sazu.si

1 Uvod

Preučevanje ljudske glasbe na podlagi zvočnih zapisov ima pred preučevanjem notnih zapisov veliko prednosti, saj omogoča temeljno pri zaznavanju glasbe – poslušanje. Vendar pa tudi zvočni zapis ni popoln znanstveni vir (prim. Schüller, 1992; Schüller, 1994), saj na zaznavanje in dojemanje posnete glasbe močno vplivajo ne samo številni tehnični in metodični dejavniki pri snemanju, ampak tudi akustični, psihološki in sociološki dejavniki pri poslušanju posnetega gradiva.

Digitalizacija arhivskih zvočnih posnetkov pomeni nov izziv pri uporabi in interpretaciji zvočne vsebine. Treba se je namreč zavedati, da se na digitaliziranem gradivu ne odraža le izvirni zvočni zapis, ampak tudi ves metodološki in tehnični postopek digitalizacije, tako vsa uporabljena oprema, njeno stanje in nastavitve kakor tudi morebitno urejanje, »čiščenje«, »popravljanje« in drugo poseganje v zvočni zapis. Digitaliziran posnetek sam po sebi še ni enoumen in jasen vir zvočnih informacij, ampak dobi svojo pravo vrednost šele s poznavanjem snemalnih okoliščin in postopka digitalizacije ter potrebno spremno dokumentacijo.

2 Namen članka

Danes je večina zgodovinskega etnomozikološkega zvočnega gradiva, ki je bilo zapisano na mehanskih nosilcih, dostopno le v digitalni obliki. Zaradi skromnih tehničnih možnosti prvih snemalnih naprav ter današnjega pomanjkanja poznavanja nekdanje tehnologije in znanja pri predvajanju in interpretaciji zgodnjih zvočnih posnetkov, je lahko takšno digitalizirano zvočno gradivo pomanjkljiv in celo zavajajoč znanstveni vir.

Prispevek bo predstavil izkušnje pri digitalizaciji in uporabi digitalnega etnomuzikološkega zvočnega gradiva iz arhiva Glasbenonarodopisnega inštituta ZRC SAZU ter na primeru zgodnjih etnomuzikoloških posnetkov podal temeljne napotke pri uporabi digitalnih zvočnih posnetkov kot znanstvenih virov za etnomuzikološke in folkloristične raziskave. Prav tako bo na podlagi izkušenj opozoril na nekatere možne napačne interpretacije digitalizirane zvočne vsebine, do katerih lahko pride zaradi pomanjkanja podatkov in znanja v postopku digitalizacije, ki jih bo ponazoril z zvočnimi in s slikovnimi primeri.

3 Inštitutske izkušnje

V Zvočnem arhivu Glasbenonarodopisnega inštituta ZRC SAZU v Ljubljani se že več kot dve desetletji ukvarjamo z zaščito, digitalizacijo in arhiviranjem zvočnih zapisov, s poudarkom na terenskem in zgodovinskem zvočnem gradivu, ki služi kot znanstveni vir različnim raziskovalcem. V postopku digitalizacije zvočnega gradiva kot znanstvenega vira poskušamo zajeti in ohraniti čim več zvočno zapisanih informacij, pri čemer nas vodijo strokovni in ne estetski kriteriji. Takšna digitalizacija je zahtevna, saj ob tem poleg ustrezne opreme potrebujemo tudi veliko znanja o tem, kako predvajati specifične nosilce, da se bo reproducirana zvočna slika čim bolj približala izvajani in posneti. Poleg tega je zelo pomembno, da pred digitaliziranjem zberemo vso že obstoječo spremno dokumentacijo (metapodatke) in morebitne dodatne vire o posnetkih, v procesu digitalizacije in arhiviranja pa ne posegamo v zvočni signal ter dobro dokumentiramo celoten postopek digitaliziranja in uporabljeno opremo z njenimi nastavitvami (parametri presnemavanja). S tem opredelimo digitalni presnetek in omogočimo raziskovalcem boljše razumevanje slišanega. Le skupaj z vsemi metapodatki lahko digitalni zvočni presnetki postanejo dober vir tako v sedanjih kakor tudi v prihodnjih raziskavah.

3.1 Hitrost predvajanja (frekvenca vrtenja) mehanskih nosilcev zvoka

Eden najpomembnejših tehničnih parametrov pri digitalizaciji mehanskih nosilcev zvoka je frekvenca vrtenja nosilca. Osnovno pravilo pri določanju hitrosti predvajanja mehanskih nosilcev določa, da se nosilci predvajajo z isto hitrostjo, kot so bili posneti. Le tako je namreč mogoče zagotoviti enak zvočni učinek posnetka, kot ga je imel izviren zvok (prim. Bradley, 2009). Vendar v zgodnjem obdobju zvočnih snemanj še ni bilo dogovorjenih norm in standardov snemanja, zato so se snemalci odločali predvsem na podlagi intuicije, priporočil in lastnih izkušenj ter v skladu tehničnimi možnostmi uporabljenih naprav. Tudi potem, ko se pri studijskem snemanju fonografskih valjev uveljavila standardizirano frekvenca vrtenja 160 obratov na minuto (prim. Sage, 2005), jih mnogi v glasbeni industriji, predvsem pa raziskovalci na terenu, niso upoštevali. Zato ni nenavadno, da so bili valji posneti s hitrostmi od 80 pa vse do 250 o/min (Wiedmann, 2000). Takšen razpon snemalnih hitrosti pomeni, npr. pri izbiri navedenih skrajnih vrednosti, spremembo hitrosti posnetka in njegovega trajanja za več kot trikrat, poleg tega pa se spremeni intonacija posnetega za več kot oktavo in pol.

Hitrost predvajanja nosilca pri digitalizaciji opredeljuje mnoge dejavnike, ki določajo zvočno sliko ter odločilno vpliva na zaznavanje in dojemanje slišanega: predvsem na tempo izvajanja glasbe in hitrost govora, intonacijo posnetega, zven oz. zvočno barvo posnetka, način podajanja izvajalca (npr. glasbenih okraskov, dramatičnost govora) idr. Vpliva tudi na subjektivno zaznavo posnetega gradiva, kot je razumljivost govora in besedil pesmi, občutek »naravnega« zvena glasbil in glasov, estetsko dojemanje posnete glasbe in govora, na interpretacijo vsebine in izvajalce, kakor tudi na zaznavanje različnih tehničnih motenj, kot sta raven in barva šuma.

3.2 Interpretacija digitaliziranih zgodnji zvočni posnetki

Raziskovalci so kmalu spoznali, da dokumentarni zvočni posnetki ne prinašajo samega zvoka, ampak dejansko »interpretacijo zvoka« (Seeger, 1986), na katero močno vplivajo metodološke in tehnične okoliščine zvočnega zapisa. Zato smo na GNI izvedli raziskavo, s katero smo želeli ugotoviti, kako poslušalci zaznavajo in interpretirajo nekatere značilnosti digitaliziranega zvočnega gradiva s starih zvočnih nosilcev. Pripravili smo vprašalnik, pri katerem so anketiranci pri vsakem vprašanju poslušali različne zvočne primere (vzorce), jih primerjali med sabo in odgovarjali na zastavljena vprašanja. Zvočni primeri pri posameznem vprašanju so bili praviloma del istega posnetka, ki pa je bil digitaliziran na različne načine. V anketi smo jih spraševali po različnih značilnostih, ki bi jih bilo s posnetkov moč zaznati in interpretirati, kot sta npr. tudi spol in starost pevcev. Rezultati so pokazali, da na različne načine digitalizirani posnetki pri poslušalcih sprožajo možnosti spekulativne interpretacije slišanega, kar lahko privede do napačnih hipotez in zaključkov.

4 Literatura

Kevin Bradley, ur. 2009. *Guidelines on the production and preservation of digital audio objects. IASA Technical Committee - Standards, Recommended Practices and Strategies, IASA-TC 04,* second edition. Canberra: International Association of Sound and Audiovisual Archives (IASA).

Glenn Sage. 2005. *Early Recorded Sounds and Wax Cylinders*. http://www.tinfoil.com/earlywax.htm.

Dietrich Schüller. 1992. Phonographische Dokumentationsmethoden in der Ethnomusicologie. V: Wolfgang Lipp, ur., *Gesselschaft und Musik. Wege zur Musiksoziologie*, str. 505–517. Berlin: Dunker & Humblot.

Dietrich Schüller. 1994. Mikrophonverfahren für ethnomuzikologische Schallaufnahmen. V: Elisabeth Th. Hilscher in Theophil Antonicek, ur., *Vergleichend-systematische Musikwissenschaft. Beiträge zur Methode und Problematik der systematischen, ethnologischen und historischen Musikforschung. Franz Födermayr zum 60. Geburtstag*, str. 119–144. Tutzing: Hans Schneider.

Anthony Seeger. 1986. The Role of Sound Archives in Ethnomusicology Today. *Ethnomusicology*, 30(2):261–276.

Albrecht Wiedmann. 2000. A few Technical Remarks on the Digital Conservation of the Old Inventory of the Berlin Phonogramm-Archive. V: Artur Simon, ur., *The Berlin Phonogramm-Archive 1900–2000. Collections of Traditional Music of the World,* str. 203–208. Berlin: VWB – Verlag fűr Wiessenschaft und Bildung.

Trirazsežno dokumentiranje v službi varovanja nepremične kulturne dediščine

Aleš Lazar‡, Sonja Ifko*

‡ Magelan skupina d.o.o., Glavni trg 13, Kranj
* Fakulteta za arhitekturo, Univerza v Ljubljani, Zoisova 12, Ljubljana

1 Uvod

Nepremična kulturna dediščina je tisti del prostora, ki uokvirja in opredeljuje naše razvojno zgodovinske značilnosti, kulturno samobitnost in je ključen gradnik identitete tako družbe kot vsakega posameznika. Zato je pomembno, da z dediščino odgovorno ravnamo in jo ustrezno varujemo. Njena učinkovita identifikacija in dokumentiranje sta ključna elementa varstva. V prispevku želimo opozoriti na dokumentiranje vsaj najpomembnejše dediščine, ki se mu v naši državi žal ne posvečamo sistematično in celostno.

Dejstvo je namreč, da se arhitekturna in arheološka kulturna dediščina hitreje izgublja kot se dokumentira. Glavna dejavnika za to sta človeške narave, in to sta vojna in »nekontroliran« razvoj. Če temu procesu dodamo še naravne nesreče, zanemarjanje in neprimerno skrb, dobimo razlog, zakaj izgubljamo tako pomemben potencial, kot je kulturne dediščina. Danes obstajajo metode in tehnike, ki omogočajo celovito trirazsežno (3D) dokumentiranje objektov kulturne dediščine s pomočjo učinkovitih digitalnih orodij.

V prispevku bodo predstavljeni pristopi, s pomočjo katerih lahko zagotovimo učinkovito dokumentiranje, ki je podlaga za odločanje pri izvjanju varstvenih ukrepov ter predstavlja hkrati pomemben arhivski dokument stanja določenega spomenika, oziroma enote dediščine v času, ko je bil posnetek izveden.

2 3D dokumentiranje

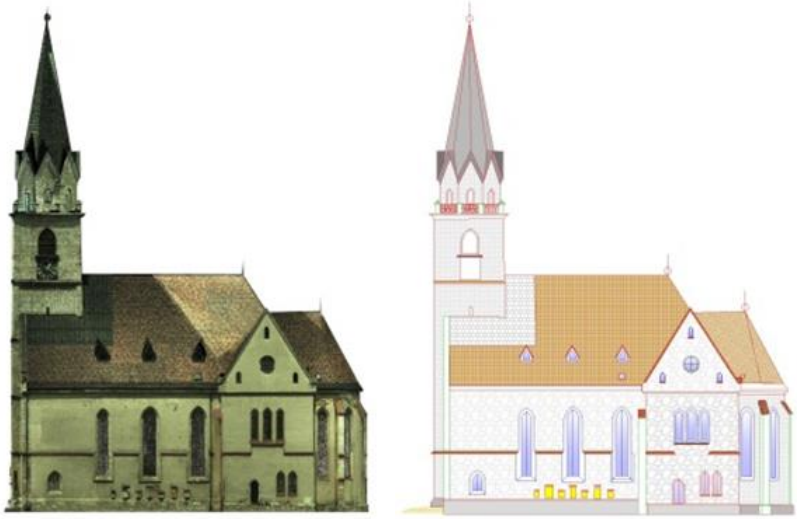
Integrirana metoda 3D laserskega skeniranja in fotogrametrije je uveljavljena kot najkvalitetnejša metoda zajema prostorskih podatkov za arhitekturne objekte, saj zagotavlja celovit pristop za tehnično dokumentiranje, analizo površinskih deformacij na objektih ter možnosti atraktivnih vizualizacij, animacij. Gre za metodo, ki združuje prednosti obeh merskih metod. Terestrično 3D lasersko skeniranje (TLS) zagotavlja hitre, natančne in celovite 3D meritve na varen, brezkontakten in neinvaziven način. Fotogrametrija, znanost in tehnologija pridobivanja kvantitativnih informacij iz fotografskih posnetkov, omogoča vključitev teksture na 3D modele prostorskih podatkov.

Osnovni produkt integrirane uporabe 3D laserskega skeniranja in fotogrametrije je fotorealističen oblak točk. **Fotorealističen oblak točk** (tudi **fotorealističen točkovni 3D model**) predstavlja temeljni 3D digitalni arhiv, ki je uporaben tako za izdelavo klasičnih izdelkov tehnične dokumentacije kot za sodobne 3D prikaze in analize. Terestrični 3D skener sistematično in z visoko mero natančnosti izmeri vse kar vidi človeško oko. Gostota točk zajema podatkov sega do milimetrskih razponov in prav tako razpoznavnost posameznih detajlov.

Slika 1: Fotorealističen oblak točk Predjamskega gradu (desno) in oblak točk obarvan z vrednostjo intenzitete odboja laserskega žarka (levo).

3 Rezultati

Fotorealističen oblak točk predstavlja digitalno, računalniško oz. virtualno repliko objekta kulturne dediščine v naravi. Iz teh podatkov se lahko izdela 2D in 3D tehnična dokumentacija, ki obsega situacijske načrte, topografske načrte, ortofoto načrte, karakteristične profile oz. prereze, digitalni model reliefa, 3D CAD modele... Na podlagi natančnega zapisa površin se lahko kartirajo poškodbe na objektu, izdela stavbna analiza ipd. V primeru meritev istega objekta v različnih terminih, se lahko izvaja monitoring, temeljita analiza spremembe strukturnih poškodb in deformacij na površini.



Slika 2: Ortofoto (levo) in CAD izris posameznih detajlov (desno) cerkve Sv. Kancijana v Kranju.

4 Diskusija in zaključek

Dokumentacija je izhodišče za razumevanje in opredelitev pomena kulturne dediščin. Omogoča natančno pripravo grafičnih podlag za nadaljnje delo. Prednosti natančnega in celovitega 3D dokumentiranje objektov stavbne dediščine so uporabne tako v znanosti in stroki kot tudi v smislu popularizacije objekta za širšo množico, saj se lahko s pomočjo fotorealističnega oblaka točk izdelajo atraktivne vizualizacije, animacije, 3D in 4D interaktivne aplikacije ipd.

Iz predstavljenega sledi, da je pomembno tako učinkovita orodja čim širše uporabljati in da je nujno v državi vzpostaviti sistematičen sistem digitalnega dokumentiranja stanja objektov kulturne dediščine,

saj bi tako vzpostavili kvalitetno osnovo za varstvo dediščine kot enega ključnih nacionalnih razvojnih potencialov.

5 Literatura

David Andrews, Jon Bedford, Bill Blake, Paul Bryan, Tom Cromwell, Richard Lea. 2010. *Measured and Drawn: Techniques and practice for the metric survey of historic buildings. Second edition.* London, English Heritage, Kemble Drive, Swindon. https://www.historicengland.org.uk/images-books/publications/measured-and-drawn/.

Aleš Lazar. 2012. *Sodobno tehnično dokumentiranje grajske arhitekture na primeru gradu Lož*. Diplomska naloga, UL FGG. http://drugg.fgg.uni-lj.si/4099/1/GEV914_Lazar.pdf.

Robin Letellier. 2007. *Recording, Documetation, and Information Management for the Conservation of Heritage Places*. The Getty Conservation Institute. Los Angeles, ZDA. http://www.getty.edu/conservation/publications_resources/pdf_publications/recordim.html.

John Mills, David Barber. 2006. *An addendum to the metric survey specifications for english heritage - the collection and archiving of point cloud data obtained by terrestrial laser scanning or other methods.* London, English Heritage.

Analysing Spatial Distribution of Linguistic Variables in Geoencoded Tweets from Croatia, Bosnia, Montenegro and Serbia

Nikola Ljubešić,[*] Tanja Samardžić,[†] Maja Miličević[‡]

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

[†]CorpusLab
University of Zurich
Plattenstrasse 54, CH-8032 Zurich
tanja.samardzic@uzh.ch

[‡] Faculty of Philology
University of Belgrade
Studentski trg 3, SR-11000 Belgrade
m.milicevic@fil.bg.ac.rs

1. Introduction

Twitter has recently become a popular medium for spatial analysis of language (Doyle, 2014; Jørgensen, et al, 2015), given that (1) it has an open, high-quality API, and (2) a small, but significant percentage of the messages it contains is geoencoded.

In this demo we will present a tool which is currently under development (https://github.com/scopes-reldi/geotweet) that enables researchers interested in spatial variation of language to define a geographic perimeter of interest, collect data from the Twitter streaming API published in that perimeter, filter the obtained data by language and location, define and extract variables of interest, and analyse the extracted variables by one spatial statistic and two spatial visualisations.

We will demonstrate the tool on the area and a selection of languages spoken in former Yugoslavia. By defining the perimeter, the languages and a series of linguistic variables of interest we will illustrate the tool's data collection, processing and analysis capabilities. The linguistic variables we focus on are those that are known to vary in our area of interest, more specifically across the highly similar languages of Croatian, Bosnian, Montenegrin and Serbian.

The only previous work on Twitter data for the linguistic areas mentioned above that we are aware of is Ljubešić and Kranjčić (2015), where the focus is on discriminating between the languages. Our intention in this paper is to address a new topic, namely, the spatial distribution of specific linguistic phenomena often recognised as characteristics of the languages / varieties in this area. The two main goals of our research activities in this domain are (1) further development of the methodology for analysing linguistic variation via geoencoded social media, and (2) a comparison of actual data with the distributions expected on the basis of the literature, or widely accepted beliefs.

The remainder of this abstract gives an overview of the tool functionality.

2. Data Collection

In order to start the data collection, the user needs to enter his/her Twitter API credentials (obtained from the Twitter Developer site) and define the geographic perimeter of interest. Once started, the data collection component communicates with the Public Twitter Streaming API and stores the messages satisfying the perimeter criterion into a relational database.

3. Data Processing

There are two main functionalities of the data processing module: data filtering and variable extraction.

3.1. Data Filtering

Currently there are three user filtering techniques implemented in the tool: filtering by the minimum number of posts collected from a user, filtering by the country in which most of a user's tweets were posted, and filtering by the language of the majority of a user's tweets.

3.2. Variable Extraction

There are three main mechanisms for defining variables to be extracted. The first one enables extracting metadata from the Status objects, e.g. number of retweets, posting time, whether the tweet is a reply to another tweet etc. The second mechanism enables defining a variable based on a lexicon / token list, while the third one allows the user to define the variable via regular expressions. During the demo session each of the mechanisms will be showcased on a series of example variables relevant for the languages in question.

The final result of the data processing module is a simple tab-delimited file that can be used either in the data analysis module, which is described next, or in some other tool chosen by the user.

4. Data Analysis

The analysis module consists of three functionalities: point visualisation, spatial trend detection, and the identification of dominant regions per variable level.

4.1. Point Visualisation

The point visualisation functionality allows the user to gain an initial impression of the spatial distribution of all levels of a linguistic feature. Each tweet containing a value for the inspected variable is represented on a map as a point, with the value of the variable level encoded by colour. The text of the tweet can be obtained by clicking on the point.

4.2. Spatial Trend Detection

The spatial trend detection functionality comprises a measure that quantifies the spatial dependency in the data, often referred to as spatial autocorrelation. We compare the spatial distances as computed between all tweets of one linguistic feature (expected distances) with the distances as calculated for each feature level separately (observed distances). Aggregating these two sets of distances into what we call a relative distance measure allows us to distinguish feature levels that are spatially clustered (observed distance < expected distance) from levels that are scattered in space (observed distance > expected distance).

4.3. Dominance Maps

Dominance maps visualise the dominant levels of a variable throughout a map. They are particularly useful when many measurements are available and points start to overlap, making by-point visualisation hard to decipher.

All three analysis functionalities will be showcased during the demonstration on variables extracted with the data processing module.

5. References

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 98–106, Gothenburg, Sweden, April. Association for Computational Linguis- tics.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In Proceedings of the Workshop on Noisy User- generated Text, pages 9–18, Beijing, China, July. Association for Computational Linguistics.

Nikola Ljubešić and Denis Kranjčić. 2015. Discriminating between Closely Related Languages on Twitter. Informatica, 39(1):1–8.

The First World War on the Web - The Case of Serbia

Slobodan Mandić*
*Historical Archives of Belgrade
Palmira Toljatija 1, Belgrade, Serbia
slobodan.mandic@gmail.com

1. Introduction

Even though the implementation of computers in the historical sciences began in the 60s and 70s of the 20th century, together with the development of new methods in the studies of economic history and demography, and again in the 80s with the appearance of personal computers, only with the expansion of Internet and digital revolution in the 90s, great number of historians started to use modern technologies. Internet, as the world's biggest global informational resource and unique way of communication in real time, has offered until then unbelievable possibilities in the field of research activities, but also introduced numerous problems concerning the use of Internet information in historical science. In the moment when 3, 5 billion people have access to the Internet, celebration of the World War I, the first war in historiography considered to be global and total, represents a challenge and raises important questions for the historical science.[1] In the previous period numerous projects on the WWI have been initiated and implemented worldwide, with a goal to connect experts from various fields, to facilitate the access to historical sources and archival material, to present results of researches, to organize virtual exhibitions, to advance teaching process and education, all this by using new technologies and media.

2. Goal of the Paper

This paper aims to study phenomena on using new technologies in historical science based on analysis of form, structure and content of several most representative Web presentations, and most of all, to study the development in perception of historical sources, archival material and the representation of the results of the scientific researches. Also, within the global picture, example of Serbia and Serbian cyber space is analyzed separately, since Serbia has dedicated a lot of attention and space in media, science and culture to celebrate WWI.

3. Digital History from the Perspective of the First World War on the Web

A wide range of international projects includes large Internet sites, sites like International Encyclopedia of the First World War, made as a result of a common project of more than a thousand experts from more than fifty countries, or Europeana 1914–1918 project, includes projects of individual institutions like internet presentations of Austrian State Archives and Historical Archives of Belgrade and individual historians' presentation (i.e. Otto Vervaart' blog) and nongovernmental organizations (i.e. Remembering the ancestors from the First World War album of the Association of the descendants of Serbian warriors 1912–1920). This material today easily accessible for the users from all around the world illustrates in the correct way changes and challenges modern generations of historians are faced with. It introduces numerous heuristic possibilities that can be used to solve one of the biggest problems of a historian related to the representativeness of a used material, when the research results cannot be generalized but have to be limited to a certain area and period, because of the lack of sources. However, it should be underlined that this opens series of methodological problems and questions about use of information and historical sources taken from the Internet, like

[1] Internet penetration in Serbia is 54%, 4.758.861 Internet users. Internet Live Status. 2016. Internet Users by Country (2016) <http://www.internetlivestats.com/internet-users-by-country> (8. June 2016).

partial digitalization of fonds and collections or fetishism of the documents, often without any selection in reconstructing of the past (Mandic, 2008; Jovanovic 2009)[2].

Numerous publications (monographs, reprints, proceedings, exhibition catalogues) relating to the First World War have been published in Serbia in the past few years, many exhibitions, conferences, public manifestations and television shows have been organized and recorded, but only several Web presentation have been released, which allowed full perception of the historical events that are of great importance in the collective memory. Although these projects have different organizational and institutional starting points (being part of a greater international projects, within individual institutions of culture or within the NGO sector), it can be concluded that several important projects have been realized significantly enriching the possibilities of researches, studies and work in digital environment. Experiences and results achieved in the projects of digital collections of National Library of Serbia and Yugoslav Film Archives, online digital thematic guide of the Historical Archives of Belgrade (in Serbian and English language), Remembering the ancestors from the First World War album of the Association of the descendants of Serbian warriors 1912–1920 and presentations focusing on a specific chronological and geographical point (Valjevo – The Hospital City 1914–1915, Belgrade in the Great War, On the Streets of the Sava area in 1914), made with the intention to celebrate the centenary of the First World War, will also serve as a starting point to redefine many aspects of digital humanities in the following period. It can be already concluded that more organized aspect of intersectoral cooperation of cultural institutions would have greater effect in the country and on international scene.

It is interesting the fact that most of the Web presentations on the First World War in Serbia are not supported by corresponding Web 2.0 tools, first of all in the most popular social networks, with a goal to promote and attract wider audience. Except for two examples, Historical Archives of Belgrade's which promoted Online digital Thematic Guide on the Archives' official Facebook and Flickr account and project Remembering the ancestors from the First World War album of the Association of the descendants of Serbian warriors 1912–1920 also used Facebook to promote its activities, it is evident that all other projects did not take advantage of this kind of communications and large number of users.

One of the basic characteristics of projects with Internet presentations concerning the First World War, that will eventually turn out to be a problem at the same time, is cooperation of different experts. Interdisciplinary approach implies cooperation between historians and computer scientist, librarians, archivists, museologists, resulting in new interdisciplinary science – digital humanities. Global character of the First World War is determined by the fact that countries from all continents participated in the conflict, by the fact that WWI was also an industrial war and that entire societies were engaged in the background of the frontline, introducing massive changes in everyday life. Numerous Web presentations and large amount of material on the World Wide Web represent potentially a good material for retrospection of relations between historical science and challenges of new technologies' implementation today and in the future. Historian William Thomas (2014) indicated the necessity to reconstruct historical science for digital era and to introduce more complex attitude towards social reality encountering and arising from total and complex social reality of the

[2] Definition of the criteria for selection of the documents intended for digitalization, their evaluation and selection of the priorities, that is partial digitalization of the certain fonds and collections, can lead to the unbalanced impressions of those who are using specific material or even to the creation of inappropriate attitude of those less familiar with the material and with that historical period. It is extremely important to underline this problem of using digitalized archival material, both in researches and in wider use, especially in history teaching process. See: S. Mandić. 2008. Kompjuterizacija i istoriografija 1995-2005. Beograd, str. 42, 87. The problem of document fetishism was discussed by professor Miroslav Jovanovic, who stated that in the problem of „document inflation " historians are at the same time accomplices and victims, and emphasized that each decade of the 20th century produced more information than the human civilization in six previous millenniums. Traditional positivism's tendency towards fetishism of the documents in most cases does not reconstruct even the simplest facts in the past, but leads towards plain rewriting and accumulation of the documents. This is especially important in the situation when a large amount of documents and information on archival material becomes available in digital form, easily available on Internet.
See: Мирослав Јовановић, Радивој Радић. 2009. *Криза историје. Српска историографија и друштвени изазови краја 20. и почетка 21. века*, Београд, стр. 31-32, 67, 75-81.

past. He also pointed out that small number of reviews and critiques in the field of digital history existed and that larger interpretation and implementation of results of already finalized projects were necessary, because thematic archives and sophisticated projects of digitalization are often unnoticed, unquoted and uninvolved in the scientific streams. Besides this, it is important to motivate and involve wider audience in the field of digital history, in the segments that potentially have material for family history and genealogy research.

Table: Website locales statistic for *Project First World War in Fonds and Collections of the Historical Archives Of Belgrade – Online Digital Thematic Guide*. February 2016.

Locales	Domain	Pages	Hits	Bandwidth
Republic of Serbia	sr	6,079	54,674	4.15 GB
Germany	de	272	1,582	164.32 MB
Bulgaria	bg	242	2,097	286.47 MB
Bosnia-Herzegovina	ba	194	1,307	157.53 MB
United States	us	187	833	89.63 MB
Croatia	hr	153	946	107.95 MB
Austria	at	148	1,163	149.75 MB
Hungary	hu	66	396	51.58 MB
Slovenia	si	66	387	36.85 MB
Great Britain	gb	64	430	36.22 MB
France	fr	61	183	20.09 MB
Japan	jp	43	117	2.54 MB
Netherlands	nl	42	227	28.29 MB
Italy	it	40	286	35.03 MB
Romania	ro	34	193	24.20 MB
Poland	pl	30	160	20.19 MB
Albania	al	26	286	44.91 MB
Macedonia	mk	24	212	18.27 MB
Australia	au	22	131	12.99 MB
Belgium	be	19	162	15.23 MB
Spain	es	16	120	7.06 MB
Turkey	tr	15	169	22.53 MB
Switzerland	ch	14	105	11.63 MB
Czech Republic	cz	11	111	14.32 MB
Montenegro	me	11	42	1.69 MB
Others		54	483	55.38 MB

4. References

Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, Jeffrey Schnapp. 2012. *Digital_Humanities*. Massachusetts Institute of Technology.

Мирослав Јовановић, Радивој Радић. 2009. *Криза историје. Српска историографија и друштвени изазови краја 20. и почетка 21. века*. Удружење за друштвену историју, Београд.

Slobodan Mandić. 2008. *Kompjuterizacija i istoriografija 1995-2005*. Istorijski arhiv Beograda, Beograd. http://www.udi.rs/articles/mjovanovic_Kriza%20istorije.pdf.

Слободан Мандић. 2010. Архиви и "Web 2.0" окружење. Историјска баштина. 271-278.

Slobodan Mandić. 2013. Intersektorska saradnja baštinskih institucija kulture – neiskorišćeni potencijal veba. Читалиште. 22: 30-33.

Lawrence J. McCrank, 2001. *Historical Information Science: An Emerging Unidiscipline*. Medford, New Jersey, Information Today.

Мирослав Перишић, Александар Марковић, Љубинка Шкодрић, Бранко Богдановић. 2015. *Први светски рат у документима Архива Србије*, Том I, 1914. Архив Србије, Београд.

4.1. Internet

Албум сећања на наше претке из Првог светског рата@albumsecanja. *Facebook.* <https://www.facebook.com/albumsecanja/?fref=ts> (20. IV 2016).

Милош Брун, Немања Калезић. *Београд у Великом рату.* <http://www.beogradvelikirat.org> (12. V 2016).

Freie Universität Berlin. *1914-1918-online. International Encyclopedia of the First World War*. <http://www.1914-1918-online.net> (15. IV 2016).

Historical Archives of Belgrade. 2015. *Project First World War in Fonds and Collections of the Historical Archives Of Belgrade – Online Digital Thematic Guide*. <http://ww1.arhiv-beograda.org/?language=eng&navmenu=1> (18. IV 2016).

Historical Archives of Belgrade. *Flickr*. <https://www.flickr.com/photos/125431192@N07/> (20. IV 2016).

Hrvatski državni arhiv. *Prvi svjetski rat 1914.-1918. – pogled iz arhiva*. <http://prvisvjetskirat.arhiv.hr> (15. V 2016).

International Committee of the Red Cross Historical Archives. *Prisoners of the First World War*, <http://grandeguerre.icrc.org> (2. XI 2015).

Istorijski arhiv Beograda. *Facebook*. <https://www.facebook.com/Istorijski-arhiv-Beograda-180908415273554/> (20. IV 2016).

Jugoslovenska kinoteka. *EFG1914 projekt*. <http://www.kinoteka.org.rs/di/efg/Http/EFG/00-efg_S_index01.htm> (7. VIII 2015).

С.Н. Камышев, А.А. Мелитонян, В.В. Петраков. 2014. *1914-1918, Великая война, документы и фотографии Первой мировой войны*. Российское военно-историческое общество, «Народный архив». <http://pomnimvseh.histrf.ru> (12. V 2016).

Verena Moritz. *Forschungsprojekt: Kriegsgefangene in Österreich(-Ungarn) 1914-1918 Zwangsarbeit und Gewalt*. <http://www.pows-ww1.at/> (18. II 2016).

Ministry of Culture and Information, Republic of Serbia. *Serbia Remembers 1914 – 2014: Program of Marking First World War Jubilee (1914-1918).* <http://www.srbijapamti.rs/eng> (4. VI 2016).

The National Archives (UK). *First World War - First World War portalFirst World War portal*. <http://www.nationalarchives.gov.uk/first-world-war> (22. V 2016).

National Library of Serbia. *The Great War.* <http://velikirat.nb.rs/en/> (18. I 2016).

The National Museum of Valjevo. *Valjevo-City-Hospital (1914-1915).* <http://valjevo-hospital.org/index.html> (18. I 2016).

Österreichisches Staatsarchiv. *1914-2014 - 100 Jahre erster Weltkrieg*. <http://wk1.staatsarchiv.at> (15. IV 2016).

Radio televizija Srbije. *Specijal o Velikom ratu*. <http://www.rts.rs/page/stories/ci/Velikirat.html> (18. I 2016).

Савез потомака ратника Србије 1912–1920. године, *Албум сећања на наше претке из Првог светског рата*, <http://славним-прецима.срб> (17. I 2016).

GO Savski venac, Tačka komunikacije. *Walk of the Century - The Streets of Belgrade's Sava Neighbourhood in 1914*. <http://setnjaveka.rs> (18. I 2016).

Союз возрождения родословных традиций – Геральдика. Первая мировая война. <http://www.svrt.ru/1914/1914.htm> (12. V 2016).

William Thomas. 2014. *The Future of Digital History*. November 16, 2014. <http://railroads.unl.edu/blog/?p=1146> (19. IV 2016).

Otto Vervaart. *digital 1418 - Digital projects around the First World War*. <https://digital1418.wordpress.com> (19. IV 2016).

Od znanstvenokritične izdaje do repozitorija rokopisov

Dosežki, možnosti in načrti

¸Matija Ogrin,* Tomaž Erjavec,# Jan Jona Javoršek‡

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

1 Pregled dela

Od začetka prvega slovenskega raziskovalnega projekta o znanstvenih izdajah v elektronskem mediju v letu 2001 je minilo 15 let. V tem času se je zvrstila vrsta iniciativ in projektov, v katerih smo kolegi z ZRC SAZU in IJS razvijali in aplicirali postopke elektronske obdelave, analize, prikaza in objave besedil. S temi postopki, vedno so temeljili na smernicah konzorcija Text Encoding Initiative (TEI), smo obdelovali razne vrste humanističnih tekstov s področja več ved, zlasti literarne zgodovine oz. širše literarne vede, občega zgodovinopisja in celo muzikologije. Ta besedila so v prostem dostopu z licencami CC objavljena na spletu v elektronskih znanstvenih izdajah, ki jih sorodne konceptualne osnove povezujejo v zbirke eZISS (Elektronske znanstvenokritične izdaje slovenskega slovstva), portal eZMono (Elektronske znanstvene monografije) z zbirkami SGD (Slovenska glasbena dediščina), SD18 (Osemnajsto stoletje na Slovenskem), v razvoju je zbirka eZD (Elektronska Zbrana dela slovenskih pesnikov in pisateljev), med osrednjimi dosežki vseh teh prizadevanj pa je portal NRSS (Neznani rokopisi slovenskega slovstva 17. in 18. stoletja).

2 Kritične ugotovitve

Prispevek predstavi pregled nad raznolikimi oblikami elektronskih izdaj in zbirk z omenjenih področij zlasti s treh gledišč:

1. kaj se po 15 letih uporabe in razvoja elektronskih metod zdi še danes uporabno, kaj pa se kaže kot zastranitev ali slepa ulica;
2. uporabne in perspektivne elektronske izdaje je treba spričo tehnoloških sprememb prenoviti, jih prenesti v novo tehnološko okolje;
3. kaj se je za področje literarnih in/ali historičnih ved izkazalo za posebej uporabno, a za to še ni razvita vsa zaželena tehnološka podpora.

S prvo točko želimo nakazati aktualnost in koristnost elektronskega označevanja in prikaza besedil, zlasti za namene znanstvenokritičnih izdaj temeljnih virov (eZISS). Na drugi strani pa se poskusi povezovanja elektronskega besedila z multimedijskim okoljem zdijo vsaj za zdaj tehnološko brez prave perspektive.

S točko dve želimo nakazati, da je tip elektronske znanstvenokritične izdaje, kot se je z raznimi manjšimi razlikami razvil v zbirki eZISS, trajno uporaben in za humanistične vede razvojno perspektiven. Prispevek nakaže, kako bomo obstoječe izdaje eZISS prenesli v novo tehnološko okolje (Fedora Commons), kjer že deluje vrsta naših zbirk, zlasti repozitorij rokopisov NRSS.

V tretji točki bo izpostavljeno, da je ustroj repozitorija, kakršen je NRSS, sicer uporaben za analitične opise primarnih virov, v našem primeru rokopisov, in za njihove digitalne faksimile, kar se v podobnih oblikah uporablja široko po svetu. Vendar bo repozitorij še mnogo bolj uporaben, ko mu bomo dodali načrtovani modul s prepisi celotnih rokopisnih besedil s primernimi iskalnimi možnostmi. Sočasno razvijamo z metodami strojno podprtega prevajanja tudi programsko rešitev, ki bo diplomatično prepisana rokopisna besedila do določene stopnje predelala v kritično redigirane tekste z minimalnimi, toda računalniško zahtevnimi posegi, kakor je denimo pretvorba (pogosto nestandardne) bohoričice v gajico. Z vsem tem bo širši javnosti bistveno olajšan dostop do vsebine starejših slovenskih rokopisnih besedil, do njihove tako dokumentarno-podatkovne kakor literarno-estetske in duhovne dimenzije.

3 Literatura

Tomaž Erjavec, Matija Ogrin. 2005. Digital critical editions of Slovenian literature: an application of collaborative work using open standards. V: DOBREVA, Milena (ur.), ENGELEN, Jan (ur.). *From author to reader : challenges for the digital content chain : proceedings of the 9th ICCC International Conference on Electronic Publishing,* Arenberg Castle, June 8-10, 2005. Leuven: Peeters, str. 151–156.

Matija Ogrin, Jan Jona Javoršek, Tomaž Erjavec. 2013. A register of early modern Slovenian manuscripts. *Journal of the Text Encoding Initiative,* ISSN 2162-5603, March 2013, issue 4, str. 1-13, ilustr. http://jtei.revues.org/715.

TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Version number]. [Last modified date]. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

The Use of Semantic Word Classes in Document Classification

Stevan Ostrogonac,*† Branislav Popović,*† Milan Sečujski,*†

* Faculty of Technical Sciences, University of Novi Sad
Trg Dositeja Obradovića 6, 21000 Novi Sad

†AlfaNum – Speech Technologies Ltd,
Polgar Andraša 38a/61, 21000 Novi Sad
ostrogonac.stevan@uns.ac.rs
bpopovic@uns.ac.rs
secujski@uns.ac.rs

1. Introduction

Document classification and topic modeling represent some of the biggest challenges in the fields of natural language processing and information retrieval. Many of the techniques developed for these purposes are language-independent (Sanderson and Bruce Croft, 2012). However, language resources are needed for each language, along with domain-specific data sets for particular applications, and every new language introduces a specific set of problems. In this paper, a method for addressing the problem of data sparsity in document classification for under-resourced, highly inflective languages, is proposed. The case of Serbian is considered, but the method is applicable to other languages as well. The approach includes training a language model on a large textual corpus, using it to create semantic word classes and using the extracted semantic information to obtain a more robust document classifier. As a topic model, Latent Dirichlet Allocation can be used, as well as its variants or other types of topic models.

2. Semantic Information Extraction

The method for semantic word class derivation has been described in a previous research (Ostrogonac et. al, 2015) and here will be described briefly. A textual corpus for Serbian, which contains over 20 million word tokens, which correspond to around 360 thousand word types, 180 thousand lemmas and around 1000 morphologic classes (Ostrogonac et. al, 2012) was used to train a language model (LM). The LM was lemma-based, since morphologic information was available for Serbian (Sečujski, 2002) and could be restored after semantic lemma classes were derived. The semantic classes were created by applying a greedy clustering algorithm (Mikolov, 2012) to the lemmatized textual corpus, which was based on lemma colocation as a basis for determining semantic similarity measure. The clustering algorithm leans on the probabilities obtained from the LM for hypotheses created by replacing a lemma with other lemmas from the dictionary. The lemmas for which the replacement causes the smallest change in probabilities are likely to be semantically similar to the original word. After the entire corpus is processed, and morphologic information is restored to derive words from lemmas, semantic word classes are created. The parameters for clustering should be fine-tuned by iteratively observing the results and adjusting the values so that the classes are optimized for a particular application. A semantic class can, therefore, represent only synonyms, but it can also represent all the words that can be placed in certain positions within sentences and result in semantically correct sentences, or it can represent something in between.

3. Semantic Word Classes in LDA

An LDA is a generative model which can be used for document classification (Blei et. al, 2003). One of the most popular document classification tasks is e-mail classification into regular messages and spam, which will be used in the following text in order to illustrate the effect of semantic word clustering. In LDA, a document is considered to be a mixture of a number of topics, which is similar to the bag of words (Mikolov, 2012) concept. Each word may belong to many topics, to each with a certain probability. In order to define those probabilities and the topics themselves, a great amount of data is needed. The main problem is that two spam messages can contain similar or the same topics, but consist of very different sets of words. For example, two spam messages containing the same advertisement may contain corresponding sentences such as “Buy now at lower price and enjoy the trip!” and “Purchase immediately, experience an exciting travel with our discount!”. This problem is emphasized in highly inflective languages. The lack of data results in poor classifiers. However, even though textual data of specific content may not be enough to train highly accurate classifiers, other textual resources can be used to obtain additional information. Semantic classes derived from

a large textual corpus which contains many different types of documents can be used to make a document classifier more robust. By using semantic class IDs instead of words, an LDA can model topics quite well even with a small amount of application-specific data, since for each word that is observed within the training data set, an entire semantic class is included in the modeling process. Semantic word clustering described in Section 2 insures that words with the same meaning but different morphological features are grouped together and therefore eliminates morphology as a cause of data sparsity in topic modeling. However, a morphologic dictionary is not available for all inflective languages. In those cases, other methods may be used to deal with this problem. Semantic classes may be grouped manually, or by applying a rule-based approach including word-stem derivation, for which the implementation is language-dependent. Other suboptimal solutions may be used as well. Furthermore, semantic classes include words with similar meaning, which reduces the number of topics to be modeled, resulting in more accurate topic representations. Spam detection is a fine example of how a classification process can benefit from semantic information extracted from an external source, but the application of the described approach is far more broad and includes other information retrieval tasks.

4. Further Research and Application

Semantic word clustering itself can be improved by implementation of a probabilistic approach, meaning that words would belong to more than one semantic class with specific corresponding probabilities. Furthermore, even though semantic classes obtained in the described way may contain words with similar meaning, no information about the correlation between the classes is extracted. For example, semantic class A = {malaria, flu, meningitis, AIDS, cancer…} and class B = {drug, medicine, therapy, cure, pill…} are highly semantically correlated, but this information is not extracted. Obtaining this higher-level semantic information requires wider context analysis, which will be the main topic of further research.

The applications of the extracted semantic information are numerous and represent the basis for creation of advanced dialogue systems, which would be able to mimic natural dialogue. The most important pursuit in this area would be to develop the possibility of determining the menaning of a word that a dialogue system has not seen before.

5. Acknowledgements

The work described in this paper was supported in part by the Ministry of Education, Science and Technological Development of the Republic of Serbia, within the project TR32035: “Development of Dialogue Systems for Serbian and Other South Slavic Languages”.

6. References

David M. Blei, Andrew Y. Ng, Michael I. Jordan, “*Latent Dirichlet Allocation*”, Journal of Machine Learning Research 3 (4–5): pp. 993–1022, January 2003. doi:10.1162/jmlr.2003.3.4-5.993

Mark Girolami, A. Kaban, “*On an Equivalence between PLSI and LDA*”, Proceedings of SIGIR 2003. New York: Association for Computing Machinery. ISBN 1-58113-646-3

Mark Sanderson, W. Bruce Croft, “*The History of Information Retrieval Research*”, Proceedings of the IEEE 100: 1444–1451, 2012. doi:10.1109/jproc.2012.2189916

Milan Sečujski, “*Accentuation Dictionary for Serbian Intended for Text-to-Speech Technology*”, Proceedings of DOGS, pp.17-20, Novi Sad, Serbia 2002.

Stevan Ostrogonac, Branislav Popović, Robert Mak, Milan Sečujski: “*Automatic Word Clustering Based on Semantics - an Approach for Serbian*”, 3rd International Acoustics and Audio Engineering Conference, TAKTONS 2015, Novi Sad, Srbija: Radio-televizija Vojvodine, Fakultet tehničkih nauka, Univerzitet u Novom Sadu, Srpska sekcija AES (Audio Engineering Society), Dirigent Acoustics, Beograd, 18-21. novembar 2015, pp. 36-37, ISBN: 978-86-7892-758-4.

Stevan Ostrogonac, Dragiša Mišković, Milan Sečujski, Darko Pekar, Vlado Delić: “*A Language Model for Highly Inflective Non-Agglutinative Languages*”, 10. SISY, International Symposium on Intelligent systems and Informatics, Subotica: IEEEXplore, 20-22.09.2012, ISBN: 978-1-4673-4749-5, pp. 177-181.

Tomáš Mikolov, “*Statistical language models based on neural networks*”, in PhD Thesis, Brno University of Technology, 2012.

Xiaogang Wang, Eric Grimson, “*Spatial Latent Dirichlet Allocation*”, Proceedings of Neural Information Processing Systems Conference (NIPS), 2007.

Dolgotrajno ohranjanje raziskovalnih podatkov v manjših raziskovalnih infrastrukturah
Uporaba odprtokodne aplikacije Archivematica

Andrej Pančur,* Bogomir Rožman†

* Inštitut za novejšo zgodovino
Kongresni trg 1, 1000 Ljubljana
andrej.pancur@inz.si
† UniPort – DR, računalniški inženiring, d.o.o.
Fosterjeva 40, 1000 Ljubljana
bogomir.rozman@uniport.si

1 Uvod

Evropska raziskovalna politika je v zadnjih letih začela intenzivno uvajati načela odprtega dostopa do raziskovalnih podatkov. Program Obzorje 2020 tako uvaja pilot za odprte raziskovalne podatke, v okviru katerega morajo udeleženci pripraviti načrt za ravnanje s podatki. Ta načrt mora zajemati življenjski cikel vseh raziskovalnih podatkov, pridobljenih ali ustvarjenih v okviru projekta. Naposled morajo udeleženci izbrati primeren raziskovalni podatkovni repozitorij, ki bo trajno hranil njihove podatke in metapodatke (European Commission, 2016). Države članice Evropske unije naj bi v skladu s priporočili Evropske komisije enaka določila kot za Obzorje 2020 uveljavile tudi za nacionalno financiranje raziskovalnih dejavnosti. V skladu s temi priporočili je Vlada Republike Slovenije septembra 2015 sprejela Nacionalno strategijo odprtega dostopa do znanstvenih objav in raziskovalnih podatkov v Sloveniji 2015-2020, v kateri je med drugim tudi predvideno, da bodo morali udeleženci pilotnega programa odprtega dostopa do raziskovalnih podatkov le-te predati področnim, institucionalnim ali splošnim repozitorijem raziskovalnih podatkov (Vlada Republike Slovenije, 2015, 20).

Pričakujemo lahko, da bodo v prihodnosti v Sloveniji poleg dveh obstoječih repozitorijev raziskovalnih podatkov (Arhiv družboslovnih podatkov[1] in CLARIN.SI[2]) začela nastajati še nova področna podatkovna središča, katera bodo zadovoljevala potrebe posameznih disciplinarnih področij v vseh fazah življenjskega cikla podatkov (Štebe et al., 2013, 7-8). Pri tem se v humanistiki glede na zelo različne vire in raziskovalna vprašanja življenjski cikli podatkov med seboj precej razlikujejo (Puhl et al., 2015). Zato nadaljnji razvoj podatkovne infrastrukture za humanistiko zahteva fleksibilnost, »ki bo zagotovila specializirano obravnavano glede na vsebino in tip podatkov« (Štebe in Bezjak, 2014, 11). Posledično lahko pričakujemo, da bo v Sloveniji (in verjetno tudi v drugih manjših članicah Evropske unije) v naslednjih letih nastalo več manjših specializiranih podatkovih središč, katera bodo lahko optimalno pokrivala celoten življenjski cikel raziskovalnih podatkov posameznih raziskovalnih področij.

Ker bodo denarna sredstva za razvoj teh podatkovnih središč razmeroma omejena, je nujno, da se bodo pri razvoju svoje infrastrukture čim bolj naslonile na obstoječe odprtokodne programske rešitve.

2 Dolgotrajno ohranjanje raziskovalnih podatkov

V okviru življenjskega cikla raziskovalnih podatkov je dolgoročno arhiviranje ponavadi sicer umeščeno na konec raziskovalnega procesa, vendar priprave nanj potekajo že od začetka (izbor metapodatkov in formatov). V predstavitvi se bova osredotočila zgolj na končno arhiviranje. Pri tem bova predpostavljala, da mora vsak zaupanja vreden sistem dolgotrajnega ohranjanja raziskovalnih podatkov temeljiti na standardu odprtega arhivskega informacijskega modela (OAIS).[3]

2.1 Archivematica – odprtokodna aplikacija za dolgotrajno arhiviranje

V zadnjih letih se manjše ustanove, ki se ukvarjajo s hranjenem digitalne kulturne dediščine, prav tako kot raziskovalne ustanove srečujejo s številnimi ovirami (tehnična in intelektualna kompleksnost ter visoki stroški) pri implementaciji najnovejših standardov s področja dolgotrajnega ohranjanja njihovih zbirk. Kot

[1] Arhiv družboslovnih podatkov, http://www.adp.fdv.uni-lj.si/.

[2] CLARIN.SI Repository, https://www.clarin.si/repository/xmlui/.

[3] ISO 14721:2012, Space data and information transfer systems – Open archival information system (OAIS) – Reference model, http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284.

odgovor na te potrebe je leta 2009 nastala Archivematica (Garderen, 2010), katero aktivno razvijajo še danes.[4] Archivematica je integrirana zbirka odprtokodnih programskih orodij, ki uporabniku omogoča obdelavo digitalnih objektov v skladu s funkcionalnim modelom OAIS (zajem, vzdrževanje arhiva, dostop). Uporabnik preko nadzorne plošče upravlja in nadzira izbrane mikrostoritve. Te so razdrobljen sistem opravil, katere delujejo na konceptualnem nivoju OAIS informacijskih paketov: sprejemni informacijski paket (SIP), arhivski informacijski paket (AIP) in dostavni informacijski paket (DIP). Informacijski paketi vsebujejo datoteke, XML metapodatke, dokumentacijo, checksum, log podatke ipd. Poleg originalno razvitih programskih rešitev Archivematica pri tem uporablja še mnoga odprtokodna programska orodja (bulk_extractor, Clam AV, ElasticSearch, ExifTool, FITS, fido, JHOVE, MediaInfo, Tesseract, Imagemagick, md5deep itd.). Archivematica uporablja METS, PREMIS, Dublin Core, BagIt in druge priznane standarde, na podlagi katerih tvori zaupanja vredne, avtentične, zanesljive in sistemsko neodvisne arhivske informacijske pakete (AIP), namenjeni hrambi v poljubnih repozitorijih. Trenutno je Archivematica že integrirana v repozitorije Islandora, dSpace in DuraCloud, ki jih uporabljajo infrastrukture s področja digitalne humanistike. Hkrati je integrirana še v številne sisteme, ki jih uporabljajo v glavnem ustanove s področja kulturne dediščine (CONTENTdm, LOCKSS, AtoM, OpenStack, Archivists' Toolkit, Arkivum, ArchivesSpace).

2.2 Raziskovalna infrastruktura Slovenskega zgodovinopisja

Archivematico smo za dolgotrajno arhiviranje začeli uporabljati tudi v okviru Raziskovalne infrastrukture Slovenskega zgodovinopisja na Inštitutu za novejšo zgodovino (RI INZ). Na RI INZ tako upravljamo portal Zgodovina Slovenije – SIstory,[5] v okviru katerega je najbolj opazna digitalna knjižnica digitaliziranih in digitalnih publikacij ter posnetki konferenc in predavanj. Poleg le-te pa razpolagamo še z različnimi zbirkami raziskovalnih podatkov v relacijskih bazah in v XML zbirkah. Vsi ti raziskovalni podatki niso shranjeni v enem repozitoriju. Hkrati je lahko digitalna kulturna dediščina, ki je najprej dostopna v okviru digitalne knjižnice, kasneje tekom (ponovne) uporabe v življenjskih ciklih raziskovalnih podatkov kot povsem nova zbirka raziskovalnih podatkov dostopna v drugih specializiranih repozitorijih ali bazah podatkov.

3 Zaključek

Na konferenci bo podrobneje predstavljena implementacija Archivematice v RI INZ. Sprva so Archivematico uspešno testirali in začeli uporabljati v številnih repozitorijih s področja ohranjanja kulturne dediščine. Šele pred kratkim pa so bolj intenzivno preizkusili uporabo Archivematice pri hranjenju raziskovalnih podatkov (Mitcham, 2016). Smatramo, da bodo na konferenci predstavljene izkušnje in testne meritve konstruktivno prispevali pri vzpostavljanju dolgotrajne hrambe raziskovalnih podatkov v manjših specializiranih podatkovnih centrih v Sloveniji in drugje po Evropi.

Literatura

European Commission. 2016. *Guidelines on Data Management in Horizon 2020*, verzija 2.1. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Peter Van Garderen. 2010. Archivematica: Lowering the Barrier to Best Practice Digital Preservation. V: Archiving 2010 Final Program and Proceedings, str. 39-41, Society for Imaging Science and Technology.

Jenny Mitcham, Chris Awre, Julie Allinson, Richard Green in Simon Wilson. 2016. Filling the Digital Preservation Gap: A Jisc Research Data Spring project Phase Two report – February 2016. http://dx.doi.org/10.6084/m9.figshare.1481170.

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. http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2015-4-4.

Janez Štebe in Sonja Bezjak. 2014. Nastavki odprtih podatkovnih zbirk kot podlaga za družboslovno in humanistično raziskovanje. *Glasnik (Slovensko etnološko društvo)*, 54(1/2): 8-16.

Janez Štebe, Sonja Bezjak in Sanja Lužar. 2013. *Odprti podatki: načrt za vzpostavitev sistema odprtega dostopa do raziskovalnih podatkov v Sloveniji.* Fakulteta za družbene vede, Založba FDV, Ljubljana. http://www.dlib.si/details/URN:NBN:SI:DOC-US3XRRB2.

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.

[4] Archivematica, https://www.archivematica.org/en/.

[5] http://sistory.si/.

Razvoj aplikacije za spodbujanje trajnostne mobilnosti

Dan Podjed,* Saša Babič,* Tatiana Bajuk Senčar,* Alenka Bezjak Mlakar,‡ Gregor Burger,†
Jurij Fikfak,* Jože Guna,† Marko Maver,‡ Matevž Pogačnik,†
Emilija Stojmenova Duh,† Uroš Žolnir‡

* Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti, Inštitut za slovensko narodopisje
Novi trg 2, 1000 Ljubljana
dan.podjed@zrc-sazu.si
† Univerza v Ljubljani, Fakulteta za elektrotehniko
Tržaška cesta 25, 1000 Ljubljana
matevz.pogacnik@fe.uni-lj.si
‡ CVS Mobile, informacijske rešitve, d.d.
Ulica Gradnikove brigade 11, 1000 Ljubljana
alenka.bezjak@cvs-mobile.com

1 Uvod

V prispevku predstavimo razvoj in delovanje aplikacije za pametne telefone *1, 2, 3 Ljubljana*, ki smo jo na podlagi primerjalne raziskave v Ljubljani, Beogradu, Budimpešti, Newcastlu in Durhamu razvili v triletnem interdisciplinarnem projektu *DriveGreen: Razvoj aplikacije za spodbujanje eko-vožnje pri prehodu v nizkoogljično družbo*. Najprej namenimo pozornost raziskovalnim izzivom projekta, s katerim smo povezali humanistiko in tehniške vede, nato pa predstavimo rezultate razvoja ljudem prijazne rešitve za spodbujanje trajnostne mobilnosti.

2 Povezovanje pristopov

Prvi izziv projekta *DriveGreen* je bil vzpostaviti sodelovanje med antropologi in etnologi ter inženirji, ki imajo povsem različne metodološke in epistemološke pristope pri raziskavah vožnje. Predstavniki omenjenih humanističnih področij navadno uporabljajo kvalitativne metode ter skušajo identificirati in analizirati tudi neizmerljive dejavnike, ki oblikujejo vozniške navade, inženirji pa se bolj zanašajo na kvantitativne pristope, denimo merjenje načinov vožnje s telematskimi napravami in zbiranje ter analizo podatkov o premikanju vozil in ljudi. Prvi pomembnejši dosežek projekta je bil zato skupen raziskovalni pristop, v katerem smo prepletli kvalitativne in kvantitativne metode in tehnike ter zagotovili relevantne in primerljive rezultate. Temelj kvalitativnega dela novega pristopa je bila t. i. večkrajevna etnografija (opazovanje z udeležbo, polstrukturirani intervjuji, fokusne skupine, video-etnografija itd.), ki smo jo prepletli s kvantitativnimi merjenji porabe goriva in izpustov CO_2 ter analizo načinov vožnje in gibanja s telematskimi napravami. S takšnim pristopom smo zagotovili podlago za večplastno primerjalno analizo mobilnostnih praks v urbanih središčih, s pomočjo katere lahko ugotovimo, kako se na obravnavanih lokacijah oblikuje *vozniški habitus* (prim. Bourdieu, 2002), kateri so glavni dejavniki, ki vplivajo na načine vožnje, kaj si ljudje v različnih mestih mislijo o prometu, kako dojemajo in sprejemajo druge udeležence na cesti ter kako se medsebojno sporazumevajo.

3 Razvojni obrat

Pri razvoju aplikacije smo poleg raziskovalnih izsledkov upoštevali izkušnje drugih raziskovalno-razvojnih skupin, ki so s pomočjo aplikacij za mobilne telefone skušale spodbujati ekološki in ekonomičen način vožnje. Preprost primer je Toyotina aplikacija za iPhone *A Glass of Water*, ki prikazuje vodo v virtualnem kozarcu, katere gladina se odziva na pospeševanje ali zaviranje. Količina vode, ki pljusne čez rob, se med vožnjo beleži, voznik pa tako pridobi povratne informacije o svoji učinkovitosti na cesti. Podobne rešitve se uveljavljajo v gospodarskih in osebnih vozilih (Shaheen et al., 2012), po podatkih raziskav pa se lahko z njihovo pomočjo poraba goriva zmanjša za 10–15 odstotkov, za toliko pa se zmanjšajo tudi izpusti CO_2 (Barkenbus, 2010; Podjed et al., 2013).

Na podlagi terenskih raziskav, fokusnih skupin, intervjujev in testiranja domačih in tujih aplikacij smo v prvem letu projekta *DriveGreen* naredili obrat od sprva načrtovanega razvoja rešitve za izboljšanje vožnje z osebnimi vozili k razvoju aplikacije za spreminjanje življenjskega sloga, pri čemer smo posebej

izpostavili zdravje, dobro počutje in pomen gibanja. Uporabnikov z aplikacijo nismo želeli spodbujati, naj se z osebnimi vozili prevažajo bolj varčno in varno, temveč smo jih skušali prepričati, naj se na pot raje odpravijo peš, s kolesom ali z javnim prevozom. Podatki o tem, kako malo se gibljemo, so, kot je pokazala raziskava, boljše motivacijsko sredstvo za spreminjanje navad kot ozaveščanje o negativnih vplivih prometa na okolje.

Tako je nastala aplikacija *1, 2, 3 Ljubljana,* ki pokaže, koliko je uporabnik v zadnjem dnevu, tednu, mesecu in letu hodil, tekel, kolesaril, uporabljal javni prevoz ter se vozil z avtomobilom (Slika 1). Dosežke aplikacija izmeri in prikaže s podatki, ki jih pridobi s senzorji telefona, kar pomeni, da uporabnik ne potrebuje dodatne naprave za merjenje razdalje in trajanja gibanja, temveč le mobilni telefon z operacijskim sistemom Android.

Slika 1: Aplikacija prikaže, kako se je uporabnik gibal v zadnjem dnevu, tednu, mesecu in letu.

Slika 2: Gibanje in trajnostne oblike mobilnosti aplikacija spodbuja z različnimi akcijami.

Pri razvoju smo posebej poudarili preprosto in celovito uporabniško izkušnjo (Krug, 2009; Tullis in Albert, 2013) ter dolgoročno motiviranje uporabnikov, kar smo skušali zagotoviti s *poigritvijo* (angl. *gamification*), kot jo opisujeta Zichermann in Cunningham (2011). Aplikacija *1, 2, 3 Ljubljana* tako spodbuja gibanje in trajnostne oblike mobilnosti z dnevnimi nasveti, točkovanjem dosežkov in lestvicami ter različnimi akcijami – tako individualnimi kot skupinskimi (Slika 2). Slednje so posebej pomembne, saj spodbujajo sodelovanje in solidarnost na mestni ravni ter zavedanje, da uporabnik ni sam na poti v nizkoogljično družbo.

4 Sklep

Raziskovalni izsledki in razvojni rezultati projekta *DriveGreen* kažejo, da lahko tehniške vede in humanistika uspešno sodelujejo pri razvoju novih tehnoloških rešitev. Predpogoj za interdisciplinarno sodelovanje je prilagajanje pristopov in vzpostavitev skupnega metodološkega okvira, ki omogoča poglobljene in primerljive raziskave ter vodi k razvoju izvirnih in ljudem ter okolju prijaznih rešitvam onkraj ustaljenih vzorcev in disciplinarnih praks.

5 Literatura

Jack N. Barkenbus. 2010. Eco-driving: An Overlooked Climate Change Initiative. *Energy Policy*, 38: 762–769.

Pierre Bourdieu. 2002 (1980). *Praktični čut*. Studia Humanitatis, Ljubljana.

Steve Krug. 2009. *Rocket Surgery Made Easy: The Do-It-Yourself Guide to Finding and Fixing Usability Problems.* New Riders, Berkeley.

Dan Podjed, Jernej Kosič, Hubert Benedik, Alenka Bezjak, Marko Maver in Marko Šetinc. 2013. Telematics Surveillance as a Solution to the Global Tragedy of the Commons. V: *ITS in Real Time: Proceedings of the 21st International Symposium on Electronics in Transport (ISEP 2013)*, str. 23–26. Elektrotehniška zveza Slovenije in Slovensko društvo za inteligentne transportne sisteme, Ljubljana.

Susan Shaheen, Elliot Martin in Rachel Finson. 2012. *Ecodriving and Carbon Footprinting: Understanding How Public Education Can Reduce Greenhouse Gas Emission and Fuel Use*. Mineta Transportation Institute, San José.

Thomas Tullis in William Albert. 2013. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics.* Elsevier, Amsterdam.

Gabe Zichermann in Christopher Cunningham. 2011. *Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps.* O'Reilly Media, Sebastopol.

Luščenje in jezikoslovna analiza kolokacij iz korpusa Šolar

Tadeja Rozman,*† Špela Arhar Holdt,*† Senja Pollak,§ Iztok Kosem*†

* Zavod za uporabno slovenistiko Trojina,
Trg republike 3, 1000 Ljubljana
spela.arhar@trojina.si, tadeja.rozman@trojina.si, iztok.kosem@trojina.si
† Filozofska fakulteta Univerze v Ljubljani,
Aškerčeva 2, 1000 Ljubljana
§ Inštitut »Jožef Stefan«,
Jamova cesta 39, 1000 Ljubljana
senja.pollak@ijs.si

1 Uvod

Korpus šolskih pisnih izdelkov Šolar (Kosem et al., 2012) predstavlja pomemben vir za raziskovanje pisne jezikovne zmožnosti učencev in je tako dragocen vir informacij za pripravo didaktičnih gradiv in priročnikov, namenjenih šolski populaciji. Čeprav je bil Šolar zgrajen že leta 2011, pa je bilo zaenkrat opravljenih malo korpusnih analiz, ki bi bile usmerjene v pridobivanje podatkov, pomembnih za izdelavo tovrstnih gradiv. V pričujočem prispevku bomo zato predstavili raziskavo, s katero deloma zapolnjujemo vrzel na področju leksikalnih analiz, ki omogočajo uvid v proces usvajanja besedišča v šolskem kontekstu.

2 Predhodne raziskave

Raziskava, ki jo predstavljamo, je nadgradnja raziskave Arhar Holdt in Rozman (2015). Čeprav je korpus Šolar s približno milijonom besed za analize leksike razmeroma majhen,[1] je omenjena raziskava med drugim pokazala,

a) da Šolar omogoča pridobitev novih spoznanj o rabi besedišča pri učencih ter tako predstavlja pomemben vir relevantnih informacij za izdelavo priročnikov, gradiv ter prenovo didaktičnih načel in praks na področju šolske obravnave besedišča,

b) da je za zanesljivejše zaključke potrebno analizirati večji del korpusnega gradiva, za kar moramo bolj avtomatizirati postopke luščenja informacij.

Ker so se kot gradivno zanimivejši pokazali učiteljski popravki besed, vezanih na kolokacijske omejitve rabe, bomo v pričujočem prispevku predstavili metodo luščenja kolokacij ter analizirali izbrane primere s stališča uporabnosti tako pridobljenih podatkov za pripravo šolskih priročnikov in gradiv.

3 Metoda in rezultati

Za luščenje kolokacij bomo uporabili postopek, ki je bil predhodno preizkušen za primerjavo kolokacij v korpusih GOS in Kres in za luščenje korpusnospecifičnih kolokacij korpusa uporabniških vsebin Janes (Pollak, 2015). Z metodo bomo primerjali kolokacije v korpusu Šolar in uravnoteženem referenčnem korpusu Kres (Logar et al., 2012) .

Za ponazoritev navajamo nekaj rezultatov metode luščenja kolokacij, ki se opira na orodje SketchEngine (Kilgariff et al., 2004) in z avtomatskim izvozom preko API-ja (Pollak, 2015) za določen seznam besed izvozi kolokacije, njihove frekvence in kolokacijske vrednosti ter povezavo na korpusni zgled. Luščili smo kolokatorje, ki se pojavljajo ob najpogostejših stotih samostalniških lemah v korpusu Šolar (*človek, življenje, ljubezen, otrok, čas* itd.), in sicer tiste, ki se pojavljajo na mestu pred lemo in so označeni bodisi kot pridevnik, glagol ali samostalnik. Med kolokacijami, specifičnimi za korpus Šolar, ki imajo vsaj 5 pojavitev in vrednost logDice nad 3, je po pričakovanjih najti precej primerov z lastnoimenskimi kolokatorji (*Bogomilina ljubezen, hlapec Jernej*) in primere, ki vsebujejo jezikovne popravke (o teh gl. spodaj). Za razumevanje procesa usvajanja besedišča v šolskem kontekstu je zanimiva ugotovitev, da so preostali primeri – čeprav na pogled vsebinsko širši – pogosto tesno vezani na specifično obravnavano delo, npr. *pretentati barona* (Ta veseli dan ali Matiček se ženi), *večinska vera* (Krst pri Savici), *absurdno dejanje* (Tujec). Na drugi strani je najti kolokacije, ki se pojavljajo kot terminologija v šolskih testih, npr. *aplikativni cilj, fobični človek*.

V raziskavi bomo metodo luščenja razširili s primerjalno metodo (Pollak in Arhar Holdt, 2015) in prilagodili analizo, da bo ustrezala specifikam korpusa Šolar in namenu raziskave. Za razliko od predhodnih raziskav, v katerih je bilo v središču pozornosti predvsem besedišče, ki je glede na referenčni korpus novo in

[1] Šolar se bo do leta 2018 povečal na predvidoma dva milijona besed, saj trenutno poteka projekt »Nadgradnja korpusa Šolar«, ki ga financira Ministrstvo za kulturo RS.

drugačno, nas namreč pri primerjavi s korpusom Šolar zanima tudi besedišče, ki je v korpusih enako oz. primerljivo. Podatke bomo rangirali v različne skupine: (I) kolokacije, ki se pojavljajo zgolj v korpusu Šolar; (II) kolokacije, ki se pojavljajo v obeh korpusih; (III) kolokacije, ki se pojavljajo samo v korpusu Kres. Ob (kritično ovrednoteni) predpostavki, da korpus Šolar predstavlja pisanje mladostnikov, ki pisno kompetenco šele razvijajo, Kres pa vzorec odraslih, izkušenih piscev, je mogoče podatke nadalje kategorizirati in interpretirati v iskanju zadreg in močnih točk šolskega pisanja. Druga sprememba v zornem kotu je premik od tipičnih, pogostih kolokacij do zvez, ki se pojavljajo redko, vendar v širšem naboru podobnih primerov lahko ponudijo uvid v usvajanje večbesednih enot (npr. korpusnospecifične zveze [*kovati, opredeliti, povprašati, izvedeti, dopustiti*] *mnenje* ali [*predati, upreti*] *se mnenju*).

Dodaten izziv luščenja je upoštevanje posebnih oznak korpusa Šolar, tj. oznak učiteljskih jezikovnih popravkov v besedilih. Trenutno je postopek na te oznake neobčutljiv, zato med rezultati korpusnospecifičnih kolokacij dobimo tudi pare napaka-popravek na ravni posamezne besede (npr. *dogotek dogodek; zakonik zakon*). Te primere je smiselno ločevati od primerov, kjer se učiteljska oznaka nanaša na kolokacijsko raven. V prispevku raziščemo možnosti za avtomatsko ločevanje enih in drugih primerov.

Primerjalno analizo kolokacij v korpusu Šolar in korpusu Kres, ki jo omogoča postopek luščenja kolokacij po zgornji metodi, bomo dopolnili z analizo jezikovnih popravkov v korpusu Šolar. V ta namen smo opravili postopek luščenja kolokacij, ki morajo zadostiti dvema pogojema: vsaj en del kolokacije mora biti označen kot napaka in biti hkrati popravljen, jezikovni popravek pa je uvrščen v tip 'napaka besedišča'. Podatke smo izluščili za šest tipov kolokacij (del s popravkom je pisan z velikimi začetnicami, v oklepaju je podano število najdenih zadetkov): pridevnik + SAMOSTALNIK (147), PRIDEVNIK + samostalnik (131), SAMOSTALNIK + samostalnik (176), samostalnik + SAMOSTALNIK (95), glagol + SAMOSTALNIK (171), GLAGOL + samostalnik (103). Ker pri večjem kolokacijskem razponu, npr. -5 +5, zaradi načina označenosti korpusa z jezikovnimi popravki dobimo precej šuma, smo iskanje kolokacij omejili na zaporedne kombinacije besed. Sledila je kategorizacija, podrobna analiza in primerjava s podatki prvega luščenja.

4 Sklep

V prispevku bomo predstavljeno metodo luščenja kolokacij evalvirali, na izbranih primerih opravili kvalitativno analizo in ovrednotili korpus Šolar kot vir (relevantnih) podatkov o specifičnosti rabe kolokacij pri šolski populaciji. S tem nadaljujemo prizadevanja po pridobivanju empiričnih podatkov o rabi besedišča med učenci ter želimo spodbuditi nadaljnje empirične raziskave leksikalne problematike. Te so pomembne tudi v luči prizadevanj za izdelavo sodobnih slovarjev, ki bodo upoštevali zmožnosti in potrebe ciljnih uporabnikov (Gorjanc et al., 2015), med drugim tudi mladih v procesu izobraževanja (Rozman et al., 2015), ter pozivov po spremembah poučevanja slovenščine v šolah, ki so mu deloma botrovali tudi rezultati raziskav (PISA, PIRLS, NPZ) o upadu bralne pismenosti in neučinkovitosti pouka pri razvijanju znanj na višjih taksonomskih ravneh (npr. Rozman et al., 2012; Stabej, 2011).

Literatura

Špela Arhar Holdt in Tadeja Rozman. 2015. Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov. V: M. Smolej, ur., Obdobja 34: Slovnica in slovar - aktualni jezikovni opis, 1. del, str. 67–74. Znanstvena založba Filozofske fakultete UL. http://centerslo.si/wp-content/uploads/2015/11/34_1-Arhar-Hol-Roz.pdf

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek, ur. 2015. *Slovar sodobne slovenščine: problemi in rešitve*. Znanstvena založba Filozofske fakultete UL.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz in David Tugwell. 2004. The Sketch Engine. V: G. Williams in S. Vessier, ur., *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004*, str. 105–116. Universite de Bretagne-sud.

Iztok Kosem, Mojca Stritar Kučuk, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012. *Analiza jezikovnih težav učencev: korpusni pristop*. Trojina, zavod za uporabno slovenistiko.

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

Senja Pollak. 2015. Luščenje kolokacij iz korpusa uporabniških spletnih vsebin. V: M. Smolej, ur., *Obdobja 34: Slovnica in slovar - aktualni jezikovni opis*, 2. del, str. 601–607. Znanstvena založba Filozofske fakultete UL. http://centerslo.si/wp-content/uploads/2015/11/34_2-Pollak.pdf

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

Tadeja Rozman, Iztok Kosem, Nataša Pirih Svetina in Ina Ferbežar. 2015. Slovarji in učenje slovenščine. V: V. Gorjanc, P. Gantar, I. Kosem in S. Krek, ur., *Slovar sodobne slovenščine: problemi in rešitve*, str. 150–167. Znanstvena založba Filozofske fakultete UL.

Tadeja Rozman, Irena Krapš Vodopivec, Mojca Stritar Kučuk in Iztok Kosem. 2012. *Empirični pogled na pouk slovenskega jezika*. Trojina, zavod za uporabno slovenistiko.

Marko Stabej. 2011. Jezikovni potrošnik in potrošnica. *Sodobna pedagogika*, 62=128(2), 102–113. Zveza društev pedagoških delavcev Slovenije. http://www.dlib.si/details/URN:NBN:SI:doc-RF8JNPLQ

Historična topografija Slovenije

Miha Seručnik

Znanstvenoraziskovalni center SAZU,
Zgodovinski inštitut Milka Kosa Novi trg 2, 1000 Ljubljana
miha.serucnik@zrc-sazu.si

1 Uvod

Pri raziskavah zgodnejših zgodovinskih obdobij se raziskovalec sooči s problematiko identifikacije krajevnih imen, s katerimi se sreča v virih. V obdobju fevdalnega družbenega reda je zemljiška posest predstavljala temelj gospodarske in družbene moči. Pravni promet z nepremičninami je zato igral osrednjo vlogo v srednjeveški družbi, kar se kaže tudi v strukturi ohranjenih pisnih virov. Zaradi vloge zemljiške posesti so opo imenovanja delov pokrajine, gora, vodotokov in naselij eden najbolj izpostavljenih elementov vsebine omenjenih pisnih virov.

Večinoma tujerodni pisarji so slovenska krajevna imena zapisovali tako, da so se poskušali čim bolj približati njihovemu glasovnemu učinku, hkrati pa so jim dodajali primerna latinska oziroma nemška in italijanska obrazila in predloge. Drugačen pristop je bil dobesedni prevod (kalk) slovenskega imena v jezik dokumenta. V visokem in deloma poznem srednjem veku se pojavijo še krajevna imena, ki so bila izvorno nemška (na Primorskem latinska/italijanska), danes pa jih poznamo v uradni slovenski obliki. Mednje sodijo v prvi vrsti imena gradov in urbanih naselij, na nemškem kolonizacijskem območju pa tudi imena vasi in ledin.

Zaradi navedenih dejavnikov in zaradi odsotnosti ustaljenih pravopisnih pravil v srednjem veku, je raziskovalec soočen s precejšnjo morfološko raznolikostjo poimenovanj za isto lokacijo. Hkrati so določena poimenovanja precej pogosta, zaradi česar se pojavi potreba po razlikovanju med različnimi kraji z enakim imenom (npr. Brezje ali Javor).

Na podlagi tradicije zgodovinsko-topografskih priročnikov, s katerimi so si zgodovinarji pomagali pri svojem delu (Kos 1975, Zelko 1982, Blaznik 1986-1989), smo se na Zgodovinskem inštitutu Milka Kosa odločili izdelati spletni pripomoček oziroma aplikacijo.[1] Slednja dopolnjuje klasično topografijo z novimi funkcionalnostmi, ki jih omogočajo informacijske tehnologije. Namenjena je tako zgodovinarjem in drugim humanističnim raziskovalcem (npr. etimologom) kot tudi širši javnosti. Aplikacija, ki jo predstavljamo je rezultat temeljnega raziskovalnega projekta *Slovenski toponimi v prostoru in času*, ki je bil financiran s strani Agencije za raziskovalno dejavnost Republike Slovenije in je vključeval kritično analizo in preverjanje starejših historičnih topografij ter pritegnitev dodatnih virov.

2 Namen prispevka

V uvodnem delu prispevek predstavi problematiko historične topografije in njeno predzgodovino, v nadaljevanju pa sledi predstavitev projekta *Slovenski toponimi v prostoru in času* oziroma spletne aplikacije, ki predstavlja glavni rezultat omenjenega projekta.

Jedro nove historične topografije predstavlja relacijska zbirka podatkov, ki temelji na gradivu Kosove »Topografije za Kranjsko« (1975). Vključuje 3665 lokacij, od katerih jih je večina geolociranih. Lokacije so opremljene z današnjimi uradnimi imeni in njihovimi historičnimi različicami (v trenutku pisanja 16451 zapisov), ki se uporabljajo v virih. Vsaka historična oblika je opremljena z datacijo in navedbo vira. Uporabniški vmesnik v privzetem pogledu ponudi v sse geolocirane lokacije, prikazane s pomočjo markerjev, uporabnik pa jih lahko nato filtrira s pomočjo iskalnega polja in s klikanjem dostopa do podrobnejših podatkov. Aplikacija omogoča iskanje historičnih imen lokacij (ki smo jih poimenovali *paleonimi*) z vidika današnjih krajevnih imen in tudi v obratni smeri od paleonimov k današnjim krajevnim imenom. Aplikacija tako omogoča primerjavo prostorske pojavnosti določenih krajevnih imen ter njihove postosti. V prihodnosti želimo omogočiti še funkcionalnost naprednega iskanja, ki bo med drugim omogočala tudi zamejitev prikaza po časovni komponenti.

[1] http://topografija.zrc-sazu.si.

Pri oblikovanju zbirke podatkov je bilo treba razrešiti vrsto problemov, povezanih s podatkovnim modelom ter domensko specifiko historičnih virov. Izhodiščna točka projekta je bilo skenirano besedilo Kosove *Topografije*, na katerem je bila opravljena optična prepoznava besedila (OCR), ki je bila nato deležna še »ročnega« pregleda. Izhajajoč iz zgoraj omenjene predloge smo prvotni podatkovni model zasnovali tako, da smo oblikovali dva osnovna objekta – *toponime*, ki predstavljajo danes obstoječe lokacije s sodobnimi krajevnimi imeni, ter *paleonime* oziroma v historičnih virih izpričane oblike poimenovanj lokacij. Tekom projekta smo ugotovili, da je prvotni model preveč vezan na tiskano predlogo in smo ga morali za potrebe geolokalizacije prilagoditi. Sedanja rešitev se zgleduje po modelu, ki je bil razvit za potrebe antičnega krajevnega imenika Pleiades.[2]

Poseben izziv so predstavljale datacije, saj podatki vsebujejo veliko datumov, ki niso natančni, ampak je znano na primer le leto omembe ali časovno razdobje nastanka vira. Obstoječe platforme za relacijske zbirke podatkov ne podpirajo formata zapisa datumov, ki bi zadovoljeval vse potrebe historične datacije. Problem smo rešili z implementacijo datumsko-časovnega formata, ki ga je ustvarila kongresna knjižnica iz Washingtona – Extended date/time format.[3] Takšna rešitev je seveda dvorezna, saj na ravni relacijske zbirke (MySql strežnik) po meni, da so podatki zapisani v obliki znakovnih nizov (string) in je pravilnost zapisa potrebno nadzorovati na ravni aplikacije.

S tehničnega vidika je spletna aplikacija kombinacija Javascript in PHP skriptov. Lokacije so prikazane na Google Maps spletnem zemljevidu s pomočjo markerjev.

Kot že rečeno spletno aplikacijo še dograjujemo in ji postopoma dodajamo funkcionalnosti. Registrirani uporabnik/sodelavec bo imel dostop do spletnih obrazcev, s pomočjo katerih bo mogoče ustvarjati nove zapise oziroma dopolnjevati in popravljati obstoječe. S tem želimo zagotoviti ažurnost aplikacije. Trenutno za zbirko podatkov skrbijo člani projektne skupine, v prihodnje pa upamo na pritegnitev dodatnih sodelavcev oziroma strokovnih zunanjih uporabnikov. Srednjeročno želimo tudi, da bo aplikacija omogočala funkcionalnost servisa za druge projekte,[4] zato načrtujemo, da jo bomo opremili z API-jem, ki bo omogočal pridobitev izpisov v RDFa in JSON formatih. Hkrati želimo topografijo nadaljevati tudi v vsebinskem smislu ter pos topoma pokriti celotno slovensko državno ozemlje.

3 Literatura

Pavle Blaznik. 1928. Bitenj : Historično-geografska študija. *Geografski vestnik*, 4, str. 88–98. Zveza geografskih društev Slovenije, Ljubljana.

Pavle Blaznik. 1952-53. Doneski k historični topografiji ljubljanske okolice. Zgodovinski časopis, letnik 6-7, str. 391-397. Zveza zgodovinskih društev Slovenije, Ljubljana.

Pavle Blaznik. 1966. Topografija vitanjskega urada v luči urbarja iz 1404. *Časopis za zgodovno in narodopisje*, n. v. letnik 2, str. 96-103. Založba »Obzorja«, Maribor.

Pavle Blaznik. 1986–1989. *Historična topografija Slovenije II. Slovenska Štajerska in jugoslovanski del Koroške, do leta 1500* (3 deli). Maribor.

Milko Kos. 1965–1966. Doneski k historični topografiji Kranjske v srednjem veku. *Zgodovinski časopis*. Letnik 19–20, str. 139–147. Zveza zgodovinskih društev Slovenije, Ljubljana.

Milko Kos. 1966. *»Vas« in »selo« v zgodovini slovenske kolonizacije. Razprave I. razreda SAZU 5*. str. 77–98. Slovenska akademija znanosti in umetnosti, Ljubljana.

Milko Kos. 1975. *Gradivo za historično topografijo Slovenije : (za Kranjsko do leta 1500)*. Inštitut za občo in narodno zgodovino Slovenske akademije znanosti in umetnosti, Ljubljana.

Joseph von Zahn. 1893. *Ortsnamenbuch der Steiermark im Mittelalter*. A. Hölder, Wien.

Ivan Zelko. 1982. *Historična topografija Slovenije I. Prekmurje do leta 1500*. Pomurska založba, Murska Sobota.

[2] http://pleiades.stoa.org/help/technical-intro-places

[3] http://www.loc.gov/standards/datetime/

[4] Geolokacija bi lahko dopolnjevala prepis in označevanje (mark-up) v TEI formatu, kot ga ponuja npr. monasterium.net.

Language Technologies in Humanities: Computational Semantic Analysis in Folkloristics

Gregor Strle,* Matija Marolt†

* Institute of Ethnomusicology ZRC SAZU
Novi trg 5, 1000 Ljubljana
gregor.strle@zrc-sazu.si
† University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana
matija.marolt@fri.uni-lj.si

1. Introduction

The paper discusses computational methods for natural language processing (NLP) and possibilities they offer to folkloristics.[1] As folkloristic materials are very challenging for NLP, due to their specific semantic-syntactic structure, inherent dialectical diversity and strong intertextuality, a robust NLP method is needed that can account for topical distribution, detect general heterogeneity, and context. The focus of this paper is on computational semantic analysis (such as word-sense disambiguation, topic recognition) and its ability to uncover latent semantic structure of folkloristic corpora.

2. Objectives

Our aim is to determine the appropriatness of NLP for analysing general patterns and relationships on the level of song types and genres, and also specify a general procedure (necessary steps) for conducting computational analysis of folkloristic materials. Two main approaches are presented and compared on the large-scale corpus of Slovenian folk songs: a statistical associative approach using Latent Semantic Analysis (LSA) and a probabilistic topic modelling approach using Latent Dirichlet Allocation (LDA). Emphasis is being placed on the practical applications of LSA and LDA models for analysing folkloristic corpora.

3. Methods

3.1. Latent semantic analysis (LSA) and Latent Dirichlet allocation (LDA)

Latent semantic analysis (LSA; Landauer and Dumais, 1997) and *Latent Dirichlet allocation* (LDA; Blei et al., 2003) are two of the most known methods used in NLP. They differ in theory and their approach to the semantic analysis. LSA does computations on high-dimensional similarity-space representations of associations between words, extracting the 'meaning' of individual words based on their proximity to other words in the semantic space. LDA, on the other hand, is a generative probabilistic topic model – it tries to uncover the blend of latent topics as distributions over documents and words. The central question for topic modeling approach is, 'what is the hidden structure behind these documents?'

The analysis and visualization of song types was made by Matlab topic modeling toolbox (TMT). In LDA, the analysis is an optimization process initialized randomly for the number of topics given as an input parameter. Consequently, multiple calculations yield slightly different results. The input parameter for the number of topics impacts the representation of semantic structure – smaller number of topics results in a more general overview of topic distributions in the corpus, whereas higher number of topics gives greater segmentation and detail.

3.2. Corpus

Both methods were tested on a large corpus of 3449 folk song variants from the collection of Slovenian folk songs (SLP I-V)[2]. The songs date back to the 19th century, with some variant types represented by only one variant and others consisting of up to 180 variants. Moreover, strong intertextuality is present throughout the corpus, which reflects characteristic phenomenon of Slovenian folk song: traveling of verses, motifs, and

[1] This work discusses previous research on semantic analysis of folkloristic materials (Strle and Marolt, 2014)
[2] Part of the EthnoMuse multimedia archive (Institute of Ethnomusicology, ZRC SAZU): www.ethnomuse.info

thematic patterns from one song to the other, within and across variant types. This consequently affects the results, as most occurring topics and motifs dominate over lesser ones.

3.3. Document preprocessing

Lemmatization of the documents was performed in two steps (Figure 1). First, special characters used for encoding characteristics of dialect groups (such as semivowels, diphthongization, pitch accent etc.) are replaced by their grammatical equivalents. A dialect dictionary containing over 18.000 entries, specifically built from the folk song corpus, was used to translate the resulting words into literary language. In the second step, we used the statistical morphosyntactic tagger for the Slovenian language Obeliks (Grčar et al., 2012) to lemmatize the text.

A	Nəč predowga, nəč prekratka, sej ne bom plesala u nji.
B	Nič predolga, nič prekratka, sej ne bom plesala v nji.
C	nič predolg nič prekratek saj ne biti plesati v on

Figure 1: Lemmatization: A shows the original text, B the text after removal of a dialect and C the lemmatized text

4. Results

Two cases of using NLP analysis in folkloristics are presented: the analysis of general semantic structure of the corpus by the respective approach and the hierarchical clustering of folk song variants into folk song family types (the latter by using LDA only). Both cases give insight into thematic and semantic relationships, with the LDA clustering of folk song variants proving especially useful for building folk song typologies on the fly.

The results of general topic analysis show differences in the semantic structure of the corpus generated by LSA and LDA. There is a significant difference between both methods in their ability to the detect topics across the corpus and within various song families. Due to its simple design, LSA can only detect more prevalent topics across the semantic space. Arguably, this is the main limitation of LSA. As LSA model cannot account for topic distribution, it has difficulty detecting heterogeneity and the resulting semantic space repeatedly generalizes towards most salient (frequent) topics of the corpus. LDA is better in detecting the heterogeneity of the corpus and provides a more balanced representation of the semantic space (see Figure 2). Furthermore, the topic clusters generated by LDA correlate with the division of songs into song families and the general typology of the corpus.

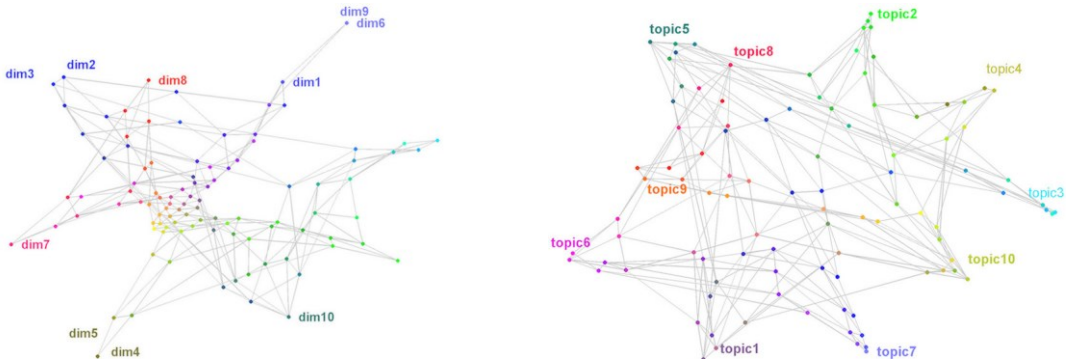


Figure 2: Comparison of semantic spaces: LSA (left) and LDA (right)

For the purpose of hierarchical clustering and visualization of song variant types only a subset of the corpus was chosen with the analysis limited to 5 clusters and two song family types, the narrative poems about fate and conflict in love and family. Our goal was to train the model on a smaller scale and investigate how it deals

with the challenge of strong intertextuality present in both song types. We first calculated the average topic vector for each individual variant type by averaging the vectors of all the variants belonging to the respective variant type. The purpose of averaging is to reduce the disparity between the disproportionate number of variants representing particular variant type. The output is single (average) topic vector representing individual variant type, totalling 88 topic vectors for the song types analysed in the collection.

The method of hierarchical clustering was then used to divide variant types into clusters similar to the two folk song family types (love vs. family song type). The similarity of song types was calculated as the cosine similarity between topic vectors. Hierarchical clustering in Figure 3 shows the division of song types within the cluster, as well as the relationship between the individual clusters. Branches of the dendrogram for all five clusters are composed of 30 sub-groups, dividing all song variant types (88) into love (36) and family (52) types, with the former prevailing in clusters 2 and 5, and the latter in clusters 1, 3, and 4. This division indicates approximately 60% dominance of family narrative poems, which corresponds to the division between love and family songs in the corpus.

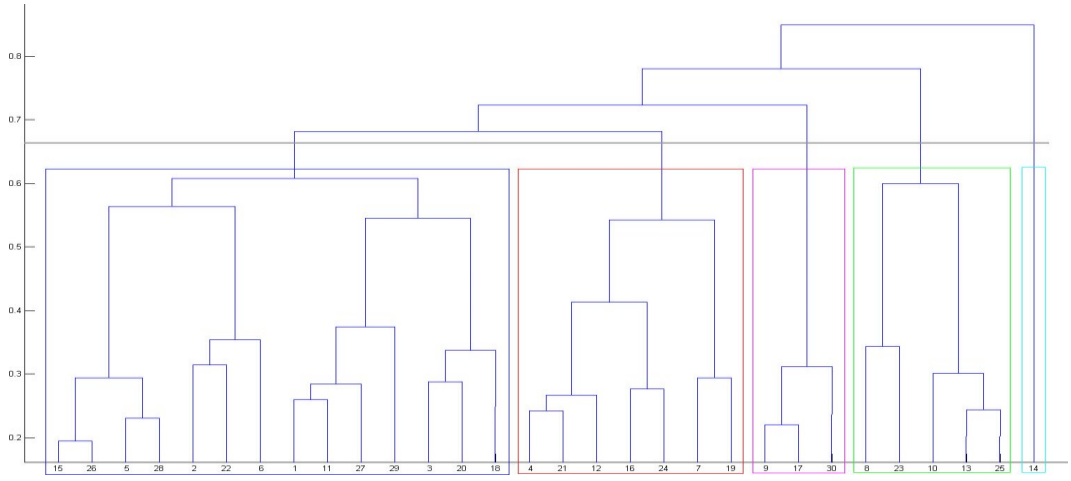


Figure 3: Hierarchical clustering of folk song poems about love and family fate. Colour boxes indicate 5 clusters: the songs about family fate prevail in clusters 1, 3, and 4, whereas the songs about love prevail in clusters 2 and 5

5. Conclusion

The main advantage of using NLP in folkloristics is the ability to analyze the semantic structure of large corpora, going beyond the limitations of traditional methods used in the office and fieldwork. Additional advantage of generative probabilistic models (such as LDA) is the ability to learn and generalize on new information, and thus expand existing analyses with new examples. This is especially handy in situations where we need to follow the semantic and typological transformations both chronologically and thematically. Future investigations will consider how computational methods can be used in folkloristics for more complex semantic and typological analyses.

6. References

Gregor Strle and Matija Marolt. 2014. New approaches: uncovering semantic structures in ethnological materials | Novi pristopi. Odkrivanje semantičnih struktur v etnoloških vsebinah. *Glasnik SED*, vol. 54/1-2, pp. 17-21.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, pp. 211-240.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5), pp. 993-1022.

Miha Grčar, Simon Krek, and Kaja Dobrovoljc. 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In: T. Erjavec, J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan.

Slikovna retrospektiva porušenega Breginja in analiza pokrajinskih sprememb

Tatjana Veljanovski,* Žiga Kokalj*†

* Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti
Novi trg 2, 1000 Ljubljana
tatjana.veljanovski@zrc-sazu.si
† Center odličnosti Vesolje, znanost in tehnologije
Novi trg 2, 1000 Ljubljana
ziga.kokalj@zrc-sazu.si

Ključne besede: zgodovinske letalske fotografije, slikovna rekonstrukcija, kulturna dediščina, Breginj, 3-razsežna rekonstrukcija naselja, analiza pokrajinskih sprememb

Keywords: historic aerial photography, image-based reconstructions, cultural heritage, Breginj, settlement 3D reconstruction, landscape change analysis

1 Uvod

S tem sestavkom želiva prikazati potencial zgodovinskih fotografij za digitalno slikovno retrospektivo krajev, ki v prvotni obliki ne obstajajo več, vendar so pustili pomembne sledove v človeški in pokrajinski zgodovini. Kljub temu, da študija temelji na primeru Breginja, želiva pozornost usmeriti na splošen pomen in potencial tovrstnih slikovnih rekonstrukcij v luči doprinosa k ohranjanju arhitekturne in kulturne dediščine, natančneje, zmožnostim in potencialu zgodovinskih fotografij za digitalno retrospektivo naselij in preteklih pokrajin.

Breginj, odročno hribovsko naselje v bližini meje z Italijo, zahodno od Kobarida, je bil nekdaj samostojna občina s široko avtonomijo. Stoletja samozadostna in izredno organizirana lokalna skupnost je znala izkoristiti tako naravne danosti za kmetijstvo kot obmejno lego (tihotapljenje, obramba, začasno delo čez mejo) (Pipan, 2011). Tedanji Breginj je bil eno redkih večjih naselij z dobro ohranjeno arhitekturo stavb 18. in 19. stoletja, ki je pripadala beneškoslovenskemu arhitekturnemu tipu in je bila zaščitena kot spomenik prve kategorije (Lipušček, 1995; Celarc in Erjavec, 2012). V to živahno okolje sta leta 1976 globoko zarezala zaporedna potresa, sledile so politično spodbujene odločitve in izvajati so se začele popotresne dejavnosti. Življenje se je spremenilo in pustilo osupljive posledice v pokrajinski sliki naslednjih desetletij.

Stari Breginj so porušili domala čez noč; pustili so le nekaj stavb, ki naj odražajo stari slog gradnje kamnitih hiš v ogradih. Novo vas modernih stavb pa so v nekaj mesecih zgradili na drugi strani potoka proti vzhodu.

Za pridobitev vpogleda v izgled starega naselja in njegove pomembnosti v času in prostoru sva uporabila zgodovinske letalske fotografije. Z dvema različnima pristopoma sva želela izluščiti različne zgodovinske prostorske informacije, in sicer:

- s postopkom grajenja strukture iz gibanja (angl. Structure-from-motion photogrammetry – SfM) detajlno trirazsežno rekonstrukcijo starega vaškega jedra in
- z objektno usmerjeno slikovno analizo vpogled v pokrajinske spremembe v zadnjih štirih desetletjih.

2 Digitalna slikovna retrospektiva

Kulturna dediščina je zabeležena na številnih fotografijah. Če je bila spremenjena, poškodovana ali uničena, so fotografije pogosto edini dokaz o njeni vizualni podobi in poteku sprememb. V primeru Breginja je vas popolnoma spremenila svojo lego in strukturo, kar je, skupaj s splošno depopulacijo ruralnih območij, povzročilo postopno spremembo načina življenja v vasi in rabo okolišnje zemlje. V tej študiji naju je na eni strani zanimalo kako, kje in do kakšne mere se lahko pokrajinski videz spremeni v nekaj desetletjih ter kako uspešno to lahko ugotavljamo iz zgodovinskih letalskih fotografskih virov? Na drugi strani pa ali je mogoče iz razpoložljivih letalskih fotografij pridobiti digitalni model naselja ter kakšno natančnost upodobitve stavb pri tem dosežemo?

Sodobne metode obdelave slikovnih podatkov (tj. fotografij, letalskih in satelitskih posnetkov) podpirajo izdelavo različnih približkov zgodovinskih objektov. Odvisno od virov in namena slikovne rekonstrukucije je objekt pri tem lahko posamezen predmet ali arheološka najdba, lahko pa tudi stavba, najdišče oziroma naselje ali vsepogosteje pretekla pokrajina. Digitalna rekonstrukcija porušenega Breginja je trenutno vzočni primer tega, kar potencialno lahko slikovna rekonstrukcija ponudi. Pomembna je predvsem v luči naslednjih zmožnosti:

- omogoča prostorsko in časovno retrospektivo zgodovinskih procesov na opazovanem območju,
- ponuja obnovo večjih zgodovinskih objektov (starega vaškega jedra) v digitalnem prostoru ter s tem ohranjanje spomina na naselje, ki je bilo pomemben spomenik arhitekturne dediščine in
- poda izbrane merljive informacije o učinkih popotresne sanacije in političnih odločitev na strukturo in funkcijo pokrajine in ljudi.

3 Zgodovinske letalske fotografije in 3-razsežna rekonstrukcija naselja

Slovenija ima široko znanje za uporabo fotogrametričnih tehnik na področju ohranjanja in dokumentacije arhitekturne dediščine (Kosmatin Fras 1996). Od leta 1993 naprej se v okviru projekta IZMERE pod okriljem spomeniškovarstvenih služb izvaja nacionalni projekt preventivnega fotogrametričnega snemanja objektov in območij kulturne dediščine (Grobovšek 2002). Glavni namen je snemanje objektov državnega pomena in vseh tistih, ki so ogroženi, ter vzpodbujanje vsestranske uporabe natančnih položajnih in 3R podatkov o objektih. Z vidika splošnega varstva nepremične kulturne dediščine je vrednost dokumentiranja predvsem v zbiranju vsebin iz različnih slikovnih in neslikovnih virov. Vendar so bili mnogi objekti porušeni ali predelani še pred uveljavitvijo nacionalnega projekta dokumentacije. Prav fotografije so pogosto edini dokaz o njihovem obstoju in razvoju in našteto velja tudi za staro vaško jedro Breginja. Ker fotogrametrični zajem ni več mogoč, smo preizkusili pristop prilagojene fotogrametrične obravnave fotografij za 3-razsežne rekonstrukcije objektov, s postopkom grajenja strukture iz gibanja.

Po prvem potresu maja 1976 so območje za potrebe opazovanja opustošenja fotografirali iz letala. Nastala je zbirka nizkih preletov naselij v zelo visoki ločljivosti. Niz šestih navpičnih letalskih fotografij smo uporabili za digitalno rekonstrukcijo porušenega naselja. Fotografije so bile zajete v enem samem preletu in vsako točko na tleh lahko vidimo iz največ treh kotov gledanja. Skenirali smo jih z običajnim namiznim skenerjem velikosti A3, v ločljivosti 1200 pik na palec. Tako majhno število kotov gledanja, kot način skeniranja nudijo slabše pogoje za rekonstrukcijo z grajenjem strukture iz gibanja, vendar smo lahko kljub temu izdelali dva detajlna modela višin: prvega iz fotografij v polni ločljivosti in drugega iz fotografij zmanjšane ločljivosti (prevzorčeni na ločljivost 800 pik na palec). Kljub enakim nastavitvam algoritma so stavbe na drugem modelu bolje definirane. Ker rekonstrukcija trenutno temelji le na letalskih fotografijah v eni smeri, lahko z razširitvijo nabora fotografij iz drugih smeri izračunana modela še izboljšamo. Vendar že taka kot sta, dajeta dobre obete za retrospektivo strukture naselja, tako posameznih hiš kot ulične mreže med njimi.

4 Izdelava kart pretekle rabe tal z objektno usmerjeno klasifikacijo

Daljinsko zaznavanje je razvilo tehnologijo in različne metode za stroškovno učinkovito kartiranje pokrovnosti zemeljskega površja na velikih območjih. Ključni dejavnik za razpoložljivost in zanesljivost teh kart za uporabo v okoljskih znanostih je razvoj učinkovitih postopkov za analizo in klasifikacijo satelitskih in zračnih posnetkov. Glavni cilj klasifikacije ali razvrščanja je odkrivanje in razvrščanje elementov (geografskih objektov in pojavov) na zemeljskem površju. Objektno usmerjena klasifikacija je način semantične slikovne analize, ki je bila osnovana za prepoznavanje in razvrščanje elementov geografske stvarnosti na zelo visoko ločljivih satelitskih in letalskih posnetkih (Blaschke, Lang in Hay, 2008). S postopkom pridobimo poljubne entitete pokrovnosti (lastnosti površja), ki so na posnetku razpoznavne in tvorijo/odslikavajo geografski prostor. Postopek sestavlja več korakov: segmentacija (tvorjenje enovitih območij), razvrščanje oz. klasifikacija (semantično združevanje segmentov v ciljne razrede) ter preverjanje. Rezultat je klasificiran geografski prostor, glede na naravne elemente, pokrovnost ali rabo tal, v obliki karte stanja. Uporaba postopka na posnetkih iz različnih časov nadalje omogoča časovne primerjave stanja ter kvantitativno in kontekstualno sledenje sprememb v pokrajini.

Klasifikacijo rabe tal smo izvajali na letalskih posnetkih Geodetske uprave Republike Slovenije, in sicer iz arhivov lastnih snemanj (od leta 1971) ter iz programa cikličnega aerosnemanja Slovenije (CAS), ki se redno izvaja na tri leta vse od leta 1985. Stanje smo spremljali v treh časovnih presekih: 1976, 1998 in 2011. Spremljali smo spremembe v stanju gozda, njiv, travnikov, cest ter naselij. Navkljub dejstvu, da so bili posnetki zajeti v različnih obdobjih, na različne načine, z različnimi inštrumenti in so zato razlikujejo po lastnostih in kakovosti (ostrina, barvni prostor, ločljivost, raven vidnega detajla), smo z izbranim postopkom dobili natančne karte stanj ter časovno primerljive rezultate. Najbolj očitne spremembe so nastale v razredu

grajeno, kjer spremljamo spremembe v legi in strukturi naselij na tem območju ter vsesplošen porast gozdnih površin na račun zaraščanja pašnikov in od naselij bolj oddaljenih njiv. Rezultati tovrstne obravnave zgodovinskih letalskih fotografij nam tako podajajo karte stanja rabe tal v različnih obdobjih, poleg tega pa omogočajo tudi izmero površin in pridobivanje kvantitaivnih ocen raznoterih sprememb. Upoštevaje, da arhivi zgodovinskih letalskih posnetkov hranijo fotografije stare tudi več kot sto let, se odpirajo številne nove možnosti proučevanja prostora, zgodovinskih dogodkov ter spremljanja procesov tako v naravi kot urbanih predelih v zadnjem stoletju.

5 Zaključek

Breginj, odročno obmejno hribovsko naselje, ki je bilo v potresih 1976 precej poškodovano, zato so ga na novo zgradili na drugi lokaciji, je edinstven primer uničenja pomembnega spomenika naše kulturne dediščine. Starega Breginja ne moremo več fotografirati. S študijo različne uporabe in obdelave zgodovinskih letalskih fotografij sva pokazala, da lahko stare fotografije iz zraka služijo za trirazsežno rekonstrukcijo starega vaškega jedra Breginja in tako prispevajo nove poglede k upravljanju s kulturno dediščino in povezanimi storitvami. S sodobnimi tehnologijami in programi za obdelavo slikovnih podatkov se odpirajo nove možnosti za obnovitev spomina na ta kraj, kot je nekdaj bil. Z metodo grajenja strukture iz gibanja sva iz šestih letalskih posnetkov s prekrivanjem v eni smeri, izdelala trirazsežen model starega vaškega jedra – prvi približek digitalne rekonstrukcije starega Breginja. V prihodnosti želiva pozornost usmeriti v integracijo različnih virov slikovnih podatkov, predvsem razširiti nabor zgodovinskih letalskih posnetkov ter fotografij iz tal ter jih smiselno vključiti v model. Izboljšave so najbolj pričakovane v vsebinski in geometrični izpopolnjenosti modela naselja. V primeru uspešne nadgradnje modela s teksturo pa bi se že zelo približali pravi 3R rekonstrukciji starega Breginja, ki bi lahko popestrila tudi vsebine muzejev in spomeniškovarstvenih centrov.

Uspešna retrospektiva pokrajinskih sprememb iz zgodovinskih letalskih fotografij pa nakazuje, da lahko z uporabo sodobnih postopkov za obdelavo slik spremljamo kompleksen proces zgodovinskega odtisa dogodkov v pokrajini. To nadalje dokazuje, da so zgodovinske letalske fotografije, tudi povsem različnih izvorov, namenov ter lastnosti, lahko pomemben in poseben vir informacij za geografe, gozdarje, biologe, zgodovinarje, arheologe in druge. Mednarodni arhivi zgodovinskih letalskih fotografij hranijo fotografije, ki so jih resda snemale vojaške službe v preteklosti. Danes so te fotografije dostopne in nam omogočajo vpogled v naselja in pokrajine za sto in več let nazaj. S primerom pokrajinske analize Breginja v zadnjih štirih desetletjih sva želela pokazati osnovne zmožnosti tovrstne slikovne retrospekcije pokrajine, kjer je naravna nesreča in posledični popotresni ukrepi zaznamovali razvoj območja – možnosti opazovanja je seveda mnogo več.

6 Literatura

Aljaž Celarc in Tea Erjavec. 2012. Breginjski kot. V: D. Kladnik, ur., *Slovenija VI. Vodniki Ljubljanskega geografskega društva*, str. 25–42. Založba ZRC, ZRC SAZU, Ljubljana.

Jon Grobovšek. 2002. *Preventivno fotogrametrično snemanje gradu Snežnik v okviru nacionalnega projekta 'IZMERE'*. Geodetski vestnik 46-4. Ljubljana.

Mojca Kosmatin Fras. 1996. *Architectural photogrammetry in heritage preservation - a description of methods and products*. Vestnik / Zavod RS za varstvo naravne in kulturne dediščine.

Primož Pipan. 2011. *Primerjava popotresne obnove v Italiji in Sloveniji po potresih v Zgornjem Posočju in Furlaniji*. Doktorsko delo. Koper.

Radovan Lipušček. 1995. Breginj. *Krajevni leksikon Slovenije.* Ljubljana, DZS.

Thomas Blaschke, Stefan Lang, Geoffrey J. Hay (uredniki). 2008. *Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications*. Lecture Notes in Geoinformation and Cartography. Springer.

Gradnja in analiza petjezičnega korpusa podnaslovov govorov TED*

Miha Helbl,† Žiga Domevščck‡

† Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
† miha.helbl@gmail.com, ‡ ziga.domevscek@gmail.com

Povzetek

Namen članka je preučevanje funkcionalnosti prevoda slovenskih prostovoljnih prevajalcev za nepridobitno organizacijo TED, ki niso nujno strokovno izobraženi prevajalci, temveč zgolj ljubitelji ali poznavalci teme govora. Analizirali smo odstopanja med angleškimi izvirniki in njihovimi slovenskimi prevodi. Analiza obsega pregled prevajalski strategij posploševanja, konkretizacije, zamenjave samostalnika z zaimkom ter izpustov znotraj slovenskih in angleških segmentov izbranih reprezentativnih govorov. Za namen raziskave so v petjezični korpus vključene izvirne angleške transkripcije ter slovenski, nemški, italijanski in francoski podnaslovi analiziranih govorov. Pričakujemo, da bodo prevajalci za doseganje ustreznosti prevodnih rešitev pogosto nezavedno uporabljali vrsto različnih prevajalskih strategij.

Compilation and Analysis of a Multilingual Corpus of TED Subtitles

The purpose of this article was to analyse the functionality of the Slovenian translations made by volunteers for the non-profit organization TED. These volunteers are not necessarily educated as translators, but are often fans or experts on the talk topic. We analysed the discrepancies between the English originals and their Slovenian translations. The focus was on the following translation strategies: generalization, specification, substitution of the noun with a pronoun, and omissions. The multi lingual corpus, created for the purpose of this research, consists of the original English transcripts and the corresponding Slovenian, French, Italian and German subtitles. We expect that the translators, in order to achieve the suitable translational solution, will often unknowingly use a variety of translational strategies.

1 Uvod

Projekt gradnje in analize petjezičnega korpusa podnaslovov govorov TED je bil zasnovan z namenom raziskave prevedene vsebine s strani prostovoljnih prevajalcev, ki opravljajo delo za nepridobitno organizacijo TED. Želeli smo preučiti funkcionalnost prevodov slovenskih prostovoljnih prevajalcev, ki niso zagotovljeno šolani prevajalci ali jezikoslovci, temveč so najpogosteje zgolj ljubitelji ali poznavalci teme določenega govora. V ta namen smo zbrali podnaslove 316 govorov v petih različnih jezikih. Glavni kriterij izbora je bil, da so podnaslovi prevedeni v slovenski jezik iz angleškega izvirnika. Pri analizi smo uporabili samo slovenske in angleške podnaslove. Želeli smo prikazati razhajanja med izvirnimi angleškimi transkripcijami govorov TED in njihovimi slovenskimi prevodi na podlagi odstopanja pri izpuščanju in zgoščevanju. Z analizo teh odstopanj smo želeli izpostaviti najpogostejše sporne segmente pri prevajanju in pretvorbi transkripcij v podnaslove.

Nepridobitna organizacija TED je bila ustanovljena leta 1984 s ciljem, da bi širila ideje na področjih tehnologije, zabave in oblikovanja (Technology, Entertainment, Design). Danes njene konference pokrivajo skoraj vsa področja od znanosti do gospodarstva in do globalnih vprašanj in so od junija 2006 na voljo za sprotno gledanje prek interneta. Spletni arhiv TED omogoča ogled govorov s podnaslovi, ki jih prispevajo prostovoljni prevajalci. Skupno si je od januarja 2016 na spletišču TED možno ogledati govore s podnaslovi v 110 jezikih, ki jih je pomagalo prevesti 16.114 prostovoljnih prevajalcev. Tako imenovani Projekt odprtega prevajanja (Open Translation Project) je TED vzpostavil na pobudo obiskovalcev spletišča, ki so želeli TED govore deliti s svojimi prijatelji in družinskimi člani in so organizatorje vprašali, če lahko oni sami prevedejo te govore in jih opremijo s podnaslovi. Organizacija TED je v ta namen vzpostavila sistem, ki je omogočil prostovoljcem prevajanje njihovih najljubših govorov v poljuben jezik. V času opravljanja raziskave je bilo del prevajalske ekipe tudi 106 prevajalcev za slovenščino, ki so prevedli več kot 300 govorov.[1]

Na svojem spletišču so organizatorji objavili tudi smernice za vse prevajalce, da bi tako uskladili podnaslove v vseh jezikih. Prva in hkrati osrednja točka smernic je sodelovanje najmanj dveh prostovoljcev pri prevajalskem projektu – prevajalca in lektorja. Druga zajema problematiko podnaslavljanja in prostovoljcem predstavi standarde glede dolžine podnaslovov. Tretja

* Projekt gradnje korpusa je potekal v okviru magistrskega modula Korpusi in baze podatkov na Oddelku za prevajalstvo FF UL (štud. l. 2015/2016) pod mentorskim vodstvom doc. dr. Darje Fišer. Besedilo je nastalo pri predmetu Slovensko strokovno besedilo pri prof. dr. Vojku Gorjancu.

[1] TED. https://www.ted.com/people/translators?sort=translations&country=slovenia (Dostop 31. 1. 2016)

pa zajema slog TED, ki je osrednji slogovni del smernic in med drugim spodbuja uporabo neformalnega, sproščenega in vsebinsko ter jezikovno natančnega sloga prevajanja. Prevajalci imajo poleg smernic v besedilu na voljo tudi vodič v video formatu.

Tak način prevajanja spada med tako imenovano množičeno prevajanje (crowdsourced translation). Množičeno prevajanje izkoristi interes, ki združuje člane prevajalskih projektov. Ti člani sodelujejo pri prevajanju, lektoriranju, zbiranju idej, virov in gradnji pomnilnikov prevodov. Množičenje takšnim skupinam odpira možnosti za nudenje nadaljnjih prevajalskih storitev, ki so lahko ali plačana ali pa delujejo na prostovoljski ravni.

V članku je predstavljen pregled področja in raziskave, ki so proučevale podnaslovno prevajanje. V teoretičnemu delu nato sledi zasnova projekta, prikaz najpogostejšega besedja in analiza prevodov.

2 Pregled področja in sorodne raziskave

Delia Chiaro (2013) označi avdiovizualno prevajanje kot prenos verbalnih prvin, ki jih nosijo avdiovizualna dela in produkti (filmi, opere, televizijske oddaje, muzikali, računalniške igre itn.) iz enega jezika v drugega. Razlikuje med dvema glavnima načinoma avdiovizualnega prevajanja – sinhronizacijo in podnaslavljanjem. Podnaslovi so skrajšani zapisani prevodi tistega, kar je moč slišati na ekranu. S takim izhodiščem Mona Baker (1992) definira sedem prevajalskih strategij – posploševanje, nevtralizacija, kulturna substitucija, kalkiran prevod z razlago, parafraziranje, izpust in prevod s sliko – od katerih sta za našo analizo pomembna predvsem posploševanje in izpust. Mona Baker označuje posploševanje kot eno od najsplošnejših strategij pri prevajanju, kjer ni ustreznic v ciljnem jeziku. Ta strategija je znotraj večine jezikov enako učinkovita, saj hierarhična struktura semantičnih polj ni vezana na določen jezik (Baker, 1992: 26). Če pomen določenega izraza ni ključen za razvoj nadaljnjega besedila in bi njegova razlaga zgolj zmotila bralca, se prevajalci pogosto lahko odločijo za izpust prevoda besede ali izraza (Baker, 1992: 40).

Na slovenskem področju je Urša Vogrinc Javoršek (2007) izvedla raziskavo, kjer je proučevala zgolj metabesedilne elemente v igrani ameriški nanizanki Midve z mamo (Gilmore girls). Metabesedilni elementi so tisti deli, ki ne prispevajo k vsebini besedila, temveč pomagajo gledalcu vsebino bolje in lažje organizirati ter razložiti (Vogrinc Javoršek po Halliday, 1985: 35). Izsledki raziskave so pokazali, da so podnaslovi prevodi igrane nanizanke manj kohezivni kot izvirnik. To drži, če upoštevamo zgolj verbalni vidik, a podnaslovi nastanejo na podlagi dokončanega avdiovizualnega medija, kjer je potrebno upoštevati tudi zvok in sliko. Prevajalec tako upošteva dejstvo, da informacije ne prihajajo zgolj skozi en kanal. Dopušča, da se določeni elementi, ki se pojavijo v drugih kanalih, ne obdržijo v podnaslovu (Vogrinc Javoršek, 2007). Irena Kovačič je več kot desetletje pred tem (1995) izpostavila, da današnje norme v podnaslovih zelo očitno odsevajo odnos do jezika in še posebno do registra, ki ignorira najdbe v moderni lingvistiki, zlasti na področju razlik med pisnim in govornim jezikom ali funkcionalnimi vrednostmi različnih jezikovnih struktur. Razlogi za to so očitni: podnaslovi so prikazani v pisni obliki, ki je že po naravi bolj organizirana in bolj zgoščena kot govor, tako je idealna za podnaslavljanje, kjer je redukcija oblik ena osnovnih zahtev. Pisni jezik je relativno standardiziran, medtem ko govorni jezik ni. Prevajalci se morajo ne le odločiti, kako prevesti besedilo kot celoto in kako prenesti elemente v ciljni jezik, ampak v prvi vrsti se morajo tudi odločiti, kaj sploh prevesti in kaj izpustiti (Kovačič, 1996). Te odločitve so tesno povezane z dvema določenima lastnostma podnaslavljanja: a) potreba po zgoščevanju sporočila zaradi omejenega prostora, b) prenos iz govorjenega v pisni diskurz. Kovačič po Hallidayju definira tri funkcije zgoščevanja – (med)osebno, besedilno in ideacijsko. Ideacijska funkcija je od naštetih najbolj odvisna od jezika, medosebna je pogosto replicirana v neverbalni in nejezikovni interakciji. Besedilna funkcija pa postane pri podnaslovih drugotnega pomena, ker slika poskrbi za kontinuiteto in kohezijo. Prevajalska strategija je odvisna od vrste programa in pričakovanega ciljnega občinstva. Za ciljno občinstvo govorov TED se s strani organizacije predvideva, da imajo sposobnost hitrejšega branja, širše besedišče, je že seznanjeno s tematiko ali pa pozna ozadje govorov. Zato so podnaslovi lahko gostejši, vsebujejo več dobesednih izrazov, hkrati pa morajo izraziti ne zgolj zgodbo ampak tudi osebnosti govorcev ter psihološke odnose in pozicije moči med osebami (Kovačič, 1996).

Pri podnaslovnem prevajanju prihaja do kompromisa med normami pisanega in govorjenega jezika pri čemer se prevod in priredba zlijeta in potekata istočasno. Pri tem naj bi se prevajalci trudili, da uporabijo polno strukturo, če je le mogoče (Halliday, 1994). Organizacija TED se v svojih smernicah za prevajalce manj ukvarja s strukturo kot s pomenom, saj med drugim narekuje da naj prevajalci kljub časovnim in prostorskim omejitvam predvsem natančno posredujejo pomen govorjenega besedila. Tu moramo opozoriti na to, da podnaslovi govorov TED tako po času trajanja in dolžini posamezne vrstice odstopajo od ustaljenih norm, ki se uporabljajo pri televizijskih in kinematografskih podnaslovih pri nas. Prav tako odstopajo po najvišjem dovoljenem številu znakov na sekundo. Tako TED v svojih smernicah prevajalcem[2] narekuje, da je v posamezni vrstici dovoljenih največ 42 znakov, pri čemer prevajalec ne sme vstaviti več kot dveh vrstic. Najvišja dovoljena bralna hitrost prikazovanja podnaslovov je po normah njihovih smernic 21 znakov na sekundo. Slovenski standardi za

[2] TED. https://www.ted.com/participate/translate/guidelines. (Dostop 31. 1. 2016)

podnaslavljanje, ki jih je tudi moč zaslediti pri nekaterih televizijskih hišah in prevajalskih agencijah (na primer prevajalski studio Milenko Babič, s. p.), v povprečju narekujejo največ 36 znakov na posamezno vrstico in bralno hitrost 12 do 15 znakov na sekundo. Ti standardi se lahko razlikujejo med naročniki.[3]

Da pa se prevajalci lahko držijo zgoraj omenjenih prostorsko-časovnih omejitev, morajo prevodno besedilo ustrezno prilagoditi. Pri tem si pomagajo z zgoraj omenjenimi prevajalskimi strategijami. Mi smo se pri naši analizi osredotočili na posploševanje, konkretizacijo, zamenjave samostalnika z zaimkom ter izpuste, ki smo jih razdelili na tri sklope: vsebinske izpuste, izpuste diskurzivnih označevalcev in izpuste ponavljanj. Odločitev za opazovanje diskurzivnih označevalcev utemeljujemo z njihovo naravo, saj so to jezikovna sredstva, ki imajo v prvi vrsti pragmatično vlogo (niso tako pomembni za posredovanje informacij ter vsebine, ampak razvijajo medosebni odnos sogovornikov pri organizaciji poteka diskurza (Verdonik, 2006), označujejo prehode med tematskimi sklopi (prav, okej, no, dobro, v redu …), prehod v zaključek govora (dobro, okej, prav, torej), popravljanje ali druge spremembe v strukturi izjave ... Gre torej za jezikovne elemente, za katere lahko glede na napotke pri TED-u, pa tudi zaradi omejitve pri prostoru in povezave med sliko in podnaslovom, pričakujemo, da jih bodo prevajalci v večji meri izpuščali.

3 Zasnova projekta

Projektno delo je obsegalo gradnjo korpusa, pridobivanje metapodatkov in dve jezikoslovno-prevodni analizi. Za osnovo korpusa smo zbrali vse slovenske podnaslove na spletišču TED, ki jih je bilo v času pisanja članka 316. Prenesli smo slovenske, angleške, nemške, francoske in italijanske podnaslove za vsakega od 316 govorov. Vse govore smo zbrali v arhiv in jih razporedili glede na leto nastanka. Vsi podnaslovi so vsebovali tudi časovne kode, ki nas pri gradnji in analizi korpusa niso zanimale, zato smo iz njih ustvarili ločen arhiv.

Hkrati smo na spletišču TED za vseh pet jezikov izpisali tudi metapodatke, ki bi jih vključili v poseben korpus. Med te podatke smo vključili naslov govora, ime govorca, prevajalca, lektorja, opis govora, leto in mesec konference ter ime konference na kateri je bil govor predstavljen. Metapodatki bodo služili kot dodaten referenčni podatkovni material pri uporabi korpusa.

Ustvariti smo želeli paralelni korpus, ki vsebuje poravnana besedila v več jezikih. Za poravnavo podnaslovov smo uporabili orodje LF Aligner. V orodje smo za vsak posamezen govor naložili podnaslove v vseh petih jezikih v obliki .TXT. Program nam je podnaslove poravnal in ustvaril en sam Excelov dokument z vsemi petimi podnaslovi, poravnanimi horizontalno, ter tudi pomnilnik prevodov v datotečni obliki .TMX. Postopek smo ponovili za vseh 316 govorov. Pomnilnike prevodov smo naložili na odprto korpusno spletišče SketchEngine, kjer smo paralelni korpus ustvarili za vseh pet jezikov. Zaradi omejenosti dostopa do korpusa v času raziskave smo bili s paralelnim korpusom omejeni zgolj na milijon besed. V tistem času še nismo vedeli, da bi lahko prosili za povečanje pomnilnika in bi tako lahko naložili celotno zbrano zbirko besedil – od takrat je bil pomnilnik povečan na štiri milijone besed in je že pripravljen na razširitev raziskave. Korpus smo zapolnili s 131 pomnilniki slovenskih in angleških prevodov. Ker so govori nastali v različnih letih – od leta 2004 do 2015 – smo za korpus naredili selektiven izbor govorov glede na nastanek. Kronološko smo tako izbrali prvih 30 govorov v arhivu, preskočili naslednjih 60, ponovno vključili naslednjih 30 govorov in postopek ponavljali dokler nismo korpusa napolnili do omejene količine podatkov, pri čemer smo pazili tudi na zastopanost prevodov iz različnih časovnih obdobij.

Nastali korpus smo najprej analizirali glede na obseg celotnega korpusa, na število pojavnic za posamezni jezik in opravili pregled, v okviru katerega smo opazovali najpogostejše polnopomenske besede v slovenščini in angleščini, v nadaljevanju pa smo se osredotočili na analizo prevodnih rešitev. Pri tem pa je bil naš glavni namen pregled odstopanj med izvirnikom in prevodom zaradi izpustov in posploševanja besedila. Izpuste smo pri analizi razdelili na tri sklope: izpuste diskurzivnih označevalcev, izpuste ponavljanja in vsebinske izpuste. Končni cilj je bil določiti funkcionalnost prevodov s strani prostovoljnih prevajalcev, ki morebiti niso vedno izobraženi prevajalci. Pri analizi smo se omejili na govore, kjer so bila največja odstopanja pri številu segmentov. Opazili smo namreč, da je prišlo do bistvenih razlik v številu segmentov le pri omejenem številu govorov. Kot ločen segment smo opredelili en podnaslov, ki je lahko tako enovrstičen kot dvovrstičen.

Že pred začetkom analize smo pričakovali, da bo zaradi prostovoljne narave prevajalskega projekta pri organizaciji TED kljub natančnim navodilom s strani organizacije prihajalo do odstopanj pri številu segmentov slovenskih podnaslovov. Predpostavljali smo tudi, da bo v primerjavi z angleškim izvirnikom v slovenskih podnaslovih pogosteje prihajalo do izpuščanja besedila kot dopolnjevanja, saj je besedilo kognitivno lažje izpuščati kot pa dodajati.

4 Analiza

Analiza korpusa je sestavljena iz dveh delov. Najprej smo opravili krajši pregled najpogostejšega besedišča v zgrajenem vzporednem korpusu govorov TED za slovenski in angleški jezik v konkordančniku SketchEngine. Med analizo korpusa smo morali rešiti dilemo, kako ravnati, saj slovenski nabor besedil ni bil jezikovno označen (med opravljanjem raziskave nismo bili seznanjeni s slovenskim označevalnikom). Zaradi

[3] Lektorsko društvo Slovenije. http://www.lektorsko-drustvo.si/predavanja/na-istem-bregu. (Dostop 31. 1. 2016)

tega ni bilo možno izvesti popolnoma reprezentativnega prikaza slovenskega korpusa. Slovenski prikaz najpogostejšega besedišča namreč ne odseva najpogostejših pojavnic, temveč najpogostejše različnice. Tako smo poiskali deset najpogostejših slovenskih različnic (samostalnike, glagole, pridevnike in prislove), ki so se pojavile na frekvenčnem seznamu in smo opravili prevodno primerjavo z angleškimi pojavnicami. Prikaz angleškega korpusa pa je bolj reprezentativen, saj je za razliko od slovenskega korpusa označen in smo zato zbrali deset najpogostejših pojavnic, ki smo jih tudi označili glede na besedno vrsto.

Sledila je analiza prevodov, kjer smo opazovali najpogostejše tipe uporabljenih prevajalskih strategij, hkrati pa smo v sprotni diskusiji podali lastna domnevanja, zakaj je do teh posegov v besedilo med prevajanjem prišlo.

4.1 Prikaz najpogostejšega besedišča

V celotnem petjezičnem korpusu je 131 govorov za vsak posamezen jezik. Skupno teh 131 govorov v petih jezikih zajema 968.559 različnic, kar je zgolj tretjina celotnega korpusnega arhiva 316 govorov. Od tega smo se za analizo osredotočili na slovenski in angleški del. Skupno število slovenskih besed znaša 181.536 različnic oziroma 222.276 pojavnic, medtem ko število angleških besed znaša 226.935 različnic oziroma 270.741 pojavnic, skupna velikost slovensko-angleškega korpusa je potemtakem 408.471 besed. Že po končnem številu besed za posamezen jezik je razvidno, da je bilo v slovenskih podnaslovih veliko izpuščanja, kar zgolj potrjuje obe naši začetni hipotezi.

	Pojavnica	Besedna oblika	Število
1	so	RB	1629
2	people	NNS	866
3	do	VV	865
4	now	RB	703
5	just	RB	682
6	very	RB	634
7	really	RB	516
8	going	VVG	496
9	then	RB	465
10	know	VVP	463

Tabela 1: Najpogostejše angleške pojavnice.

	Različnica	Število
1	ljudje	816
2	stvar	505
3	smeh	396
4	videti	323
5	način	254
6	aplavz	235
7	leta	201
8	hvala	193
9	primer	136
10	svet	122

Tabela 2: Najpogostejše slovenske različnice.

Najpogostejše pojavnice v angleškem in različnice slovenskem jeziku so prikazane v tabelah. V tabeli z angleškim naborom (gl. Tabela 1) so v drugem stolpcu zapisane pojavnice, ki so razvrščene po pogostosti, tretji stolpec označuje besedno obliko pojavnic v konkordančniku SketchEngine[4], četrti stolpec pa prikazuje absolutno frekvenco pojavnice. Najvišjo frekvenco pojavitev so pri tem imeli prislovi.

Seznam najpogostejših polnopomenskih angleških pojavnic smo poiskali s pomočjo funkcije *word list* znotraj konkordančnika, kjer smo lahko besede razdelili glede na oznako. Iskanje najpogostejših polnopomenskih slovenskih različnic pa je tako rekoč potekalo ročno, saj kot že omenjeno, slovenski korpus v času opravljanja raziskave še ni bil označen. Nabor najpogostejših različnic je prikazan v Tabeli 2.

Po primerjavi zbranih najpogostejših pojavnic in različnic ter prevodni primerjavi besedil znotraj korpusa smo ugotovili, da odstopanja pri številu najpogostejših besed niso tako velika. Glavne polnopomenske besede imajo odstopanja do približno 100 besed, kar je deloma tudi posledica tega, da nismo zbrali celotnega nabora besed.

4.2 Analiza prevodov podnaslovov

Analizo prevodov smo razdelili na šest večjih sklopov. Pri vsakem sklopu bomo podali po pet primerov za določen pojav, hkrati pa bomo ob tem ugotavljali, zakaj je do teh odstopanj prišlo. Sklopi so razporejeni po pogostosti pojavitev v podnaslovih od najpogostejših do najredkejših.

Iz dvajsetih govorov je bilo skupno izpisanih 202 primera odstopanja. Najpogostejši so bili izpusti diskurzivnih označevalcev s 46 odstotki, sledili so jim vsebinski izpusti s slabimi 20 odstotki, posploševanje s 14 odstotki, ostale strategije pa so zamenjave samostalnika z zaimkom z 11 odstotki, ponavljanje s 5 odstotki in konkretizacija s 4 odstotki. Skupno je bilo izpustov 71 odstotkov, ostalih strategij pa 29 odstotkov.

Največ odstopanja pri slovenskih in angleških podnaslovih je bilo pri izpuščanju diskurzivnih označevalcev v slovenskih podnaslovih. Do tega je prišlo zaradi razlikovanj v primarni funkciji besedila, saj je angleški del bolj transkripcija govora, slovenski pa je dejanski podnaslov. Poleg tega so diskurzivni označevalci značilni za govor, saj povezujejo

[4] SketchEngine. https://www.sketchengine.co.uk/penn-treebank-tagset/. (Dostop 31. 1. 2016)

posamezne segmente le-tega, imajo zelo nizko informacijsko vrednost in so pri krajšanju besedila za podnaslov prvi, ki odpadejo. Prevod označevalcev ni prioritetni, kar je lepo razvidno iz prvega primera, kjer je *well,* slovensko *torej* izpuščen, saj gledalcu ne poda nove informacije (gl. Tabela 3).

Angleščina	Slovenščina
Well it's very simple	To je enostavno.
And then finally – you know, /…/	Na koncu --najboljša možnost-- /…/
Now, when psychologists show you bars, /…/	Ko vam psihologi kažejo stolpce, /…/
So thank you very much.	Najlepša hvala.
»Oh, I created this great product, /…/«	»Ustvaril sem odličen produkt, /…/«

Tabela 3: Primeri izpuščanja diskurzivnih označevalcev.

Drugi najpogostejši način krajšanja besedila so bili vsebinskih izpusti, kjer je gledalcu zamolčan del informacij, ki lahko vplivajo na razumevanje podnaslova. To je dobro razvidno pri prvem primeru, kjer je pridevnik, ki določa tip očal, zamolčan. Morda je bil to zgolj lapsus ali pa so *varnostna* očala bila predolga za podnaslov (gl. Tabela 4).

Angleščina	Slovenščina
/…/ a pair of safety glasses, /…/	/…/ ker lahko s parom očal, /…/
/…/ extra-virgin olive oils /…/	/…/ vrst ekstra deviškega olja /…/
My goal is to better understand these violent actors /…/	Ta nasilna gibanja želim bolje razumeti /…/
/…/ that they can actually have experiences in their heads /…/	/…/ da lahko dejansko doživlja dogodke /…/
/…/ from your life.	/

Tabela 4: Primeri vsebinskega izpuščanja.

Posploševanje je tretji najpogostejši pojav krajšanja besedila. Do njega pride takrat, ko je v izvirniku opisana točno določena stvar, ki morda naši kulturi ni najbližja ali pa bi bil njen prevod preprosto predolg. To lepo ponazarja prvi primer, kjer v izvirniku rečejo *pop quiz*, kar pomeni nenapovedan preizkus znanja, slovenski prevajalec pa se odloči zgolj za *preizkus znanja*, saj bi bil točen prevod predolg (gl. Tabela 5).

Angleščina	Slovenščina
You failed the pop quiz, and you're hardly five minutes into the lecture.	To je enostavno. Po petih minutah predstavitve ste že padli na preizkusu znanja.
(Laughter) Because she was the main reason we were leaving the country.	Na koncu --najboljša možnost-- ker je bila ona glavni razlog za naš odhod.
and it's not because they whipped some up, tried it and went, "Yuck."	Ko vam psihologi kažejo stolpce, Ne zato, ker bi ga kdo naredil, poizkusil in se zgrozil.
so that they can feel better about the worlds in which they find themselves.	/…/ zato, da se lahko bolje počuti v danih okoliščinah.
Like Sir Thomas, you have this machine.	Kot g. Thomas imate tudi v to sposobnost.

Tabela 5: Primeri posploševanja.

Pri zamenjavah samostalnika z zaimkom, ki je po pogostosti na četrtem mestu, gre za podajanje že znanih informacij v krajši obliki. Pri prvem primeru že vemo, da je govora o profesorjih, zato tega ni potrebno ponovno napisati. Podnaslov razumemo, hkrati pa ga tudi preberemo hitreje (gl. Tabela 6).

Angleščina	Slovenščina
There's something curious about professors in my experience -- /…/	Nekaj zelo zanimivega je na njih po mojih izkušnjah -- /…/
But if I do this with amnesiac patients, /…/	Na koncu --najboljša možnost-- Če pa to naredim pri naših bolnikih, /…/
He has been working on the topic for 20 years.	Ko vam psihologi kažejo stolpce, Na tem je delal že 20 let.
We need to know what makes These organisations tick.	Vedeti moramo, kaj jih vodi.
On the morning of the hemorrhage, I could not walk, talk, read, write or recall any of my life.	To jutro nisem mogla hoditi, govoriti, brati, pisati ali se spomniti česarkoli v svojem življenju.

Tabela 6: Primeri zamenjave samostalnika z zaimkom.

Ponavljanje, peti najpogostejši način krajšanja, se pri podnaslovih zaradi časovno-prostorskih omejitev večinoma izpusti, saj je poglavitna funkcija podnaslovov podajanje informacij in ne ponazarjanje

stila govora govorca. Krajše je boljše. Pri prvem primeru vidimo, da se *miselni procesi* pri drugi omembi izpustijo (gl. Tabela 7).

Angleščina	Slovenščina
A system of cognitive processes, largely non-conscious cognitive processes, that help them change	Sistem miselnih procesov, večinoma nezavednih, /…/
Third: make somebody else really, really rich.	Tretje: Nekoga drugega naredite zelo bogatega.
/…/ where a tiny, tiny fraction of the world /…/	/…/ kjer se lahko le zelo majhen del sveta /…/
WK: In Malawi, Kasungu. In Kasungu.	WK: V Malaviju v mestu Kasung.
You see the thing is, the thing is /…/	Vidite, stvar je v tem /…/

Tabela 7: Primer izpuščanja ponavljanj.

Pri konkretizaciji, ki je na zadnjem mestu po pogostosti, gre predvsem za jasnejše sporočanje izvirnika. Najbolj ilustrativen je četrti primer, kjer je slovenski prevajalec zamenjal izvirni angleški *he* s točno določenim samostalnikom, ki podrobneje opisuje njega, torej *kolega* (gl. Tabela 8).

Angleščina	Slovenščina
Here's two different futures that I invite you to contemplate.	To je enostavno. Razmislite o dveh različnih življenjskih poteh, /…/
from the one they like the most, to the one they like the least.	Na koncu --najboljša možnost-- po vrsti, glede na to, koliko so mu všeč.
/…/ from the one you like the most to the one you like the least."	Ko vam psihologi kažejo stolpce, slike glede na to, koliko so vam všeč."
So he recognizes that I need help and he gets me help.	Najlepša hvala. Kolega dojame, da potrebujem pomoč in mi jo pošlje.

Tabela 8: Primeri konkretizacije.

5 Zaključek

Petjezični korpus sestavljajo prevodi v slovenščini, nemščini, francoščini, italijanščini ter angleško izvirno besedilo. Med analizo prevodov smo prišli do zaključkov, da se prostovoljni prevajalci za TED med prevajanjem sicer pogosto poslužijo strategij posploševanja, konkretizacije, zamenjave samostalnika z zaimkom ter izpustov, toda ne v tolikšni meri, kot smo prvotno pričakovali. K temu so v večji meri pripomogla prevajalcem ustrezno predstavljena navodila na spletišču TED ter dobro zamišljen sistem sodelovanja več ljudi pri posameznih prevajalskih projektih – prevajalec, lektor in tudi koordinator posamezne prevajalske skupnosti.

Pri analizi prevodov je bila najpogostejša prevajalska strategija izpust diskurzivnega označevalca, na drugem mestu pa je bil vsebinski izpust. Sledili so jim posploševanje, zamenjava samostalnika z zaimkom, izpust ponavljanja in konkretizacija. Iz tega je razvidno, da se prostovoljni prevajalci, tudi tisti, ki niso strokovno izobraženi prevajalci, poslužujejo velikega števila ustreznih prevajalskih strategij za podnaslavljanje. Ni pa jasno ali se teh strategij poslužujejo zavedno ali nezavedno, kar odpira možnosti za nadaljnje raziskave v tej smeri.

Izdelava tovrstnih specializiranih korpusov vsekakor odpira mnogo možnosti za prevajalsko stroko, saj bi lahko služili kot jezikoslovni referenčni vir za prevajalce na posameznih področjih. Že ta korpus odpira možnosti za boljše usklajevanje prevodov celotne ekipe prevajalcev za TED, saj jim daje vpogled v bazo podatkov, ki je usmerjena zgolj za njihovo področje. Vsekakor pa obstaja priložnost, da se korpus dopolni s preostalimi govori in ostalimi prevodi govorov, saj je bila v korpus bila vključena le dobra tretjina vseh zbranih govorov za angleški in slovenski jezik.

6 Literatura

Mona Baker. 1992. *In Other Words. A Course in Translation*. London and New York: Routledge.

Delia Carmela Chiaro 2012. *Audiovisual Translation. The Encyclopedia of Applied Linguistics*. http://www.researchschool.org/documents/Chiaro_Audiovisual%20Trl.pdf. (Dostop 31. 1. 2016)

M. A. K. Halliday 1985. *An introduction to functional grammar*. London, Caulfield East, Baltimore. Edward Arnold.

M. A. K. Halliday 1994. *An introduction to functional grammar, 2nd ed*. London. Arnold.

Irena Kovačič. 1995. *Reinforcing or changing norms in subtitling*. V: C. Dollerup in V. Appel (eds) Teaching Translation and Interpreting III. Amsterdam. Benjamins, str. 105–109.

Irena Kovačič. 1996. Subtitling strategies: a flexible hierarchy of priorities. V: C. Heiss in R.M. Bollettieri Bosinelli (eds) *Traduzione multimediale per il cinema, la televisione e la scena Bologna*, Clueb, str. 297–305.

TED http://www.ted.com/talks (Dostop 17.1.2016)

Darinka Verdonik. 2006. *Diskurzivni označevalci v telefonskih pogovorih*. Ljubljana. Slavistično društvo Slovenije.

Urša Vogrinc Javoršek. 2007. *Metabesedilnost v okviru strategij podnaslovnega prevajanja*. Slavistično društvo Slovenije, letnik 52, številka 3/4, str. 79–94.

#Analiza novih komunikacijskih elementov na družbenem omrežju @Twitter

Katerina Pertot, Maja Petrovčič, Nika Strojan

Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva cesta 2, 1000 Ljubljana maja.petrovcic86@gmail.com,
nikica.spika@gmail.com,
kpertot@gmail.com.

Povzetek

V pričujočem prispevku so navedeni izsledki analiz novih komunikacijskih elementov na družbenem omrežju Twitter (heštegi, omembe uporabniških imen in izrazov čustev - emojiji in emotikoni), ki smo jih opravili za slovenski in angleški del korpusa spletnih uporabniških vsebin Janes. V prispevku je podrobneje opisan potek analiz ter kriteriji za analize. Analize so bile izvedene na podlagi frekvenčnih seznamov in naključnih vzorcev rabe, ki smo jih pridobili v korpusu Janes.

#Analysis of New Communication Elements on the @Twitter Social Network

This paper reports on the analysis of new communication elements used on the Twitter social network (hashtags, mentions and emotional expressions - emojis and emoticons). The analysis was conducted with data, acquired from the English and Slovene parts of the corpus of user-generated Slovene called Janes. The analysis procedure and its criteria are also described in this article. The basis for our analysis were frequency lists and random usage samples from the Janes corpus.

1 Uvod

Jezik, ki ga uporabljamo na internetu, je drugačen od tistega, ki ga uporabljamo v vsakdanjem življenju. Na internetu smo namreč lahko anonimni, v vsakdanji komunikaciji pa to anonimnost zelo težko dosežemo, sploh če z nekom komuniciramo iz oči v oči. O preiskovanju jezika, ki ga uporabljamo na internetu, je v svoji knjigi pisal tudi David Crystal (2001, str. 5): »Naša naloga je preiskovati lingvistične značilnosti tako imenovane »elektronske revolucije« in podrobneje preučiti, ali se uporaba jezika na internetu res spreminja.«

V pričujočem prispevku se osredotočamo na jezik, ki ga uporabljajo uporabniki družbenega omrežja Twitter[1] , za katerega je značilna vrsta novih komunikacijskih elementov. Na Twitterju ima vsak uporabnik svoje uporabniško ime, ki je lahko njegovo lastno ali pa izmišljeno. Uporabniška imena se uporabljajo tudi v sporočilih pri omenjanju drugih uporabnikov (ang. mention). Za lažje sledenje objavam na Twitterju uporabniki uporabljajo t. i. heštege (ang. hashtag). Za izražanje svojih čustev in trenutnega počutja pa se uporabljajo izrazi čustev – emojiji (ang. emoji) in v zadnjem času vedno bolj pogosti emotikoni (ang. emoticon).

Kljub temu, da so tovrstni elementi v komunikaciji precej novi in so jih uvedli razvijalci družbenih omrežij, so med uporabniki postali zelo priljubljeni in se tudi hitro ustalili v računalniško posredovani komunikaciji. Vendar temeljitih jezikoslovnih raziskav razširjenosti, rabe in pomena omenjenih elementov še ni veliko niti za večje svetovne jezike, še toliko bolj pa to velja za slovenščino, ki je ponudniki aplikacij pri razvoju funkcionalnosti niso imeli v mislih. Zato nas je v okviru raziskave, ki jo predstavljamo v tem prispevku, zanimalo, kako omembe, heštege in čustvene elemente uporabljajo slovenski uporabniki družbenega omrežja Twitter, kadar tvitajo v slovenščini, za primerjavo pa smo analizirali tudi njihove objave v angleščini.

Pred analizo smo si postavili 3 hipoteze:

1. V angleškem delu korpusa bodo omembe večkrat uporabljeni v tujem jeziku kot v slovenskem.
2. Čustvene izraze (emojije in emotikone) bodo bolj pogosto uporabljale ženske kot moški.
3. Heštegi bodo v obeh delih korpusa večkrat enobesedni kot večbesedni.

2 Sorodne raziskave

Uporabniška imena na internetu so lahko drugačna od pravega imena osebe. Na ta način si oseba zagotovi neke vrste anonimnost, gradi svojo identiteto in izkazuje kreativnost. Uporabniška imena na Twitterju je proučeval J. Olivier (Olivier, 2014), ki ugotavlja, da jih je potrebno obravnavati kot posebno podkategorijo psevdonimov, ki se razlikujejo od vzdevkov. Ob tem poudarja, da uporabniška imena na Twitterju zapisujemo kot druga lastna imena, opaža pa tudi nekatere posebne značilnosti: vrstni red pri imenu in priimku je pogosto obrnjen, v imenih se lahko pojavljajo številke, velike začetnice pogosto niso upoštevane, v imenih se uporabljajo tudi drugi simboli. Kadar uporabnike omenjamo (ang. mention) v tvitih, pred uporabniško ime zapišemo znak @, s čimer je imetnik računa o omembi samodejno obveščen (Twitter, 2016). V naši raziskavi smo opazovali, kako omembe uporabniških imen slovenski uporabniki Twitterja uporabljajo v slovenskih in angleških tvitih.

Družbeni dejavniki, kot so spol, starost, stopnja izobrazbe, socialno-ekonomski razred in mnogi drugi vplivajo na jezik. Študije o jezikovni rabi na družbenih omrežjih so pokazale, da obstajajo razlike v rabi heštegov med ženskami in moškimi (Cunha, 2014). Poleg tega so v sorodnih raziskavah (Gotti) ugotovili, da je raba heštegov zelo pogosta, saj se v povprečju v več kot 50 % primerov tvitov pojavlja vsaj en hešteg. Gotti je v raziskavi proučeval tudi postavitev heštegov v tvitih. Rezultati so

[1] URL: https://twitter.com/

pokazali, da se je kar 87 % heštegov pojavilo v vlogi epiloga oziroma na kocu tvita.

Glede emotikonov in emojijev so sorodne raziskave opredelile emotikone (1) kot indikatorje čustev (npr. vesel ali žalosten), (2) kot nečustvene pomene (npr. kot šala), in (3) in kot znak (npr. :)) za ublažitev izreka. S povečano priljubljenost emojijev so začeli raziskovalci raziskovati vlogo emojijev v tekstovni komunikaciji. Ugotovili so, da z njimi ne izražamo zgolj čustev, ampak se uporabljajo tudi za druge namene, kot so vzdrževanje pogovora, igriva interakcija in ustvarjanje nekih značilnosti v komunikaciji, ki so skupne vsem udeležencem (Umashanthi Pavalanathan in Jacob Eisenstein, 2015).

3 Zasnova raziskave in metodologija

3.1 Predstavitev korpusa

Analiza, ki jo predstavljamo v tem prispevku, je bila izvedena s pomočjo slovenskega in angleškega dela korpusa Janes Tviti 0.3.4 (Erjavec et al., 2015).

Slovenski del korpusa vsebuje več kot 56 milijonov besed oz. 4 milijone tvitov, ki jih je napisalo okoli 7.600 avtorjev. Med njimi prevladujejo moški (53 %), ki so v korpus prispevali tudi največji delež tvitov in enak delež pojavnic (56 %). Žensk je približno pol manj, in sicer slabih 25 %. Objavile so 27 % tvitov in enak delež pojavnic v korpusu. V podobnem deležu se pojavljajo uporabniki, ki jim ni bilo mogoče pripisati spola (22 %).

Angleški del korpusa vsebuje 8 milijonov besed oz. 1 milijon tvitov, kar je štirikrat manj kot slovenski, vendar je to pričakovano, saj so v korpus Janes zajeti samo uporabniki, ki aktivno tvitajo v slovenščini.

3.2 Potek raziskave

V raziskavi smo analizirali rabo omemb uporabniških imen, heštegov in emotikonov ter emojijev v slovenskem in angleškem delu korpusa Janes Tviti 0.3.4. Za vsakega od obravnavanih komunikacijskih elementov smo izdelali frekvenčne sezname najpogostejših in podrobno analizirali najpogostejših 1000 elementov z vsakega seznama. Za vse tri tipe obravnavanih elementov smo oblikovali čim bolj podobne kriterije jezikoslovne analize, kot so slovensko / tuje, standardno / nestandardno, enobesedno / večbesedno /okrajšano in tematika. Kriterije za posamezen tip analiziranih elementov podrobneje predstavljamo v razdelku 4.

V drugem delu analize smo rabo novih komunikacijskih elementov proučevali za različne družbene skupine uporabnikov, pri čemer smo podkorpuse izdelali s pomočjo metapodatkov, s katerimi je opremljen korpus. Upoštevali smo naslednje parametre: moški / ženske in zasebni / javni račun. Tako v slovenščini kot v angleščini smo za vse tri tipe komunikacijskih elementov v vseh štirih izdelanih podkorpusih analizo opravili za 100 najpogostejših elementov. Opazovali smo, v čem so si rezultati podobni in v čem se razlikujejo, pri čemer smo vselej upoštevali rabo v sobesedilu.

Zadnji del analize je bil namenjen podrobni analizi skladenjske, pomenoslovne in pragmatične vloge obravnavanih komunikacijskih elementov v slovenskih in angleških tvitih, ki smo jo izvedli na naključnem vzorcu 100 konkordanc za posamezni podkorpus, izdelan v drugem delu analize.

4 Analiza podatkov

4.1 Omembe uporabniških imen

Slovenski korpus vsebuje 3,3 milijone oz. 46.974 na milijon omemb uporabniških imen. Relativno gledano jih angleški korpus vsebuje le nekaj več, 595.936 oz. 51.356 na milijon. To pomeni, da se pri tvitanju v slovenščini uporabniki vključujejo v bistveno manjšo družbeno mrežo kot v angleščini, najverjetneje zato, ker je pri tvitanju v slovenščini ta družbena mreža pretežno lokalna, pri tvitanju v angleščini pa mednarodna.

4.1.1 Tipologija omemb

Pri analizi najpogostejših 1000 omemb v celotnem slovenskem in angleškem korpusu tvitov smo upoštevali tri kriterije:

- ali so omembe v slovenskem ali tujem jeziku,
- ali so omembe enobesedne, večbesedne ali okrajšane,
- ali so omembe napisane v standardnem ali nestandardnem jeziku.

Kot okrajšane smo obravnavali uporabniška imena, ki so vsebovala kratice (*nk_maribor, DARS_SI, HDDOlimpija*), uporabniška imena, kjer je bilo okrajšano ime ali priimek (*MKamarin, zzTurk, ABratusek*) ter uporabniška imena, kjer se je pojavil okrajšani doktorski naziv ali angleška okrajšava »mr.« (*drVinkoGorenak, mr_foto, MrGabbah*).

Kot standardne smo označili omembe, napisane z velikimi ali malimi začetnicami ter s podčrtajem za presledek, kot npr. *MarkoSket*, *savicdomen*, *Petra_Jansa*. Vse ostale omembe smo označili kot nestandardne.

Rezultati tega dela analize so pokazali, da se tako v slovenskem kot v angleškem delu korpusa pojavlja več omemb slovenskih kot tujih uporabniških imen: v slovenskem korpusu delež znaša 84 %, v angleškem pa 65,8 %. To ni nenavadno, saj so v angleški korpus zajeti tviti slovenskih uporabnikov, kadar ti tvitajo v angleščini in so zaradi tega pričakovano še vedno v precejšnji meri vezani na slovenski prostor.

Pri primerjavi uporabniških imen, ki se pojavljajo v slovenskem in angleškem delu korpusa, smo ugotovili, da se v angleškem delu korpusa pojavljajo uporabniška imena, ki se v slovenskem delu ne pojavljajo, in obratno. V slovenskem korpusu se ne pojavljajo izrazi, ki so vezani na angleško govoreči prostor. To so večinoma imena tujih oseb (*@joerogan*, *@lindseyvonn*) ali časopisov (*@Independent*, *@nytimes*). V angleškem delu korpusa pa se med najpogostejšimi 1000 omembami ne pojavljajo imena slovenskih političnih strank (*@stranka SLS*, *@strankaSD*). Delež omemb, ki se pojavljajo samo v angleškem delu korpusa, znaša 29,8 %, kar pomeni, da se oba frekvenčna seznama prekrivata kar v 70,2 %.

Pri obeh delih korpusa so največkrat uporabljena večbesedna uporabniška imena (65 %), sledijo jim enobesedna (18 %), najmanjkrat pa se pojavijo okrajšana (17 %). Večbesedna uporabniška imena so največkrat sestavljena iz imena in priimka (*@petrasovdat*, *@MatevzNovak*), imen podjetij ali spletnih portalov (*@uporabnastran*, *@RevijaReporter*), identificirali pa smo tudi cele besedne zveze ali stavke (*@MyBlueDragoness*, *@JsSmRenton*). Pri okrajšavah gre največkrat za krajšanje imena ali priimka (*@JJansaSDS*, *@nmusar*) ali ime, v katerem se pojavi kratica

(*@RTV_Slovenija*, *@strankaSD*). Nekajkrat se pojavljajo tudi idiosinkratična kratična ali alfanumerična poimenovanja, katerih pomena brez poznavanja ozadja ni mogoče določiti (*@Z3MQP*, *@z8_LJ*).

Pri proučevanju standardne in nestandardne rabe se je izkazalo, da je v obeh delih korpusa nekoliko pogostejša nestandardna raba (53 %). Kot nestandardno smo tu označili vsa poimenovanja, ki vsebujejo številke (*@lenci53*, *@vinkovasle1*), so zapisana v pogovornem jeziku (*@abejz_no*, *@merineseri*) ali pa so nerazumljiva (*@DC43*, *@schoo666*).

4.1.2 Analiza omemb pri različnih skupinah uporabnikov

V tem delu analize smo se osredotočili na primerjavo rabe omemb pri različnih skupinah uporabnikov:

- moški in ženski računi in
- zasebni in korporativni računi.

Rezultati so pokazali, da so v slovenskem delu korpusa v vseh kategorijah pogosteje uporabljena uporabniška imena v slovenskem jeziku, medtem ko v angleškem delu korpusa nekatere skupine uporabnikov pogosteje uporabljajo imena v tujem jeziku (npr. moški in ženske s korporativnimi računi). To nakazuje, da slovenski uporabniki pri tvitanju s korporativnimi računi uporabljajo več omemb zaradi poslovnega sodelovanja s tujimi podjetji in iz tega razloga tudi več tvitajo v angleščini kot uporabniki z zasebnimi računi.

V obeh delih korpusa se v vseh kategorijah večkrat pojavljajo večbesedne kot enobesedne omembe uporabniških imen. Analiza je pokazala, da v slovenskem delu korpusa večbesedne omembe uporabniških imen največkrat uporabljajo ženske z zasebnim računom (63 %), najmanjkrat pa moški s korporativnim računom (52 %). V angleškem delu korpusa pa večbesedne omembe imen največkrat uporabljajo ženske s korporativnim računom (73 %) in najmanjkrat moški s korporativnim računom (59 %).

V povprečju je v obeh delih korpusa uporabljenih več standardnih (52 %) kot nestandardnih (48 %) oblik uporabniških imen.

4.1.3 Analiza omemb v sobesedilu

V tem delu smo na vzorcu 100 naključnih pojavitev omemb v korpusu analizirali rabo omemb v sobesedilu. Za analizo smo izbrali naslednje kriterije:

- ali se omemba uporablja na začetku, na sredini ali na koncu tvita,
- ali se omemba uporablja skladenjsko ali neskladenjsko,
- ali tvit vsebuje eno samo omembo ali več in
- kaj uporabniško ime označuje (osebo, kraj, spletni portal, podjetje, skupino, politično stranko ali oddajo).

Glede na rezultate analize prvega kriterija smo ugotovili, da se omembe v obeh delih korpusa v povprečju uporabljajo na začetku tvita (51 %). Večinoma gre za omembe uporabniških imen s to funkcijo, da bi jih opozorili na določeno temo, ki zadeva te uporabnike ali pa so pri njej že prej sodelovali. V slovenskem delu korpusa jih na začetku največkrat uporabljajo moški z zasebnimi računi (82 %), najmanjkrat pa ženske s korporativnimi računi (40 %). V angleškem delu korpusa pa je bila situacija malo drugačna: omembe na začetku tvita največkrat uporabljajo ženske z zasebnimi računi (54 %) in najmanjkrat moški z zasebnimi računi (23 %).

Skladenjska analiza je pokazala, da so v obeh delih korpusa omembe pogosteje rabljene skladenjsko. Pri angleškem delu so zaradi skladnje angleškega jezika skladenjsko rabljene vse omembe, medtem ko se v slovenskem delu pojavlja tudi neskladenjska raba, katere delež znaša 5 %. V slovenskem delu korpusa se neskladenjska raba pojavlja predvsem takrat, ko pred omembo uporabniškega imena stojijo predlogi in bi bilo omembo potrebno pregibati, kar v tehničnem smislu ni mogoče, saj aplikacija v tem primeru omembe ne prepozna (*tale me je spomnila na @BesedaDneva*, *še malo pa se bo začel letošnji festival Morja in sonca v @avditorij Portorož*). Pojav bo zanimivo opazovati v daljšem obdobju, ko bomo lahko ugotovili, ali so se pojavile tehnične izboljšave, ali so se uporabniki tehnični omejitvi prilagodili z jezikovno rabo, ki se težavi uspešno izogne ali pa se je zaradi tehničnih okoliščin začela spreminjati slovenska skladnja in pred tradicionalno neprvosklonskimi predlogi dopušča tudi imenovalnik.

Tako v slovenskem kot tudi v angleškem delu se v povprečju večkrat pojavlja ena omemba uporabniškega imena na tvit (57 %). Pri več omembah na tvit se je izkazalo, da uporabniki v obeh delih korpusa največkrat uporabljajo 2 omembi na tvit (24 %), največje število omemb v tvitu pa je 8.

V obeh delih korpusa uporabniška imena največkrat označujejo osebe, sledijo jim spletni portali (*@YouTub*e, *@BesedaDneva*), podjetja in politične stranke. Pri tem je zanimivo, da stranke omenjajo le moški uporabniki v slovenskem delu korpusa, ki imajo zasebni račun.

4.2 Heštegi

V tem razdelku predstavljamo rezultate analize heštegov, ki so zbrani v slovenskem in angleškem korpusu Janes Tviti v.0.3.4. Heštegi so na družbenem omrežju Twitter ena izmed najbolj priljubljenih elementov izražanja. V slovenskem delu korpusa se jih pojavi dober milijon oz. 15.319 na milijon, v angleškem pa 450 tisoč oziroma 38.735 na milijon, kar je relativno gledano več kot 2,5 krat pogosteje kot v slovenskih tvitih. Ker je slovenski del korpusa tvitov šestkrat večji od angleškega, je razumljivo, da vsebuje dvakrat več različnih heštegov (209.629) kot angleški (114.943).

4.2.1 Tipologija heštegov

Pri heštegih smo pri analizi frekvenčnega seznama 1000 najbolj pogostih pojavnic opazovali naslednje:

- tema (šport, politika)
- enobesedni, večbesedni ali okrajšava,
- slovenski ali tuji jezik in
- standarden ali nestandarden

Kot nestandardne smo označili vse heštege, pri katerih avtorji niso zapisali lastnih in zemljepisnih imen z veliko začetnico, kjer so uporabljali tujke kot npr. *#Sochi2014*. Če sta se pojavili dve pojavnici, kot sta *#PLTS* in *#plts*, smo kratico, napisano z majhnimi črkami, označili kot nestandardno. V to kategorijo spadajo tudi vsi pogovorni izrazi, kot so *#gotofje*, *#fuzbal, #butale*, vse večbesedne zveze, kot sta *#TvitajmoZaNase* in *#NaDanašnjiDan*, pri katerih so vse naslednje komponente napisane z veliko

začetnico in vse besedne zveze, pri katerih avtor ni uporabil šumnikov.

V slovenskem delu korpusa so bili prvi trije najpogostejši heštegi *#junaki, #plts, #slochi*, v angleškem delu pa *#Slovenia, #Ljubljana, #slovenia*.

V slovenskem delu se je največ heštegov nanašalo na temo športa (20 %), sledi politika (11 %). Na veliko razliko pri zastopanosti tematik v heštegih smo naleteli v angleškem delu korpusa, v katerem smo zasledili bistveno višji delež heštegov s področja podjetništva in izjemno nizko število političnih heštegov (zgolj 3 %). Na podlagi tega sklepamo, da lokalno politično dogajanje ni tema, o kateri bi slovenski uporabniki Twitterja razpravljali v angleščini.

Tako v slovenskem (53 %) kot tudi angleškem (60 %) korpusu prevladujejo enobesedni heštegi, s 32 % jim sledijo večbesedne zveze v slovenskem delu *(#ligaprvakov, #nočnastraža)* in v angleškem 29 % *(#NBABallot*) ter kratice *(#EU, #lol*), ki v slovenskem delu predstavljajo 15 %, v angleškem 11 %. Ti rezultati kažejo na potrebo po še posebej ekonomičnem izražanju v heštegih.

Rezultati so pokazali, da je velika večina slovenskih heštegov (62 %) rabljenih v tvitih, napisanih v standardni slovenščini. Zdi pa se presenetljivo, da je raba šumnikov v slovenskem korpusu zelo nizka, saj zgolj 5 % heštegov vsebuje šumnike. V angleškem korpusu je odstotek heštegov, napisanih v standardni angleščini, še višji (72 %), kar nakazuje, da slovenski uporabniki angleščino uporabljajo za formalnejše oblike komuniciranja kot slovenščino.

4.2.2 Analiza heštegov pri različnih skupinah uporabnikov

V drugem delu analize smo rabo heštegov primerjali v izbranih podkorpusih, ki smo jih ustvarili glede na metapodatke, s katerimi je opremljen korpus Janes Tviti v.0.3.4, in sicer ženske (zasebno, javno) in moški (zasebno, javno).

V podkorpusih slovenskih tvitov moških in žensk (moški-zasebno, moški-javno, ženske-zasebno, ženske-javno) so rezultati pokazali, da so ženske s korporativnim računom večkrat uporabile večbesedne heštege (48 %), moški pa enobesedne (45,5 %). V angleškem delu so najpogostejše enobesedne zveze, ki so jih največkrat uporabile ženske s korporativnim računom (60 %).

Največji razkorak opazimo pri primerjavi med moškimi in ženskami glede na rabo standardnih in nestandardnih rabo heštegov v slovenskem delu, ki kaže, da moški nekoliko večkrat uporabljajo standardne heštege (56 %), predvsem tisti s korporativnimi uporabniškimi računi (61 %). Poleg tega ženske z zasebnimi računi najmanjkrat pišejo heštege v standardni slovenščini (54 %). Medtem pa oboji pogosteje uporabljajo heštege v standardni angleščini, a največkrat ženske s korporativnimi računi z 75 % .

Ugotovili smo tudi, da moški z zasebnim računom uporabljajo več (11 %) političnih heštegov kot ženske (tako imetnice zasebnih kot javnih računov), poleg tega je v obeh moških podkorpusih 26 % heštegov,največ pri uporabnikih z zasebnim računom z 30 %, ki se nanašajo na šport. V angleškem delu je zanimivo, da ženske uporabnice v angleškem podkorpusu s korporativnim uporabniškimi računi niso uporabile niti enega športnega heštega, najdemo pa veliko število heštegov s področja podjetništva *(#ecommerce, #Google*) in marketinga *(#SocialMediaMarketing, #contentmarketing*). V nasprotju z moškimi uporabniki, ki ne glede na zaseben (27 %) ali uraden (16 %) račun, uporabljajo športne heštege. Predvidevamo lahko, da uporabnice ne opravljajo služb, ki bi zahtevale uporabo športnih tegov, in na uradnih računih uporabljajo le besedišče s svojega delovnega področja, medtem ko se moški z uradnim računom opredeljujejo tudi s svojimi prostočasnimi mislimi in mnenji oz. to od njih zahteva tudi narava njihovega dela *(#srcebije, #mismomaribor*).

4.2.3 Analiza heštegov v sobesedilu

V zadnjem delu analize naključnega vzorca 100 konkordanc nas je zanimalo, kje v tvitu uporabniki heštege najpogosteje uporabljajo, ali so heštegi rabljeni stavčno oz. nestavčno ter kakšno pragmatično funkcijo v tvitu opravljajo.

Postavitev heštegov lahko razdelimo v tri kategorije:

1. začetek oz. uvod: hešteg napoveduje temo tvita;
2. sredina: v tem primeru so pogosto rabljeni stavčno; najpogosteje kot predmet, osebek, povedek in prislovno določilo kraja;
3. konec oz. epilog: hešteg je na koncu tvita, da še dodatno zaznamuje temo tvita.

Podobno kot v sorodnih raziskavah so tudi naši rezultati pokazali, da se večina heštegov pojavi na koncu tvita: 72 % heštegov v slovenskem in 61 % v angleškem korpusu. V slovenskem delu jih je kar 83 % je uporabljenih nestavčno, 80 % pa z namenom, da dodatno zaznamujejo temo tvita. Podobno so pokazali rezultati iz angleškega dela, kjer prevladuje nestavčna raba heštegov (81 %).

Največ stavčno uporabljenih heštegov najdemo v slovenskem podkorpusu moški-zasebno, in sicer 24 %; med njimi jih je 70 %, ki se jih pojavi na sredini. Med angleškimi korpusi največ stavčno rabljenih heštegov najdemo v angleškem podkorpusu ženske-javno (34 %), med njimi jih je 83 %, ki se jih pojavi na sredini. Tako kot v angleškem kot v slovenskem delu so heštegi na sredini največkrat uporabljeni kot predmet, na drugem mestu kot osebek in na zadnje prislovno določilo kraja.

Poleg tega smo ugotovili, da uporabniki v angleščini uporabljajo več heštegov v enem tvitu (78 %) kot v slovenščini (60 %). Iz tega lahko sklepamo, da se uporabniki v slovenščini tekoče izražajo v polnih povedih, medtem ko pri tvitanju v angleščini pogosteje nizajo heštege ob krajših izjavah.

4.3 Emotikoni in emojiji

V analizi smo obravnavali tudi emotikone in emojije. Z obema želi uporabnik izraziti določeno čustvo, ki bi ga v komunikaciji v živo izrazil z intonacijo, gestami ali obrazno mimiko in bi ga v pisni računalniško posredovani komunikaciji težje izrazil z besedami. V celotnem korpusu smo opazili, da v primerjavi z moškimi ženske uporabljajo več emojijev (ženske rel.frek. 180,8, moški 74,8). Tudi v angleščini so emojiji pogostejši pri uporabnicah (rel.frek. 210) kot pri uporabnikih (rel.frek. 56,7).

4.3.1 Tipologija emotikonov in emojijev

Pri emotikonih in emojijih smo za analizo frekvenčnega seznama 1000 najpogostejših pojavnic izbrali naslednje tri kriterije:

- pomen čustvenega izraza,
- enostaven ali sestavljen čustveni izraz in
- pozitiven ali negativen čustveni izraz.

Kot sestavljene smo upoštevali emotikone, ki vsebujejo več kot 2 znaka, zaradi majhne velikosti angleškega korpusa v njem ni 1000 različnih emotikonov in emojijev, zato smo v analizo vključili vse.

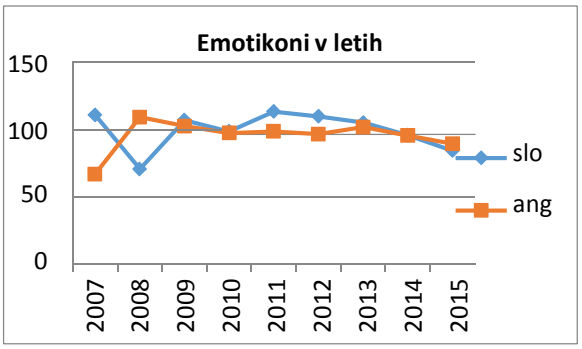
Emotikoni: Rezultati te analize so pokazali, da so emotikoni v slovenskem in angleškem jeziku po pomenu primerljivi, saj prednjačijo emotikoni s pozitivnim sporočilom, najpogostejši (5,4 %) so izrazi smeha (npr. :-) ali :)). Na drugem mestu so emotikoni, ki izražajo žalost (:(in :-( ) (4,1 %). Pri kriteriju sestavljen/enostaven opažamo, da v obeh jezikih prevladujejo sestavljeni emotikoni (88 % v angleščini, 80 % v slovenščini).

Emojiji: Na prvo mesto (v slovenščini 13,6 % in v angleščini 12,1 %) se v obeh jezikih uvrščajo emojiji, ki upodabljajo naravo (✿,🐞), na drugem mestu (v angleščini 5,1 % ter slovenščini 2,9 %) jim sledijo emojiji, ki zaznamujejo hrano (🍙, 🍤). Izmed emojijev, ki označujejo čustva, so v obeh jezikih najpogostejši (v angleščini 2,6 % in v slovenščini 2,1 %) tisti, ki izražajo ljubezen (❤ ,💘).

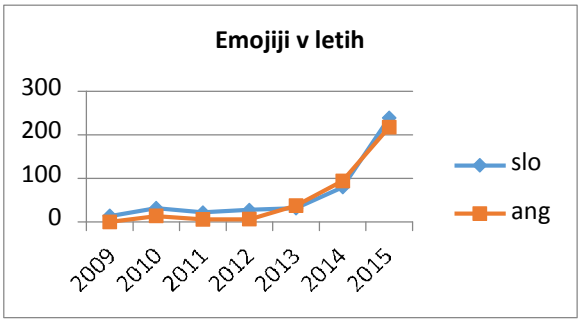
Vse emojije smo obravnavali kot enostavne, saj predstavljajo en sam Unicode znak.

Pri analizi celotnega korpusa smo se osredotočili na tisti kriterij, ki se nam je zdel najbolj zanimiv, in sicer na kronološko analizo čustvenih izrazov.

Kot kažeta Grafa 1 in 2, raba emojijev stalno narašča, še posebej strmo v zadnjih treh letih, medtem ko raba emotikonov v obeh jezikih več ali manj stagnira oz. celo počasi upada.



Graf 3: Emotikoni v letih



Graf 4: Emojiji v letih

4.3.2 Analiza emotikonov in emojijev pri različnih skupinah uporabnikov

V tem delu analize smo v izbranih podkorpusih, ki smo jih ustvarili glede na metapodatke ženske (zasebno, javno) moški (zasebno, javno), primerjali enake kriterije kot v razdelku 2.3.1 (pomen, pozitiven/negativen, sestavljen/enostaven).

V analizo smo vključili 50 najpogostejših pojavnic.

Moški z javnimi računi, ki tvitajo v slovenščini, uporabljajo več **emotikonov**, ki ponazarjajo ljubezen (6 %), a tudi več, ki jih ponazarja žalost v primerjavi z angleščino (16 % žalost v slovenščini, medtem ko je teh v angleščini le 6 %). Pri analizi **emojijev** v tem podkorpusu pa smo ugotovili, da je v slovenščini uporabljenih kar 14 % emojijev, ki zaznamujejo naravo, v angleščini pa samo 6 %. Pri **moških z zasebnimi računi** v slovenskem jeziku smo ugotovili, da je pogostost emojija, ki ponazarja žalost 6-odstotna, v angleščini pa znaša 2 %. Pri zasebnih računih smo opazili tudi razliko med moškimi in ženskami. Pri emotikonih je na primer razvidno, da uporabljajo moški, ki tvitajo v slovenskem jeziku več emotikonov za žalost v primerjavi z ženskami (16 % proti 12 %), a tudi več emotikonov, ki ponazarjajo smeh (46 % proti 36 %). V angleščini pa obratno, saj so emotikoni za smeh pogostejši pri ženskah (50 % proti 44 %). Za emojije vidimo, da izmed tistih, ki uprizarjajo ljubezen so v slovenščini bolj popularni pri ženskah kot pa ne pri moških (12 % proti 8 %) in prav tako emojiji za smeh.

Ženske z zasebnimi računi uporabljajo v angleškem jeziku več emotikonov za žalost (16 % proti 12 %) in več emotikonov, ki ponazarjajo ljubezen (10 % angleščina, 2 % slovenščina).

4.3.3 Analiza emotikonov in emojijev v sobesedilu

V tem delu analize smo za izbrane 4 podkorpuse v obeh jezikih na naključnem vzorcu 100 konkordanc opazovali, ali so emojiji in emotikoni v tvitih rabljeni namesto ločila ali namesto besede ter ali stavek omilijo ali ga podkrepijo. V zadnji kategoriji smo ločevali emojije in emotikone, ki podkrepijo stavek tako, da smo opazovali kontekst, npr.:

1) *ženitnega posrednika se pa ne grem :D*
2) *Lučkeee, juhej :)))*

1) *Tiho bodi, ves da nisem tebe mislila!!*
2) *Pr šepanju mi gre najbolj na živce to, da sm tak počasna – bolj k bolečina, res.*

Pri št. 1) smo upoštevali, da emotikon/emoji omili stavek, pri št. 2) pa, da ga podkrepi.

Pogostost uporabljanja **emotikonov** namesto ločil je večja pri moških z zasebnimi računi, ki tvitajo v slovenščini, kot pa v angleščini (52 % proti 48 %). Več emotikonov omili stavek v angleščini, kot v slovenščini (40 % proti 32 %). Če pogledamo ženske uporabnice, lahko opazimo, da ženske z zasebnimi računi v obeh jezikih uporabljajo skoraj enako število emotikonov namesto ločil (59 % v slovenščini, 58 % v angleščini). Moški z zasebnimi računi uporabljajo veliko več krepilnih emotikonov v slovenščini kot v angleščini (68 % proti 45 %), medtem ko jih ženske z zasebnimi računi uporabljajo malo več v angleščini (63 % proti 61 %). Večina uporabnikov ne nadomešča besed z emotikoni.

Pri uporabi emojijev pa smo ugotovili, da jih moški z javnimi računi v angleškem jeziku bistveno pogosteje uporabljajo namesto ločil kot v slovenščini (62 % v angleščini in 46 % v slovenščini). Razlika pri moških in ženskah z javnimi računi je tudi pri emojijih, ki omilijo stavek, saj jih moški v slovenščini uporabljajo v večjem številu kot v angleščini (12 % proti 9 %), ženske pa obratno (10 % proti 12 %). Pri ženskah z zasebnimi računi, ki tvitajo v angleščini, je pogostost emojijev, ki podkrepijo stavek, veliko večja v primerjavi s slovenščino (90 % angleščina, 73 % slovenščina), pri ženskah z javnimi računi pa smo opazili ravno obratno (79 % slovenščina, 77 % angleščina).

5 Zaključek

V prispevku smo opravili podrobno analizo novih komunikacijskih elementov na družbenem omrežju Twitter v slovenskem in angleškem korpusu Janes Tviti v.0.3.4, ki vsebuje tvite slovenskih uporabnikov. Opazovali smo rabo omemb, heštegov ter emotikonov in emojijev ter jo primerjali med jezikoma ter med ženskimi in moškimi uporabniki, ki tvitajo v zasebne ali službene namene. Omembe se tako v slovenskem kot angleškem delu korpusa pogosteje pojavljajo na začetku tvita. V obeh delih korpusa prevladuje nestandardna oblika, omembe so po večini večbesedne. V angleškem delu korpusa se pojavljajo omembe, ki se v slovenskem delu ne in ravno obratno.

Heštegi se tako v slovenskem kot v angleškem delu korpusa v veliki večini pojavljajo na koncu tvita z namenom, da dodatno zaznamujejo temo tvita, in so večinoma standardni. Pri nestandardnih v slovenščini izstopa izjemno nizka uporaba šumnikov. Najpogostejša tema heštegov je v slovenskem in angleškem delu korpusa šport. V slovenskem delu sledi politika. Na veliko razliko pri zastopanosti tematik v heštegih smo naleteli v angleškem delu korpusa, v katerem smo zasledili bistveno višji delež heštegov s področja podjetništva in izjemno nizko število političnih heštegov). Moški uporabljajo pretežno enobesedne, ženske pa večbesedne heštege.

V korpusu Janes so najpogosteje uporabljeni emotikoni in emojiji, ki upodabljajo naravo, hrano in čustva, predvsem pozitivna, njihova prevladujoča vloga pa je, da okrepijo izjavo. Medtem ko raba emotikonov počasi upada, raba emojijev strmo narašča. Po obeh tipih čustvenih izrazov v obeh jezikih pogosteje posegajo ženske, ki jih velikokrat uporabljajo namesto končnih ločil.

Z našo raziskavo smo potrdili in ovrgli nekatere naše hipoteze:

1. Omembe uporabniških imen so bile v angleškem delu korpusa večkrat uporabljene v slovenskem jeziku. S tem je bila naša hipoteza ovržena.
2. Ženske uporabljajo več čustvenih izrazov kot moški. S tem je bila naša hipoteza potrjena.
3. Heštegi so bili v obeh delih korpusa večbesedni. Tudi ta hipoteza je bila potrjena.

V prihodnje nameravamo raziskavo razširiti s primerjavo rabe novih komunikacijskih elementov v tvitih uporabnikov, ki prihajajo iz angleško govorečih držav. S tem bomo preverili, v kolikšni meri na njihovo rabo v angleščini vpliva materni jezik, kultura in konvencije računalniško posredovane komunikacije različnih govornih področij.

6 Literatura

David Crystal. 2001. Language and the Internet. Cambridge University Press. The Edinburgh Building.

Evandro Cunha. 2014. He Votes or She Votes? Female and Male Discursive Strategies in Twitter Political Hashtags. http://1.usa.gov/1P0WiR7

Fabrizio Gotti. Hashtag Occurrences, Layout and Translation: A Corpus-driven Analysis of Tweets Published by the Canadian Government. http://bit.ly/1S0BvOl

Tomaž Erjavec. Darja Fišer. Nikola Ljubešić. Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes. http://bit.ly/1OBwO9t

Jako Olivier. 2014. Twitter usernames: exploring the nature of online South African nicknames. http://bit.ly/1Qj5scM

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Emoticons vs. Emojis on Twitter: A Causal Inference Approach. School of Interactive Computing.Georgia Institute of Technology. Atlanta, GA 3030. http://arxiv.org/pdf/1510.08480v1.pdf

Twitter glossary. https://biz.twitter.com/en-gb/glossary

Razdvoumljanje besednega pomena pri strojnih prevajalnikih Amebis Presis, Google Translate in MT@EC

Jure Škerl

Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
jureskerl@gmail.com

Povzetek

V članku je predstavljena raziskava, katere namen je bil izmeriti natančnost razdvoumljanja besednega pomena za tri strojne prevajalnike: Amebis Presis, Google Translate in MT@EC. Na voljo so različni specializirani algoritmi za razdvoumljanje besednega pomena, vendar je njihova implementacija v prakso redka. Od treh preizkušanih sistemov ima le Amebis Presis vgrajene eksplicitne mehanizme za razdvoumljanje, medtem ko se ostala dva zanašata na statistični frazni model, pri katerem je razdvoumljanje posameznih besed posledica statističnega prevajanja večjih kosov besedil naenkrat. Primerjava natančnosti razdvoumljanja med prvim in drugim pristopom je dala nekaj uporabnih uvidov v prednosti in slabosti obeh. Druga raven primerjave je bila med rezultati na različnih jezikovnih področjih: književnost, publicistika, spletni dnevniki.

Word Sense Disambiguation in MT Systems: Amebis Presis, Google Translate, MT@EC

The article presents a research in which accuracy of word sense disambiguation was measured for three MT systems: Amebis Presis, Google Translate and MT@EC. There are various specialised word sense disambiguation algorithms available, however few of them are actually implemented into practical systems. Out of the three tested MT systems, only Amebis Presis uses explicit disambiguation techniques, whereas the other two rely on the statistical phrase-based model, in which disambiguation of words is the result of statistically translating longer text segments. The comparison of disambiguation accuracy between the first and the second approach provided some useful insights into advantages and disadvantages of both. Another level of comparison was done by testing disambiguation on texts from different language domains: literary, journalistic and blogs.

1 Uvod

Razdvoumljanje besednega pomena (RBP) je eden najstarejših in tudi najtežje rešljivih problemov v okviru strojnega prevajanja. Poleg tega se z njim v naravnih besedilih srečujemo pogosteje, kot je očitno na prvi pogled. Za 121 najpogostejših angleških samostalnikov, denimo, ki predstavljajo približno petino vseh besed v poljubnem besedilu, angleški WordNet v povprečju našteje 7,8 različnih pomenov (Agirre in Edmonds, 2007). Če k temu dodamo še 70 najpogostejših glagolov, ki imajo v WordNetu povprečno po dvanajst pomenov (Palmer et al., 2007), lahko zaključimo, da bomo primere večpomenskosti srečali v slehernem naravnem besedilu.

To je pomembno poudariti, ker povprečen bralec večine večpomenskih besed niti ne zazna; človeški možgani razdvoumljanje opravljajo tako rekoč trenutno in večinoma povsem podzavestno. To počnejo na osnovi vsega znanja o svetu, ki so ga nabrali tekom svojega obstoja, načini, na katerega je to znanje shranjeno, pa so abstraktni in nikakor ne na voljo v oblikah, ki bi bile dostopne računalniškim sistemom. Računalnik se zato lahko spotakne že ob tako osnovne primere večpomenskosti, kot je tista v stavku »Ana je jabolko.« Kako lahko strojni prevajalnik ve, ali gre v tem stavku za glagol »jesti« ali glagol »biti«? Brez znanja o svetu, ki človeškemu bralcu omogoča brez pomislekov odgovoriti na to vprašanje, je vse, kar mu preostane, kontekst, se pravi okoliške besede.

Načini, na katere lahko slednjega analizira in ga uporabi kot osnovo za odločanje med pomeni, so raznoliki in segajo od uporabe pomenskih virov, kot so slovarji in tezavri, pa vse do statistične analize ogromnih količin vzporednih besedil v dvojezičnih korpusih. Ta prispevek ni poskus predstavitve vseh možnih pristopov, temveč le poskus analize praktičnih rezultatov, ki jih pri razdvoumljanju dosegajo trije v naslovu omenjeni strojni prevajalniki.

Namen članka je podrobna predstavitev rezultatov razdvoumljanja treh strojnih prevajalnikov (Presis, Google Translate in MT@EC) pri smeri prevajanja iz slovenščine v angleščino. Eksperiment je bil izveden na več besedilih v skupnem obsegu 869 besed, ki so bila izbrana tako, da so bila žanrsko, tematsko in slogovno raznolika. Nabrana so bila s treh jezikovnih področij: knjižna, publicistična in spletna besedila. Natančnosti razdvoumljanja so bile izmerjene za vse skupaj in za vsak sklop posebej. To je omogočilo primerjavo ne samo med posameznimi strojnimi prevajalniki, temveč tudi med rezultati, ki jih vsak od njih dosega na različnih jezikovnih področjih. Od treh prevajalnikov ima le Presis vgrajene specializirane mehanizme za razdvoumljanje (Holozan, 2011), druga dva, ki sta statistična, pa ne. Zato bi rezultati morali razkriti tudi učinek prisotnosti oziroma odsotnosti takšnih mehanizmov za razdvoumljanje ter razlike med njimi in golim statističnim pristopom.

2 Pregled pristopov

Pristope k RBP lahko v grobem razdelimo na dva dela: korpusne in nekorpusne (Agirre in Edmonds, 2007). Prednost korpusnih pristopov v primerjavi z nekorpusnimi je višja natančnost razdvoumljanja, slabost pa, da znajo razdvoumljati le besede, ki se znajdejo v uporabljenih korpusih. Nekorpusne metode, ki se naslanjajo na slovarje in pomenske vire, kot je WordNet, po drugi strani omogočajo razdvoumljanje precej širšega besedišča.

Med nekorpusne pristope uvrščamo denimo algoritme, ki računajo *semantično sorodnost*: iz predpostavke, da naravna besedila težijo h koherentnosti, lahko sklepamo, da bodo pomeni besed iz enega besedila med seboj bolj ali

manj sorodni. Na podlagi te ugotovitve za pravilen pomen dvoumnih besed izbiramo tiste, ki izkazujejo najvišjo semantično sorodnost z okoliškimi pomeni (Mihalcea, 2007). Enega prvih postopkov, delujočih na tem principu, je razvil Philip Resnik (1995). Za začetek je vpeljal pojem specifičnosti koncepta, kar je definiral kot verjetnost, da se nek koncept pojavi v relativno obsežnem korpusu. Resnik nato definira semantično bližino med dvema besedama tako, da kvantificira razdaljo do najnižjega skupnega vozlišča, do katerega pridemo, če potujemo od obeh besed po hierarhiji navzgor.

Njegovo enačbo sta nekoliko prilagodila Jiang in Conrath (1997), in sicer tako, da sorodnost med dvema konceptoma merita z razliko v informacijski vsebini (ang. information content). Hirst in St-Onge (1998) pa sta v svojo verzijo enačbe za izračun semantične sorodnosti integrirala še smer, v katero tečejo povezave. Ta pristop temelji na ideji, da je semantična sorodnost med dvema konceptoma tem višja, čim manjkrat povezava med njima spremeni smer.

Druga skupina znotraj nekorpusnih pristopov k RBP so *hevristične metode*, ki izrabljajo določene naravne zakonitosti jezikov. Njihova prednost je v preprostosti, zaradi katere ne zahtevajo veliko računske moči, slabost pa na splošno nižja natančnost razdvoumljanja, čeprav so lahko v specifičnih primerih zelo uspešne. Najosnovnejši hevristični postopek RBP je izbiranje najpogostejšega pomena. Naravna zakonitost, ki se jo tu izrablja, je dejstvo, da distribucije besednih pomenov sledijo Zipfovemu zakonu (1949), tj. statistični distribuciji, v kateri je ena kategorija dominantna, frekvence vseh ostalih pa so bistveno nižje. Na osnovi tega lahko izdelamo zelo preprost in robusten sistem za razdvoumljanje, ki vsaki dvoumni besedi pripiše pomen, ki je v naravnih besedilih najpogostejši. Še en primer hevristične metode je izbiranje enega pomena na diskurz, ki so ga vpeljali Gale in sodelavci (1992b) in temelji na predpostavki, da večpomenska beseda ohrani isti pomen skozi celotno besedilo, v katerem se pojavlja. To pomeni, da je razdvoumljanje večjega števila njenih pojavitev trivialna naloga, če pravilno identificiramo njen pomen v najmanj eni od njih.

Druga obsežna skupina pristopov k RBP temelji na izrabi korpusov. Vanjo spadajo številne metode, tu pa se bomo na kratko dotaknili le ene: RBP na osnovi *prevodne vzporednosti*. Ta metoda je sposobna razlikovati med pomeni izključno glede na prevodne ustreznice, ki so zelo uporabna indikacija za razdvoumljanje. Pogoj za njeno delovanje je dvojezični korpus, ki mora izpolnjevati eno poglavitno zahtevo, in sicer to, da so besedila v obeh jezikih poravnana. To pomeni, da ima vsaka fraza ali vsaka beseda v izvirnem besedilu svojo ustreznico v ciljnem besedilu. Do tega pridemo tako, da najprej poravnamo stavke, nato pa še posamezne besede/fraze. Ves postopek se izvaja samodejno, saj bi bila ročna poravnava preveč zamudna. Samo razdvoumljanje nato poteka s pomočjo podatkov o besedišču in skladnji v okolici razdvoumljane besede in preverjanjem, kako je bila ta beseda prevedena v podobnih kontekstih, najdenih v vzporednem korpusu. Eni prvih, ki so preizkušali to metodo, so bili Gale in sodelavci (1992a), kasneje pa tudi Chklovski in sodelavci (2004).

Tako za omenjene kot druge tehnike RBP sicer velja, da so bile v glavnem razvite in preizkušane zgolj v raziskovalnem okolju, njihova implementacija v praksi pa je redka. V nadaljevanju predstavljen eksperiment je tako predvsem poskus splošne primerjave natančnosti razdvoumljanja med statističnimi in na pravilih temelječimi prevajalniki.

3 Metodologija

3.1. Predstavitev prevajalnikov

Preizkušani so bili trije prevajalniki, dva statistična (Google Translate in MT@EC) in eden, ki temelji na pravilih (Presis). Slednjega razvija slovensko podjetje Amebis in je specializiran za jezikovna para slovenščina-angleščina ter slovenščina-nemščina (pri tem le v smeri iz nemščine v slovenščino). Pri stavčni analizi se opira na ročno izdelano podatkovno zbirko Ases, ki vsebuje podatke o besedah, besednih zvezah, skupinah, predlogih in pomenih (Holozan, 2011), za reprezentacijo pomena izhodiščnega besedila pa uporablja vmesni jezik (interlingua).

MT@EC je interni statistični prevajalnik prevajalske službe EU, ki je bil zgrajen na odprtokodnem sistemu Moses. Prevajati zna med vsemi kombinacijami 24 uradnih jezikov EU, pri tem pa uporablja korpuse v skupnem obsegu 1,65 milijarde besed. Sklepamo lahko, da je močno specializiran za zakonodajna in tehnična besedila, ki se tičejo delovanja EU, kar se je deloma potrdilo tudi v eksperimentu.

Zadnji izmed treh preizkušanih prevajalnikov je Google Translate, ki je ravno tako kot MT@EC statističen, vendar temelji na lastniški programski opremi. Z možnostjo prevajanja iz in v 103 jezike sveta je trenutno najbolj obsežen javno dostopen strojni prevajalnik, katerega storitve dnevno uporablja prek 200 milijonov uporabnikov. Besedila za svoje čedalje obsežnejše korpuse podobno kot MT@EC črpa iz dokumentov EU, poleg tega pa še iz uradnih besedil Združenih narodov, ki so praviloma objavljana v šestih uradnih jezikih te organizacije, ter drugih eno- in dvojezičnih spletišč.

3.2. Predobdelava

Preizkušanje razdvoumljanja je potekalo na odlomkih besedil v skupnem obsegu 869 besed, ki so bila izbrana tako, da so pokrila tri različna jezikovna področja – knjižni, publicistični in spletni jezik. Za knjižna besedila so bili odlomki vzeti iz Cankarjeve povesti Krčmar Elija ter iz Mansarde Slavka Gruma. Odlomki publicističnih besedil so bili nabrani z novičarske spletne strani MMC[1], in sicer iz dveh različnih člankov z notranje- in zunanjepolitično tematiko. Zadnja tretjina besedil vključuje objave v manj formalnem jeziku iz dveh spletnih dnevnikov[2]. Cilj takšnega izbora je bil pokriti nekoliko širšo sliko jezikovne realnosti, navkljub relativno majhnemu obsegu testnih besedil.

Na besedilih je bila najprej opravljena ročna semantična analiza pomenov z upoštevanjem njihovega vpliva na izbor prevodnih ustreznic v angleščini. Subjektivnost takšne ročne analize je bila omiljena z doslednim naslanjanjem na več slovarskih ter korpusnih virov, ki so dostopni na spletu in v knjižni obliki. Glavna

[1] www.rtvslo.si
[2] http://tomazjakofcic.com/blog in http://heliopolis.si

uporabljena orodja v tej fazi so bila pregibnik Amebis Besana 4.12, semantični leksikon sloWNet 3.1 in spletni zbirki slovarjev fran.si ter thefreedictionary.com. V prvem koraku semantične analize je bila vsaka beseda vnešena v Amebisov javno dostopni sistem Besana, ki pozna vse pregibne oblike leksemov in zna za vsako poiskati vse izvirne leme. Na ta način je bilo mogoče odkriti vse obstoječe leme, ki se pregibajo v posamezno morfološko obliko. Denimo, v stavku *naročil je še en bokal* gre lahko, če ne upoštevamo skladnje, pri besedi *naročil* za samostalnik v rodilniku dvojine/množine ali pa za pretekli deležnik. V tem primeru gre torej za dvoumnost, ki je posledica dejstva, da dva leksema vsebujeta isto pregibno obliko. Po tej analizi je za vsako besedo nastal seznam vseh možnih lem, iz katerega je bilo nato mogoče izpeljati vse možne pomenske interpretacije. Določanje potencialnih pomenov je bilo izvedeno z vnašanjem lem v slovarski iskalnik fran.si in analizo vrnjenih slovarskih vnosov.

Glede na to, da je predmet raziskave razdvoumljanje za potrebe strojnega *prevajanja,* se je v zadnjem, ključnem koraku semantične analize vse dobljene potencialne pomene navzkrižno preverilo z leksikalno podatkovno bazo sloWNet 3.1 (Fišer in Sagot, 2015) ter angleškim spletnim slovarjem thefreedictionary.com. V tem koraku so bili vsi pomeni, ki imajo v angleškem jeziku isto prevodno ustreznico, združeni v eno enoto, saj je cilj uspešnega razdvoumljanja v strojnem prevajanju dejansko najti "pravilen" prevod in ne nujno "pravilen" pomen.

Končni rezultat takšne obdelave je bil seznam vseh besed v besedilih s pripisanimi števili možnih prevodov (in ne pomenov) ter seznamom le-teh za vsako besedo posebej. Vse to je bilo vnešeno v tabelo, tako da je bilo možno v naslednjem koraku vzporediti besede, njihove možne prevode ter dejanske prevode, ki so jih izbrali preizkušani strojni prevajalniki. Pred tem pa je bilo potrebno seznam še prečistiti, in sicer so bile iz njega izločene naslednje kategorije, ki so bile bodisi nerelevantne bodisi bi povzročile izkrivljene rezultate:

- vse besede, ki imajo v ciljnem jeziku le eno prevodno ustreznico
- vsi osebni zaimki
- vse pojavitve glagola biti, v katerih le-ta nastopa kot pomožni glagol
- vse pojavitve prislova "pa" (če je nastopal kot veznik, je bil ohranjen)

Poleg tega so bili nekateri večdelni vezniki in redke idiomatične fraze združeni v eno pomensko enoto, kjer je bilo to smiselno, in sicer tako, da je bila ohranjena le ključna dvoumna beseda v frazi oziroma večdelnem vezniku. Ta se je štela za pravilno razdvoumljeno, če je prevajalnik uspel pravilno razbrati skupen pomen. Po tej fazi je na seznamu iz knjižnih besedil ostalo 106 večpomenskih enot, na seznamu iz publicističnih besedil 173 in na seznamu iz spletnih dnevnikov 116 (skupno 395). Tako prečiščen seznam predstavlja celoto podatkov, na podlagi katere je bilo moč v nadaljevanju eksperimenta izračunati natančnost razdvoumljanja za vsak preizkušani sistem posebej.

3.3. Kriteriji

Ocene natančnosti razdvoumljanja so bile izračunane ločeno za vsak prevajalnik in za vsako jezikovno področje posebej (knjižna, publicistična in spletna besedila) ter na koncu še za vsa besedila skupaj. Ocena uspešnosti pomenskega razdvoumljanja je razdeljena v štiri kategorije:

(1) Pravilno razdvoumljen pomen
(2) Napačno razdvoumljen pomen
(3) Pravilno razdvoumljen pomen s prevodno napako
(4) Neprevedeno

Pod kategorijo (1) so se uvrstili vsi primeri, v katerih je strojni prevajalnik razdvoumljanje večpomenske besede izvedel pravilno, tj. kjer je našel ustrezen prevod v ciljnem jeziku in ga tudi slovnično pravilno vtkal v prevedeno besedilo.

V kategoriji (2) so združeni vsi čisti primeri napačnega razdvoumljanja, tj. primeri, kjer je prevajalnik izbral očitno pomensko napačen prevod.

Kategorija (3) je zajela primere, kjer so prevajalniki pomensko razdvoumljanje sicer opravili pravilno, vendar je pri tem prišlo do neke prevodne napake. Tipični primeri v tej kategoriji so denimo odločitev za prislov, kjer bi moral stati pridevnik, ter obratno, pa neujemanje pravilno razdvoumljane besede z izvirnikom v sklonu, številu ali osebi, izbrano deležje namesto starinske oblike glagola v 3. os. mn. itd.

Zadnja kategorija (4) zajema vse primere, v katerih je prevajalnik odpovedal in dvoumno besedo pustil neprevedeno, bodisi z izpustitvijo bodisi tako, da jo je pustil v izvirni, slovenski obliki.

Primeri, v katerih je bil pravilno razdvoumljen pomen zajet v okoliških besedah, izvirna beseda pa ni bila neposredno prevedena, so bili uvrščeni v kategorijo (1). Ravno tako primeri, ko je bila pravilno razdvoumljena beseda umeščena na napačno mesto v stavku, tako da je bil prevod napačen s skladenjskega vidika.

4 Rezultati

4.1. Razčlemba

4.1.1. Knjižna besedila

Pri knjižnih besedilih se je najbolje odrezal Presis, ki je dosegel 70,7-odstotno natančnost razdvoumljanja, v kar se štejejo primeri iz kategorije 1 (67,9 odstotka) in primeri, v katerih je bilo razdvoumljanje uspešno, vendar je prišlo do neke druge prevodne napake (kategorija 3: 2,8 odstotka). Nasprotno so bili rezultati, ki jih je na knjižnih besedilih dosegel MT@EC, z naskokom najslabši, saj je dosegel le 49,1-odstotno natančnost (kategorija 1: 43,4 odstotka, kategorija 3: 5,7 odstotka). To je hkrati tudi najnižja posamezna vrednost v celotnem eksperimentu, se pravi izmed vseh kombinacij treh besedilnih žanrov in treh prevajalnikov. S tem se delno potrjuje domneva, da je MT@EC tematsko najožje specializiran prevajalnik, kar ni presenetljivo glede na to, da so ga razvili za potrebe prevajalskih služb EU. Za ilustracijo tega dejstva lepo služi primer napačnega razdvoumljanja glagola *tožiti* v pomenu *pritoževati se*, ki ga je MT@EC prevedel z glagolom *to sue*, kar bi bilo v kontekstu uradnih evropskih

dokumentov skoraj zagotovo pravilno razdvoumljanje, v našem, knjižnem primeru pa ne. V tem primeru je bil Presis edini, ki je razdvoumljanje opravil uspešno in izbral glagol *to moan*. Kar se Googlovega prevajalnika tiče, je v kategoriji knjižnih besedil dosegel 68,8-odstotno natančnost razdvoumljanja (66,0 odstotka v kategoriji 1 in 2,8 odstotka v kategoriji 3), torej le nekoliko manj od Presisa.

4.1.2 Publicistična besedila

Najvišjo natančnost razdvoumljanja publicističnih besedil je dosegel Google Translate, in sicer 83,8 odstotka (kategorija 1: 80,3 odstotka, kategorija 3: 3,5 odstotka). MT@EC se je z 80,9-odstotno natančnostjo uvrstil tesno za njim (kategorija 1: 79,2 odstotka, kategorija 3: 1,7 odstotka). Nekoliko bolj je zaostal Presis z 72,2-odstotno natančnostjo (kategorija 1: 69,9 odstotka, kategorija 3: 2,3 odstotka). Zanimivo si je pogledati prevajanje večpomenske besede *stran*, ki se je v tem delu pojavila trikrat. Prav tolikokrat je njeno razdvoumljanje spodletelo Presisu, ki je dvakrat izbral prevod *page* in enkrat prevod *direction*. Nasprotno je Google Translate vse tri pojavitve besede razdvoumil pravilno, dvakrat s prevodom *side* in enkrat z idiomatično frazo *on the other hand*. MT@EC se je uvrstil med njiju z enim spodletelim poskusom, ko je *na drugi strani morja* prevedel z *on the other hand, the sea*. Ti rezultati namigujejo na dober potencial za razdvoumljanje, ki ga imajo statistični prevajalniki. Vendar pa se, če upoštevamo vse rezultate skupaj, razkrije tudi njihova slabost, namreč dejstvo, da natančno razdvoumljajo le besedila s tistih tematskih področij, na katerih so bili trenirani (v tem primeru so to politične teme uradnih dokumentov EU, ki jih oba uporabljata za trenažne korpuse). To domnevo potrjuje tudi primerjava natančnosti razdvoumljanja po besedilnih kategorijah, kjer vidimo, da Presis kot predstavnik na pravilih temelječih prevajalnikov dosega najbolj konsistentne rezultate skozi vse kategorije, Googlovi in zlasti MT@EC-jevi pa od kategorije do kategorije občutno bolj nihajo (slika 1).

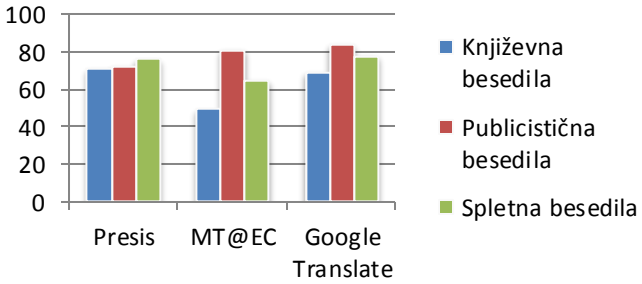
4.1.3 Spletna besedila

Tretjo kategorijo sestavljajo spletna besedila, ki so bila nabrana iz dveh slovenskih spletnih dnevnikov. V prvem je bil jezik bolj pogovoren, v drugem pa bolj knjižen; za oba je sicer značilno pisanje v prvi osebi. Najvišjo natančnost razdvoumljanja v tej kategoriji je dosegel Google Translate s 77,6 odstotki (kategorija 1: 75,9 odstotka, kategorija 3: 1,7 odstotka), sledil je Presis s 75,9 odstotki (kategorija 1: 73,3 odstotka, kategorija 3: 2,6 odstotka), precej zadaj se je uvrstil MT@EC s 64,6 odstotka (kategorija 1: 61,2 odstotka, kategorija 3: 3,4 odstotka). Slabši rezultat slednjega znova namiguje, da gre za prevajalnik, ki je ozko specializiran za potrebe prevajalskih služb EU. Za to se najde tudi konkreten primer: besedna zveza *enoglasna pritrditev*, ki je sicer povzročila težave vsem trem prevajalnikom. Nobeden je ni v celoti razdvoumil pravilno, vendar pa je bil MT@EC temu še najbližje. Medtem ko je besedo *pritrditev* preprosto prezrl, je za prvo besedo pravilno izbral pomen *unanimous*. Tu lahko sklepamo, da je MT@EC pravilen prevod našel v korpusu dokumentov EU, kjer se lema *enoglasen* verjetno res najpogosteje pojavlja v povezavi s prevodom *unanimous*. Presis je celo frazo prevedel z *enoglasna consent*, torej je pravilno razdvoumil del, v katerem je MT@EC odpovedal. Najzanimivejšo rešitev pa je ponudil Google Translate, ki je frazo (povsem napačno) razdvoumil v *monophonic mounting*.

4.1.4 Skupni rezultati

Po skupnem rezultatu se je najvišje uvrstil Google Translate (tabela 3) z 78-odstotno natančnostjo razdvoumljanja, sledi mu Presis (tabela 1) z 72,9-odstotno natančnostjo, temu pa MT@EC (tabela 2) s 67,6-odstotno natančnostjo. Že zgoraj je bilo na kratko omenjeno, da Presis izkazuje največjo konsistentnost skozi vse kategorije, kar je najbrž posledica dejstva, da temelji na pravilih in je kot tak manj občutljiv na to, s katerih tematskih področij so razdvoumljana besedila – za razliko od drugih dveh, ki temeljita na statističnem modelu in posledično prej odpovesta na besedilih s tematskih področij, na katerih ju niso trenirali.

Po drugi strani lahko glede na to, da je MT@EC precej bolj nekonsistenten kot Google, sklepamo, da je mogoče to pomankljivost omiliti s preprosto strategijo gole sile, z drugimi besedami s tem, da statistični model hranimo s karseda velikimi količinami vzporednih in enojezičnih korpusov, saj Google tu prednjači pred MT@EC.



Slika 1: Konsistentnost Presisa v primerjavi z drugima dvema prevajalnikoma.

Presis	KNJIŽNA BESEDILA	PUBLICISTIČNA BESEDILA	SPLETNA BESEDILA	SKUPAJ
Pravilno	67,9 %	69,9 %	73,3 %	70,4 %
Napačno	23,6 %	24,3 %	19,8 %	22,8 %
Pravilno s prevodno napako	2,8 %	2,3 %	2,6 %	2,5 %
Neprevedeno	5,7 %	3,5 %	4,3 %	4,3 %

Tabela 1: Rezultati za Presis.

MT@EC	KNJIŽNA BESEDILA	PUBLICISTIČNA BESEDILA	SPLETNA BESEDILA	SKUPAJ
Pravilno	43,4 %	79,2 %	61,2 %	64,3 %
Napačno	28,3 %	12,2 %	21,6 %	19,2 %
Pravilno s prevodno napako	5,7 %	1,7 %	3,4 %	3,3 %
Neprevedeno	22,6 %	6,9 %	13,8 %	13,2 %

Tabela 2: Rezultati za MT@EC.

Google Translate	KNJIŽNA BESEDILA	PUBLICISTIČNA BESEDILA	SPLETNA BESEDILA	SKUPAJ
Pravilno	66,0 %	80,3 %	75,9 %	75,2 %
Napačno	22,7 %	11,6 %	14,6 %	15,4 %
Pravilno s prevodno napako	2,8 %	3,5 %	1,7 %	2,8 %
Neprevedeno	8,5 %	4,6 %	7,8 %	6,6 %

Tabela 3: Rezultati za Google Translate.

5 Zaključek

Cilj eksperimenta je bil v praksi preveriti stanje tehnologije razdvoumljanja pri treh različnih strojnih prevajalnikih. Ob pregledu tega področja se izkaže predvsem, da je kljub obstoju številnih različnih strategij za razdvoumljanje besednega pomena njihova praktična implementacija redka. Le malokateri trenutno aktualen strojni prevajalnik pri izdelavi prevodov uporablja kakšno od obstoječih specializiranih tehnologij razdvoumljanja besednega pomena. To se zdi posledica dejstva, da so trenutno glavni trendi razvoja usmerjeni v statistične pristope s fraznimi modeli strojnega prevajanja, ki kot osnovne pomenske enote ne analizirajo posameznih besed, temveč daljše kose besedil. V tem primeru se razdvoumljanje v bistvu zgodi kot naravna posledica iskanja prevodno ustreznih daljših kontekstov v ciljnem jeziku. Z drugimi besedami, dovolj dolgi segmenti besedil se v fraznem modelu strojnega prevajanja razdvoumijo kar sami od sebe.

Takšna taktika zagotavlja sorazmerno visoko natančnost razdvoumljanja pri večini besedil, s katerimi se srečujemo pri strojnem prevajanju, vendar pa hitro odpove pri razdvoumljanju manj pogostih pomenov. V tovrstnih primerih se uporabnost različnih algoritmov za razdvoumljanje besednega pomena izkaže za nesporno. To se je potrdilo tudi v eksperimentu, ki je zajel Google Translate in MT@EC kot predstavnika statističnih pristopov ter Presis kot predstavnika na pravilih temelječih pristopov, ki ima tudi edini od trojice vgrajene eksplicitne postopke za razdvoumljanje (Holozan, 2011). Temu najbrž lahko pripišemo vsaj del zaslug za Presisovo visoko natančnost razdvoumljanja na književnih besedilih, v katerih se praviloma večkrat kot pri drugih besedilnih žanrih srečamo z redkejšimi besednimi pomeni. Druga očitna posledica pa je višja konsistentnost razdvoumljanja skozi različne besedilne žanre, ki jo izkazuje Presis v primerjavi s statističnima prevajalnikoma.

Za prihodnji razvoj in doseganje višjih natančnosti razdvoumljanja besednega pomena bi bilo morda smiselno preizkušati kombiniranje statističnih pristopov z opisanimi specializiranimi algoritmi za razdvoumljanje. Modularno dodajanje teh algoritmov že obstoječim statističnim strojnim prevajalnikom bi verjetno povzročilo povišanje natančnosti razdvoumljanja, vendar je to nekaj, kar bi zaradi specifičnosti delovanja statističnih prevajalnikov najbrž zahtevalo veliko dela pri implementaciji. Vseeno pa bi raziskovanje v to smer morda obrodilo koristne rezultate.

6 Zahvala

Zahvaljujem se prof. dr. Špeli Vintar za nasvete pri pripravi na izvedbo eksperimenta in pisanju članka, mag. Petru Holozanu za dostop do Presisa in Zoranu Zakiću za dostop do MT@EC.

7 Literatura

Eneko Agirre in Philip Edmonds. 2007. *Word Sense Disambiguation*. Springer, Dordrecht.

Amebis Besana – Pregibanje. http://besana.amebis.si/pregibanje/

Tim Chklovski, Rada Mihalcea, Ted Pedersen in Amruta Purandare. 2004. The Senseval-3 multilingual English-Hindi lexical sample task. V: *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, str. 5-8. Barcelona, Španija.

Dictionary, Encyclopedia and Thesauraus – The Free Dictionary. www.thefreedictionary.com

Darja Fišer in Benoît Sagot. 2015. Constructing a poor man's wordnet in a resource-rich world. *Language Resources and Evaluation 49*: 601–35.

Fran. www.fran.si

William Gale, Ken Church in David Yarowsky. 1992a. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. V: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, str. 249–56. Newark, ZDA.

William Gale, Ken Church in David Yarowsky. 1992b. One sense per discourse. V: *Proceedings of the DARPA Speech and Natural Language Workshop*, str. 233–37. New York, ZDA.

Graeme Hirst in David St-Onge. 1998. Lexical chains as representations of context in the detection and correction of malaproprisms. V: *WordNet: An electronic lexical database*, str. 305–32. MIT Press, Massachusetts, ZDA.

Peter Holozan. 2011. *Samodejno izdelovanje besedilnih logičnih nalog v slovenščini*. Magistrsko delo, Univerza v Ljubljani.

Jian Jiang in David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. V: *Proceedings of the International Conference on Research in Computational Linguistics*. Taipei, Tajvan.

Leposlovje. http://lit.ijs.si/leposl.html

Philip Resnik. 1995. Using information content to evaluate semantic similarity. V: *Proceedings of the International Joint Conference on Artificial Intelligence,* str. 448–53. Montreal, Kanada.

sloWTool. http://nl.ijs.si/slowtool/

George Kingsley Zipf. 1949. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley.

Korpus študentov prevajanja MetaTrans

Anja Tavčar, Ines Čeligoj Pregelj, Miha Pompe

Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana
anja.tavcar09@gmail.com, ines_celigoj@hotmail.com, miha.pompe@gmail.com

Povzetek

V pričujočem prispevku sta predstavljeni gradnja in analiza korpusa prevodnih napak študentov Oddelka za prevajalstvo Filozofske fakultete Univerze v Ljubljani. V korpusu so označene napake v študentskih prevodih poljudnoznanstvenih naravoslovnih in družboslovnih besedil iz slovenskega v angleški jezik ter njihovi popravki, ki jih je prispeval materni govorec angleškega jezika in profesor prevajanja. Napakam je pripisana tudi vrsta napake, za kar je bila uporabljena mednarodna tipologija napak v prevodih Mellange. Označevanje je potekalo s spletnim anotacijskim orodjem WebAnno. Za ugotavljanje relevantnosti rezultatov je bila narejena analiza dvojnega označevanja. Najpogostejši vir napak v prevodih je neustrezen slog, analiza dvojno označenega vzorca besedil pa kaže, da je ujemanje obeh označevalcev pri označevanju napak majhno.

The building and analysis of the corpus of translation errors made by students of the Department of Translation at the Faculty of Arts in Ljubljana are presented in this paper. The corpus is comprised of labelled errors in popular natural and social science texts translated by students from Slovene to English and the corrections made by a natural speaker of English language and professor of translation. Errors are labelled by type using the Mellange international translation mistake typology. The annotation was done with the online annotation tool WebAnno. The most common mistake in translations is inappropriate style, while the analysis of dual-labelled sample texts shows that there is little accord between different annotationists.

1 Uvod

Korpusi danes predstavljajo nepogrešljiv vir jezikovnih podatkov za jezikoslovne opise, raziskave in utemeljitve. Jezikovni podatki, ki jih dajejo korpusi, omogočajo v jeziku ločevanje med tipičnim in posebnim oz. individualnim, torej prepoznavanje osrednjih in obrobnih jezikovnih pojavov (Gorjanc et al., 2005). Za slovenščino je na voljo že veliko različnih vrst korpusov, vendar pa tako kot večina drugih jezikov še vedno nimamo na voljo reprezentativnega korpusa, ki bi služil kot pomoč pri študiju prevajalstva in bil sestavljen iz izvirnikov in njihovih prevodov ter obogaten s popravki in oznakami vrst napak, ki so jih naredili študenti med prevajanjem. Takšni korpusi nedvomno pripomorejo k poučevanju prevajalcev, saj na jasen način prikazujejo tipične jezikovne težave, ki jih imajo študenti pri prevajanju v tuji jezik, kar je še posebej pomembno v okoljih, v katerih je prevajanje v tuji jezik pogosta prevodna dejavnost. Tako okolje je tudi slovensko.

To omogoča pripravo kakovostnih didaktičnih smernic in avtentičnega gradiva za izobraževanje prevajalcev. Poleg tega pa so korpusi prevajalskih napak v veliko pomoč tudi študentom prevajalstva in tujih jezikov, saj jim nudijo nepogrešljive informacije o (ne)ustreznosti prevajalskih in jezikovnih rešitev, ki so jim v pomoč pri prevajanju.

V okviru projekta »Korpus študentov prevajanja MetaTrans«,[1] katerega cilj je bil zapolniti praznino na tem področju za slovenski jezik, je bil izdelan korpus, ki vsebuje angleške prevode slovenskih poljudnoznanstvenih besedil, ki so jih prevajali študenti 2. stopnje Oddelka za prevajalstvo na Univerzi v Ljubljani, in popravke profesorja, ki je materni govorec angleškega jezika. V skladu z mednarodno tipologijo napak v prevodih Mellange je popravkom pripisana tudi vrsta napake, ki so jo naredili študentje med prevajanjem v tuji jezik.

Čeprav sta v okviru diplomske naloge in doktorske disertacije na Oddelku za prevajalstvo na Univerzi v Ljubljani že potekali podobni raziskavi (Lavrič, 2009; Dobnik, 2011), je dodana vrednost našega korpusa ta, da smo v projekt vključili analizo dvojnega označevanja napak, ki omogoča vpogled v zanesljivost označevanja korpusa. Tako prispevek vsebuje tudi analizo in diskusijo napak v korpusu ter analizo in diskusijo medsebojnega ujemanja med označevalcema.

2 Sorodne raziskave

Eden najodmevnejših tovrstnih projektov je *MeLLANGE oz. Multilingual eLearning in LANGuage Engineering* (Castagnoli et al., 2011), katerega cilj je bil ustvariti večjezični označeni in poravnan korpus prevodov, ki vključuje prevode študentov prevajalstva in strokovnjakov s tega področja. Korpus, poimenovan Learner Translator Corpus (LTC), vsebuje prevode v desetih jezikih in je poleg tega tudi označen z oznakami o besednih vrstah in s podatki o lemi, poleg tega pa vključuje popravke in oznake napak, ki dajejo podatke o vrsti napake. Posebnost korpusa je tudi ta, da je bila za potrebe označevanja napak izdelana Mellangeva tipologija napak.[2] Korpus vsebuje metapodatke o prevajalcih besedil, saj so te informacije lektorjem, ki so prispevali popravke za sestavo korpusa, pomagale pri povezovanju napak z okoliščinami, v katerih je bilo besedilo prevedeno. Pri izdelavi korpusa je sodelovalo 440 študentov in profesionalnih prevajalcev, ki so prevajali 4 vrste besedil – pravna, tehnična, poslovna in novinarska. Ob zaključku projekta je bilo vseh prevedenih besedil 429, označenih pa 360. Projekt je bil zaključen leta 2007, končna analiza pa je pokazala, da so najpogostejše

[1] Raziskava je nastala v okviru magistrskega modula Korpusi in baze podatkov v štud. l. 2015/2016 pod mentorskim vodstvom doc. dr. Darje Fišer. Članek je bil oblikovan v okviru predmeta Slovensko strokovno besedilo, prav tako v štud. l. 2015/2016, pri prof. dr. Vojku Gorjancu.

[2] http://corpus.leeds.ac.uk/mellange/images/mellange_error_typology_en.jpg

napake pri prevajanju besedil uporaba napačnega termina, popačenje vsebine izvirnika in neskladje terminologije znotraj ciljnega besedila.

Mellangevo tipologijo napak pa si je za osnovo pri določanju napak v svojem korpusu izbral tudi študent oddelka za prevajalstvo, Davorin Lavrič, ki je leta 2009 v svoji diplomski nalogi izdelal korpus študentskih prevodov, ki ga sestavlja 122 slovenskih besedil, od tega 89 prevodov v angleški jezik. Besedila v korpusu so različno dolga in pokrivajo različna področja, popravke pa so prispevali profesorji z Oddelka za prevajalstvo Univerze v Ljubljani (Lavrič, 2009).

V Sloveniji je leta 2011 na Oddelku za prevajalstvo na Univerzi v Ljubljani potekal še en podoben projekt. V okviru doktorske disertacije je potekala gradnja korpusa študentskih napak (Dobnik, 2011), ki vsebuje prevode šestih besedil, ki so jih prevajali študenti 2. in 3. letnika dodiplomske stopnje pri predmetu Prevajanje iz francoščine v slovenščino. Vse popravke so prispevali profesorji, žal pa korpus ni označen z istimi oznakami, kar onemogoča medsebojno primerljivost rezultatov.

Podoben projekt je potekal na desetih ruskih univerzah, kjer je bil izdelan korpus *Russian Learner Translator Corpus* oziroma *RusLTC* (Kutuzov in Kunilovskaya, 2014), ki vsebuje prevode in napake v prevodih študentov prevajalstva. Študenti so prevajali v jezikovni kombinaciji ruščina-angleščina in angleščina-ruščina. Vsi prevodi so bili narejeni kot del izpita, prevajalskih tekmovanj ali kot del domače naloge. Marca 2014 je korpus RusTLC vseboval skupno skoraj 1,2 milijona besed, 258 izvirnih besedil in 1.795 prevodov, od tega pa je bilo z napakami označenih 198 prevodov v ruski jezik in 43 prevodov v angleški jezik, označevanje pa še ni bilo zaključeno (Kutuzov in Kunilovskaya, 2014).

Korpus prevajalskih napak je leta 2005 začel nastajati še v Španiji na Univerzi v Zaragozi. Korpus ENTRAD (Serrando in Sanz, 2008) vsebuje 45 angleških besedil, ki so jih v španščino prevedli študenti pri predmetu Uvod v prevajanje angleških besedil. Materni jezik študentov je večinoma španščina in francoščina, popravljavci oziroma označevalci napak pa so bili univerzitetni profesorji. Nekatere poizvedbe v korpusu so opremljene tudi z metapodatki, kot so prevajalčeva starost, spol in materni jezik. Oznake napak niso strojno berljive in temeljijo na barvni kodi in grafičnih oznakah.

Korpus prevajalskih napak je nastal tudi na univerzi Pompeu Fabra v Barceloni. *Korpus LTC-UPF* (Espunya, 2014) vsebuje besedila, ki so jih študenti prevajali iz angleščine v katalonščino. Korpus vključuje 10 izvirnih besedil in 194 prevodov. Označevanje korpusa je potekalo ročno in samodejno. Korpus se lahko uporablja za identifikacijo pogostih napak pri prevodih učencev in za analizo njihovih vzorcev jezikovne rabe. Korpus omogoča enostaven dostop do vzorcev napak in do več različic istega izvirnega besedila, kar omogoča kvalitetnejše izobraževanje prevajalcev.

Izgradnja korpusa z označenimi prevajalskimi napakami je med letoma 2009 in 2013 potekala v nemškem izobraževalnem prostoru, in sicer na univerzi Universität des Saarlandes. Zbiranje besedil za korpus *KOPTE* (Wurm, 2013) se je začelo leta 2009 pri predmetu prevajanje iz francoščine v nemščino. Večina besedil je vzetih iz časopisnih člankov, ki so jih študenti prevajali pri pouku. Za večino besedil obstaja več prevodov, saj so študenti med poukom prevajali tudi besedila z enakim izvirnikom. Korpus vsebuje več kot 77 izvirnikov in več kot 971 prevodov v nemščino. Označevanje prevajalskih napak je potekalo ročno z orodjem *UAM Corpus Tool*. V korpusu so z orodjem *TreeTagger* pripisane tudi oblikoskladenjske oznake. Za ročno označevanje napak je pobudnik izdelave korpusa KOPTE, Andrea Wurm, izdelal lastno tipologijo napak.[3] Tudi korpus KOPTE vsebuje metapodatke, kot so podatki o jeziku, ki so se ga študenti učili, podatki o času in načinu študija, podatki o poreklu staršev idr.

Na Poljskem je leta 2001 potekal projekt *PELCRA* oziroma *Polish and English Language Corpora for Research and Applications* (Uzar, 2002), v okviru katerega je bila na univerzi Uniwersytet Łódzki narejena pilotna študija, v okviru katere je bil narejen manjši korpus (15.000 besed). Korpus vsebuje prevode 180 besedil iz poljskega jezika (L1) v angleški jezik (L2). Korpus vsebuje besedila, ki so jih prevedli študenti na Oddelku za anglistiko. Od 180 besedil je bilo glede na kvaliteto besedila, berljivost besedila in kvaliteto prevoda z napakami označenih 50 besedil. Vse povedi v korpusu pa so bile označene s +, - ali 0, kar označuje, ali je prevedena poved ustrezno, neustrezno ali dokaj ustrezno prevedena glede na izvirnik.

Nekoliko drugačno strategijo od zgoraj naštetih so na portugalskem inštitutu Instituto de Engenharia de Sistemas e Computadores uporabili pri gradnji angleško-portugalskega korpusa (Costa et al., 2014), ki vsebuje oznake napak prevodov 150 povedi, ki so bile vzete iz različnih področij in prevedene v portugalščino s strojnim prevajalnikom Google Translate ali Moses, napake v njem pa je označil materni govorec portugalščine.

Podoben projekt je potekal na oddelkih za prevajalstvo na dveh univerzah v Franciji (Universite Paris Sud, Universite Paris Diderot), kjer je 46 študentov popravljalo strojne prevode, ki so jih nato označili še glede na vrsto napak (Wisniewski et al., 2014). Število vseh povedi v korpusu je 4.854, tip napake pa je pripisan skoraj polovici. Vsi dokumenti so prevedeni iz angleščine v francoščino. Vse naloge so nato pregledali še profesorji.

Omeniti velja, da v slovenskem prostoru že imamo dva nepogrešljiva korpusa z označenimi popravki in tipi napak, in sicer korpusa *Šolar* in *Lektor*, vendar sta oba enojezična, torej namenjena učenju slovenskega jezika. *Šolar* (Rozman et al., 2010) je korpus besedil, ki so jih učenci iz slovenskih osnovnih in srednjih šol samostojno tvorili pri pouku. Korpus, ki vsebuje 967.477 besed, je prvi tovrstni korpus besedil v Sloveniji, ki je narejen po vzoru korpusov usvajanja jezikov. Vsi jezikovni popravki v korpusu so delo učiteljev. Korpus *Lektor* (Popič, 2013) pa je zbirka lektoriranih slovenskih avtorskih besedil, ki vsebuje skoraj milijon besed, celoten korpus pa vsebuje 30.258 lektorskih popravkov, ki so bili ročno označeni.

3 Označevanje korpusa

3.1 Izbor besedil

V začetni fazi nastajanja korpusa smo s slovenskega spletnega portala Meta znanost izbrali slovenska besedila in njihove angleške prevode, ki jih je v angleščino v okviru projekta Slovenska znanost gre v svet prevedlo 9 študentov magistrskega programa prevajanja na Oddelku za

[3] http://fr46.uni-saarland.de/index.php?id=3702

prevajalstvo Univerze v Ljubljani. Izvirna besedila so delo slovenskih raziskovalcev in znanstvenikov. Ker smo želeli, da bi bil korpus čim bolj reprezentativen, smo izbrali 30 poljudnoznanstvenih člankov, 15 s področja naravoslovnih znanosti in 15 s področja družboslovja. Angleški del korpusa vsebuje 2.544 povedi.

3.2 Tipologija napak

V naslednji fazi smo v prevodih študentov označili popravke, ki jih je prispeval materni govorec angleščine izr. prof. dr. David Limon z Oddelka za prevajalstvo Univerze v Ljubljani. Popravljanje je potekalo na dveh nivojih. Najprej smo v vseh besedilih označili popravke, nato pa smo popravkom pripisali tudi vrsto napake s pomočjo tipologije napak, ki je bila razvita v projektu *MeLLANGE*.[4] Za uporabo tipologije Mellange smo se odločili, ker menimo, da je za primerljivost rezultatov pomembno, da se različne korpuse z označenimi tipi prevajalskih napak da primerjati, Mellangeva tipologija pa je že bila uspešno preizkušena na drugih sorodnih projektih.

Tipologija Mellange je zasnovana kot hierarhična shema, ki v osnovi temelji na razlikovanju med vsebinskimi napakami in jezikovnimi napakami. Tipologija Mellange vsebuje 39 oznak napak, ki so razvrščene v kategorije in podkategorije. Kategorije tipologije so v celoti predstavljene v nadaljevanju članka v Tabeli 1, v razdelku 4.1.

Vse podkategorije vsebujejo različne vrste napak, kot na primer preveč dobesedno, neustrezen termin, neustrezna raba ločil, neustrezen glagolski čas, neustrezen predlog, napačna raba velike/male začetnice, prevedeno preveč svobodno, preveden neprevedljiv izraz idr. Vsaka vrsta napake je označena s kodo, ki bo v korpusu stala poleg popravka. V okviru projekta smo celotno tipologijo lokalizirali v slovenščino, da bo dostopna tudi za slovenske raziskovalce. Tipologija vključuje uporabniško definirane kategorije, ki jih uporabnik izbere v primeru, da tipologija ne vsebuje vrste napake, ki jo ta želi označiti.

3.3 Označevanje napak

Ker je bilo vseh besedil 30, smo si delo razdelili tako, da je vsak od treh označevalcev označil 10 besedil v skladu s popravki maternega govorca. Pri delu smo za označevanje napak uporabljali orodje za jezikoslovno označevanje korpusov *WebAnno* (Yimam et al., 2013), ki je eno izmed najuniverzalnejših spletnih orodij za označevanje korpusov. Orodje omogoča projektno in oddaljeno delo in ne zahteva posebnih programerskih znanj, zaradi česar je zelo uporabno za jezikoslovce, humaniste, družboslovce in študente. Ena izmed glavnih značilnosti orodja je njegova prilagodljivost, saj omogoča označevanje napak na več nivojih, poljuben nabor oznak in različne načine označevanja. Prav tako pa omogoča, da na istem projektu dela več označevalcev, ki lahko vzporedno označujejo ista besedila, na podlagi česar se lahko ugotavlja tudi skladnost med ocenjevalci (ang. *inter-annotation agreement*).

Skladnost med ocenjevalci smo določali tudi pri našem projektu, saj je to zelo pomemben podatek za ugotavljanje, kako težavna je naloga in ali so bile odločitve označevalcev pri označevanju zanesljive. Označevanje prevajalskih napak je izredno problematičen in zamuden postopek, saj se mnenja o tem, za kakšno vrsto napake gre, med označevalci pogosto razlikujejo.

Vseh besedil, ki smo jih dvojno označili, je 8 (4 naravoslovna in 4 družboslovna). Analiza ugotovitev je predstavljena v 4. razdelku, 5. razdelek pa vsebuje analizo označevanja vseh 30 besedil, ki jih korpus vsebuje, in podatke o najpogostejših/najredkejših napakah študentov, količini napak glede na področje, prevajalca in na posamezne prispevke, kjer so označevalci opazili posebnosti.

4 Analiza besedil

Korpus označenih napak zajema 30 različno dolgih besedil 9 različnih prevajalcev, in sicer 15 besedil s področja družboslovja ter 15 s področja naravoslovja. Angleški del korpusa vsebuje 2.544 povedi. Vseh označenih napak je 2.458. Od teh se jih 1.546 oziroma 63 % pojavlja v besedilih s področja družboslovja, 939 oziroma 37 % pa v besedilih s področja naravoslovja.

Besedila s področja družboslovja imajo največ jezikovnih napak, ki niso podrobneje razdelane in spadajo v kategorijo drugo (234 oz. 15 %), sledijo slogovne napake (214 oz. 14 %) in napake v skladnji (192 oz. 12 %). Pri besedilih s področja naravoslovja pa je največ slogovnih napak (175 oz. 19 %), sledijo napake v kategoriji Jezik – drugo (145 oz. 15 %) ter napake v skladnji (68 oz. 7 %).

4.1 Analiza glede na tipologijo napak

V tabeli 1 so navedene kategorije napak in njihovo število od največjega do najmanjšega.

Napaka	Število napak
Jezik - neustrezen slog	389
Jezik - drugo	379
Jezik - skladnja	260
Jezik - terminologija - nepravilen pomen	163
Jezik - ločila	145
Prenos pomena - pomensko neustrezen	142
Jezik - napačna kolokacija	136
Jezik - neustrezen glagolski čas	125
Prenos pomena - preveč dobesedno	121
Jezik - napačen predlog	107
Jezik - napačna začetnica	96
Jezik - napačno število	72
Prenos pomena - dodano	39
Jezik - terminologija - drugo	36
Prenos pomena - izpust	31
Jezik - neprimerno za tip besedila	31
Jezik - istorečje	30
Jezik - črkovanje	26

[4] http://corpus.leeds.ac.uk/mellange/images/mellange_error_typology_en.jpg

Prenos pomena - preveč svobodno	25
Jezik - nesorodna beseda	23
Jezik - termin, preveden z neterminološkim izrazom	23
Prenos pomena - nejasno	20
Jezik - slog - drugo	17
Prenos pomena - drugo	11
Jezik - skloni in ujemanje - drugo	8
Jezik - terminologija - nekonsistentno znotraj ciljnega prevoda	7
Jezik - neskladno z glosarjem	7
Prenos pomena - prevedeni neprevedljivi izrazi (lastna imena ipd.)	4
Prenos pomena - poseganje v ciljni jezik - drugo	3
Jezik - nedosledno z izvirnim jezikom	3
Jezik - register - nekonsistentno znotraj ciljnega prevoda	3
Jezik - naglas ali diakritično znamenje	1
Jezik - register - drugo	1

Tabela 1: Kategorije napak tipologije Mellange

Kot je razvidno iz tabele 1, je največ napak zaradi neustreznega sloga (389 oz. 16 %), že pri drugi kategoriji (379 oz. 15 %) pa pridemo do težave, saj Mellangeva tipologija napak ni zajemala napačne rabe določnega in nedoločnega člena, zaradi česar so vse takšne napake znotraj kategorije jezikovnih napak označene kot »drugo«, kamor so se uvrstile tudi druge jezikovne napake, zato bi bila dobrodošla nekoliko bolj razdelana tipologija. Sledijo napake v skladnji (260 oz. 10 %) ter neustrezno izbiranje terminologije (163 oz. 7 %). Študentom prevajanja nemalo napak povzročajo tudi ločila (145 oz. 6 %): v večini primerov gre tu za nepravilno postavljanje vejic. Napake, ki se tičejo prenosa pomena v ciljni jezik, so šele na šestem mestu (142 oz. 6 %).

4.2 Razlikovanje med označevalci

Pri pregledu označenih napak je opaziti, da so se označevalci različno odločali za označevanje napak, kar kaže na to, da uporabljena tipologija ni najbolj jasna. Označevalec 1 ima tako na primer 298 napak označenih kot slogovne napake, medtem ko imata v tej kategoriji druga dva označevalca skupaj takšnih napak označenih le 91. Podobno tendenco opazimo pri označevalcu 2, ki je kot Prenos pomena – dodano označil 24 od skupno 39 tako označenih napak ter 18 od skupno 31 napak v kategoriji napak Prenos pomena – izpust in 144 od skupno 260 tako označenih napak v kategoriji Jezik – skladnja.
Seveda je pri tem potrebno omeniti, da so se označevalci na začetku projektnega dela zaradi narave in namena raziskave dogovorili le katero tipologijo bodo uporabljali, pri tem pa se niso dogovarjali o natančnejših smernicah ali konkretnejših primerih vrst napak. S tem smo želeli določiti, kakšna je stopnja ujemanja brez predhodnega dogovarjanja, ki bi sicer gotovo povišal stopnjo ujemanja med vsemi označevalci.

5 Analiza ujemanja med označevalcema

Poleg splošne analize označenih napak smo izvedli tudi analizo dvojnega označevanja napak, s katero smo želeli ugotoviti, kako skladni so označevalci napak, saj ti podatki vplivajo na verodostojnost rezultatov končne splošne analize in uporabnost korpusa.
Dvojno označen podkorpus zajema 8 besedil (4 naravoslovna, 4 družboslovna) od skupno 30 besedil iz korpusa za splošno analizo, kar znaša 27 % celotnega korpusa. Podkorpus vsebuje 443 povedi oz. 9.415 pojavnic.

Analizo smo izvedli na nivoju tipologije napak, in sicer tako, da je kurator s pomočjo programa WebAnno pregledal dvojno označena besedila in v primerih, kjer se označevalca nista strinjala, izbral ustreznejšo od predlaganih možnosti oz. predlagal svojo. Med različno označevanje se je štelo tudi drugačno označevanje napake v besedilu, četudi sta ji označevalca pripisala isto vrsto napake.

Analiza je pokazala, da je skupno število analiziranih napak v 443 stavkih 511, od tega sta enako označeni 102 napaki (20 %), pri 409 primerih (80 %) v 216 (49 %) stavkih pa sta označevalca napako označila različno.

Analiza neujemanj med označevalci glede na prevajalca besedila ni pokazala nobenih posebnosti, zato v nadaljevanju predstavljamo samo rezultate analize glede na področja besedila, kjer je prihajalo do večjih razlik. Stopnja ujemanja med označevalcema pri naravoslovnih besedilih je bila 7%, pri družboslovnih pa 13%. Število različno označenih napak pri naravoslovnih besedilih je bilo 167 (33 %), pri družboslovnih pa 242 (47 %).

Besedilo		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Skupaj**
Neujemanje na nivoju fraze		7	8	16	9	7	14	23	20	104 (25 %)
Neujemanje na nivoju tipa napak	Na vrhnjem nivoju	13	2	31	20	3	8	20	7	104 (25 %)
	Na srednjem nivoju	14	10	4	30	12	7	17	20	114 (28 %)
	Na spodnjem nivoju	3	6	18	18	5	5	9	23	87 (21 %)
skupaj		37 (9 %)	26 (6 %)	69 (17 %)	77 (19 %)	27 (6 %)	34 (8 %)	69 (17 %)	70 (17 %)	409

Tabela 2: Analiza področij ujemanja med označevalcema

Razlika v številu vseh napak med besedili je bila 22 %, kar pomeni, da sta bila označevalca pri označevanju naravoslovnih besedil bolj skladna.

5.1 Analiza neujemanj med označevalcema

Podrobneje smo analizirali, do kakšnih razhajanj med označevalcema prihaja najpogosteje, izdelali tipologijo napak in rezultate ovrednotili.

Iz tabele 2 je razvidno, da gre pri 25 % napak za neujemanje na nivoju fraze, pri ostalih 75 % pa za neujemanje na nivoju tipa napak. Največ neujemanj se je pojavilo na srednjem nivoju (28 %), najmanj na spodnjem nivoju (21 %).

V nadaljevanju analize se natančneje posvetimo najpogostejšim tipom razhajanja med označevalcema.

5.2 Neujemanje na nivoju fraze

Kot smo že omenili, za neujemanje štejemo tudi napake, ki sta jim označevalca pripisala isto vrsto napake, vendar jih je program zaradi različne označitve v besedilu zaznal kot neujemanje. Kot ilustrativni primer tovrstnega neujemanja bi lahko predstavili npr. označevanje manjkajočih vejic, kjer je eden od označevalcev označil besedo pred in za manjkajočo vejico, drugi pa le besedo pred ali po mestu manjkajoče vejice. Takšnih napak je 104 (25 %), torej dobra četrtina vseh napak. Če to upoštevamo pri številu enako označenih napak, se stopnja ujemanja med označevalcema zviša za 20 %, kar je bistveno izboljšanje rezultata.

5.3 Neujemanje na nivoju tipa napak

Področje analize neujemanja smo razdelili na tri dele, kot je vidno v tabeli 2:

- **Neujemanje na vrhnjem nivoju tipologije**

Sem štejemo oznake napak, ki so se razhajale na nivoju kategorij prenosa vsebine in na nivoju jezika. Gre za največje razhajanje med označevalci in napake z največjo težo, ki najbolj negativno vplivajo na uporabno vrednost korpusa. Takšnih odstopanj pri označevanju napak se je v besedilih pojavilo 104 (25 %). Pri analizi te vrste neujemanj nismo odkrili nobenih tipičnih vzorcev razhajanj, ki bi se pogosteje pojavljali, zaradi česar tudi ni mogoče predlagati možnosti za izboljšave.

Kot primer tovrstnega neujemanja pri označevanju lahko predstavimo naslednji popravek; v enem od besedil je bila beseda *diploma* prevedena kot *Bachelor thesis* in popravljena na *undergraduate dissertation*. Omenjeni popravek je eden od označevalcev označil kot napako na nivoju jezika – *pomensko neustrezno,* drugi pa kot napako na nivoju prenosa vsebine – *nepravilno (neskladno z izvirnim jezikom)* iz podkategorije terminologija in leksika.

- **Neujemanje na srednjem nivoju tipologije**

Gre za napake, katerih oznaki se razlikujeta glede na nivo znotraj prvih dveh večjih skupin napak, npr. med kategorijo *register* in *slog*. Tovrstnih razhajanj je bilo 114 (28 %).

Opazili nismo nobenih vrst napak, ki bi se pogosteje pojavljale, vendar je očitno, da bi jih bilo mogoče zmanjšati z jasnimi smernicami za prepoznavanje tovrstnih tipov napak.

Primer tovrstnega popravka je prevod besede ok. Prevedena je bila kot agreed in popravljena na okay. Prvi označevalec je popravek označil z napako nepravilno iz podkategorije terminologija in leksika, drugi pa kot neprimerno za tip besedila iz podkategorije register. Obe podkategoriji sodita v kategorijo napak na nivoju jezika.

- **Neujemanje na spodnjem nivoju tipologije**

Pri tovrstnem neujemanju napak opažamo neujemanje znotraj najožjih kategorij tipologije, npr. znotraj kategorije *slog* ali *higiena*. Razhajanj na tem nivoju je bilo nekoliko manj, pojavilo se je 87 napak (22 %). Opazili smo, da se na tem nivoju najpogosteje pojavljajo tri kombinacije napačno označenih napak (vse znotraj kategorije *terminologija in leksika*), in sicer: *nesorodna beseda* in *napačna kolokacija* (17 napak, 19 %), *napačen termin* in *nesorodna beseda* (11 napak, 13 %) ter *napačen termin* in *napačna kolokacija* (10 napak, 11 %).

Primer tovrstnega neujemanja pri označevanju je na primer prevod besede *sem,* ki je bil iz oblike *I am* popravljen na *I'm*. Eden od popravljalcev je popravek označil kot *neprimerno za tip besedila*, drugi pa kot *drugo*. Oba popravka sodita v podkategorijo *register*.

Analiza dvojnega označevanja napak je pokazala, da je odstotek nekonsistence med označevalcema precej visok, kar negativno vpliva na uporabno vrednost korpusa, zato ne moremo trditi, da je korpus v tej različici že zanesljiv. Za izboljšanje njegove uporabne vrednosti bi potrebovali natančne smernice za označevanje napak, s katerimi bi se izognili neujemanjem zaradi različnega označevanja besedila (dogovoriti bi se bilo potrebno, da se označuje npr. le prvo besedo v popravku) in odstopanja pri izbiri tipa napake na najvišjem nivoju (potreben je dogovor, kdaj gre za napačen prevod vsebine npr. *preveč dobesedno* in kdaj za jezikovno napako npr. *nerodno* v kategoriji *slog*).

6 Zaključek

Predstavljena raziskava je pilotna; njen cilj je bil izdelati zasnovo za korpus ter preizkusiti orodje, metodologijo in tipologijo za označevanje napak. V prihodnjih raziskavah je korpus potrebno povečati ter razširiti na druge tipe besedil, na prevode besedil študentov drugih jezikovnih kombinacij in drugih nivojev študija. Predvsem pa je potrebno izboljšati metodologijo označevanja in izdelati jasne smernice, da bomo izboljšali ujemanje med označevalci, zaradi česar bo korpus kvalitetnejši in bolj uporaben. Poleg vseh teh izboljšav bi se lahko poslužili tudi rešitev, ki so bile uporabljene na primer v korpusu Šolar, kjer so se odločili, da nekatere napake označijo dvojno, interpretacije le-teh pa prepustili raziskovalcem. Kljub začetnim težavam in veliko večjim neujemanjem med označevalci napak, kot smo pričakovali, smo s pomočjo izkušenj in rezultatov pridobili veliko koristnih izsledkov, ki so nam pri nadaljnjem delu s tovrstnimi korpusi v pomoč. Vsekakor pa takšen korpus odpira veliko število možnosti raziskovanja prevodnih napak študentov in lahko služi kot orodje pri prilagajanju učnega procesa prevajanja.

7 Literatura

Ana Espunya. 2014. *The UPF learner translation corpus as a resource for translator training*. V: Language Resources and Evaluation, Volume 48, Issue 1, str. 33–43, New York. http://dl.acm.org/citation.cfm?id=2598668

Andrea Wurm. 2013. *Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTE).* V: Journal of Translation and Technical Communication Research (trans-kom), zvezek 6 (2). http://www.trans-kom.eu/bd06nr02/trans-kom_06_02_06_Wurm_Eigennamen.20131212.pdf

Andrei Popescu-Belis, Margaret King in Houcine Benantar. 2002. *Towards a corpus of corrected human translations.* V: Machine translation evaluation : human evaluators meet automated metrics (LREC 2002), Third International Conference on Language Resources and Evaluation, str. 17–21, Pariz. http://www.mt-archive.info/LREC-2002-Popescu-Belis.pdf

Andrey Kutuzov in Maria Kunilovskaya. 2014. Russian Learner Translator Corpus: Design, Research Potential and Applications. V: Text, Speech and Dialogue: 17th International Conference (TSD 2014), str. 315–324, Brno. http://rus-ltc.org/references/tsd_rusltc_lnai-libre.pdf

Angela Costa, Tiago Luís in Luisa Coheur. 2014. *Translation errors from English to Portuguese: an annotated corpus.* V: Proceedings of the ninth international conference on language resources and evaluation (LREC'14), str. 1231–1234, Reykjavik. http://www.lrec conf.org/proceedings/lrec2014/pdf/199_Paper.pdf

Celia Floren Serrando in Rosa Lores Sanz. 2008. *The Application of a Parallel Coprus (English-Spanish) to the Teaching of Translation (ENTRAD PROJECT).* V: Micaela Muñoz-Calvo, Carmen Buesa-Gómez, M. Ángeles Ruiz-Moneva, ur., New Trends in Translation and Cultural Identity, str. 433–445. Cambridge Scholar Publishing, Velika Britanija.

Claudio Fantinuoli in Federico Zanettini. 2015. *New directions in corpus-based translation studies (Translation and Multilingual Natural Language Processing*. Language Science Press. Berlin.

Damjan Popič. 2013. *Je etično popravljati prevode*? V: Etika v slovenskem jeziku, literaturi in kulturi: zbornik predavanj, str. 118–123, Ljubljana.

Davorin Lavrič. 2009. *Vzporedni korpus študentskih prevodov*. Diplomska naloga. Filozofska fakulteta, Univerza v Ljubljani.

Elizaveta Kuzmenko in Andrey Kutuzov. 2014. *Russian Error-Annotated Learner English Corpus: a Tool for Computer-Assisted Language Learning.* V: Proceedings of the third workshop on NLP for computer-assisted language learning (SLTC 2014), str. 87–97, Uppsala University, Švedska. http://www.ep.liu.se/ecp/107/ecp14107.pdf

Guillaume Wisniewski, Natalie Kubler in François Yvon. 2014. *A Corpus of Machine Translation Errors Extracted from Translation Students Exercises.* V: International Conference on Language Resources and Evaluation, str. 3585–3588, Reykjavik. https://transread.limsi.fr/lrec_Wisniewskietal.pdf

Nadja Dobnik. 2011. *Analiza napak v prevodih študentov v funkciji načrtovanja in razvijanja predmetov francoskega jezika v okviru študijskega programa prevajalstva*. Doktorsko delo, Filozofska fakulteta, Univerza v Ljubljani.

Rafal S. Uzar. 2002. *A Corpus Methodology for Analysing Translation.* V: Stella Esther Ortweiler, ur., Cadernos de Tradução, str. 237–265. Lisboa.

Sara Castagnoli, Dragos Ciobanu, Kerstin Kunz, Natalie Kübler in Alexandra Volanschi. 2011. *Designing a Learner Translator Corpus for Training Purposes.* V: Natalie Kübler, ur., Corpora, Language, Teaching and Resources: from Theory to Practice, str. 221–248. Peter Lang. Berlin. http://www.eila.univ-paris-diderot.fr/_media/user/alexandra_volanschi/publi/castagnoli_et_al.pdf

Seid Muhie Yimam and Iryna Gurevych, Richard Eckart de Castilho in Chris Biemann. 2013. WebAnno: *A Flexible,Web-based and Visually Supported System for Distributed Annotations.* V: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations), str. 1–6, Sofija. http://www.aclweb.org/anthology/P13-4001.pdf

Tadeja Rozman, Mojca Stritar in Iztok Kosem. 2010. *Korpus šolskih pisnih izdelkov Šolar. V: Nova didaktika poučevanja slovenskega jezika*: Sporazumevanje v slovenskem jeziku, str. 4–35. Ljubljana.

Vojko Gorjanc, Simon Krek in Polona Gantar. 2005. *Slovenska leksikalna podatkovna zbirka*. Jezik in slovstvo, L/2: 3–19.

.

Literature Reloaded: using databases to explore literary trends

Lucija Dačić

Department of English, Faculty of Arts, University of Bristol
11 Woodland Road, Bristol BS8 1TB
lucija.dacic@bristol.ac.uk

1. Introduction

Since 2000, the speculative fiction genre has evolved significantly; new subgenres have emerged and the already existing ones underwent significant changes (Killheffer, 2000; James & Mendlesohn, 2003, 2012; Gill, 2013). Though once regarded as pulp, speculative fiction recently rooted itself in the mainstream, dominating large percentages of the book market (Vanderhooft, 2010; Chadwick, 2013). However, little has been written on the topic of how the publishing industry and its increasing dependency on trends (Thompson, 2012) impacted the genre; articles in trade journals show that the publishers seem to be trailing behind when it comes to identifying recent speculative fiction trends (Chadwick, 2012; Fox, 2012). Periodicals such as the Library Journal and Publishers Weekly, as well as Thompson in his book Merchants of Culture (2012), frequently mention the fact that no one is certain where the book market trends originate (Hollands, 2011; Fox, 2012; Chadwick, 2013). However, the publishers' trendsetting attempts likely effect subtle genre changes, and ultimately alter the ways in which a genre is perceived (Gill, 2013).

2. Goal of the paper

In order to successfully identify these changes and explore the role that the publishing industry plays in changing perceptions of genre, we need to reach beyond the traditional qualitative methods of research in the humanities. This paper presents the ways in which big data was used to inform research on genre and publishing industry, focusing on three particular databases:

2.1 Nielsen BookScan

The data on book sales in UK is collected by Nielsen BookScan, a commercial service aimed at publishing houses that gathers weekly sales figures from various UK booksellers and then compiles them into (amongst others) yearly bestselling charts. Their statistics capture over 90% of British print book market and are essential for understanding of publishing and bookselling trends.

2.2 Goodreads

Dubbed 'the world's largest site for readers and book recommendations' with over 30 million members and over 900 million books, it is a huge aggregator of data where people can, among other things, catalogue books they have read or wish to read via virtual shelves. The data regarding these shelves offers unique insight into the 'folk classification' of literature - i.e. readers' perception of literature and its genres.

2.3 Cambridge English Corpus

Compiled by Cambridge University Press, this is 'a multi-billion word collection of written, spoken and learner texts' gathered from various digital and other sources (such as books, magazines, radio and everyday conversation). Its original purpose is to inform and improve the Press' in-house English Language Teaching programme, but tracking the popularity of certain words can also tell external researchers a lot about the general public's familiarity with popular genre.

By using these databases to complement the data gathered through qualitative methods, we can get a much clearer picture of recent changes in perception and popularity of speculative fiction and its subgenres: comparing Nielsen Bookscan data (indicating popularity) to Goodreads data (indicating genre classification) can help us map past genre trends and potentially predict new ones, while Cambridge English Corpus data can shed further light on the way general public perceives genre literature.

3. References

Chadwick, Kristie. 2013. New worlds to explore: military sf and space opera stage a revival, fantasy goes dark, digital publishing is here to stay. *Library Journal*, 138 (13), p.22.

Chadwick, Kristie. 2012. Hungry for SF: genre crossovers retain fans and attract new readers. *Library Journal*, 137 (13), p.18.

Fox, Rose. 2012. Crossing the streams: with publishing in flux, genres mix and mingle. *Publishers Weekly*, 259 (37), pp.24.

Frow, John. 2006. *Genre*. London: Routledge.

Gill, R. B. 2013. The Uses of Genre and the Classification of Speculative Fiction. *Mosaic: a journal for the interdisciplinary study of literature*, 46 (2), pp.71-85.

Hollands, Neil. 2011. Sf/fantasy's epic journey: high fantasy makes a comeback, sf searches for a renaissance. *Library Journal*, 136 (13), pp.20.

James, Edward and Mendlesohn, Farah eds. 2003. *The Cambridge companion to science fiction*. Cambridge: Cambridge University Press.

Killheffer, Robert K. J. 2000. Merging, but not yielding. *Publishers Weekly*, 247 (3), pp.32.

Squires, Clare. 2007. *Marketing Literature: The Making of Contemporary Writing in Britain*. Basingstoke: Palgrave Macmillan.

Thompson, John B. 2012. *Merchants of culture: the publishing business in the twenty-first century*. Cambridge: Polity.

Vanderhooft, JoSelle. 2010. Opening New Doors. *Publishers Weekly*, 257 (15), pp.22-27.

Digitalna arheologija? Primer uporabe digitalnih orodij za analizo arheološkega najdišča

Jernej Rihter
ZRC SAZU Inštitut za arheologijo
Novi trg 2, 1000 Ljubljana
jernej.rihter@zrc-sazu.si

1 Uvod

Temelj predstavitve bo prikaz analize arheološkega najdišča v modernem digitalnem okolju, ki omogoča souporabo različnih digitalnih orodij. Gre za prikaz dobre prakse, katere končni rezultat ni zgolj (analogna) knjižna objava, temveč tudi arhiv digitalnih podatkov, ki je primeren za pre-uporabo.

2 Namen članka

Na vprašanje, kaj je digitalna arheologija, dobimo skorajda toliko odgovorov, kolikor imamo sogovornikov. Vsi se strinjamo, da gre za uporabo digitalnih orodij. Le redko pa se strinjamo o podrobnostih, kaj je digitalno orodje. Ena skrajnost je »uporaba računalnika« (za kaj več kot paket pisarniških orodij), druga skrajnost je uporaba najmodernejših tehnologij, torej trenutno na primer virtualne realnosti. Morda je ena izmed možnih definicij naslednja: digitalna arheologija je tista, katere rezultat je digitalno (pre-) uporaben, tj. primeren za ponovno uporabo v digitalnem okolju. Tak primer je (prostorska) podatkovna zbirka. Po tej definiciji arheologija, pri kateri z digitalnimi orodji proizvajamo samo začasne podatke (brez metapodatkov, datoteke ki so razumljive in uporabne samo kreatorju), končni cilj pa je izključno objava članka ali knjige, ni digitalna arheologija. To seveda ni splošno sprejeta definicija, je pa vodilo analize grobišča Župna cerkev v Kranju.

3 Primer grobišča Župna cerkev Kranj

Grobišče Župna cerkev v Kranju uvrščamo med največja srednjeveška grobišča v Evropi. Posebnost tega arheološkega najdišča je tudi izjemen časovni razpon od 7. st. do 18. st. n. š. S tem najdišče nudi potencial za vzpostavitev dobre absolutne arheološke kronologije zgodnjega srednjega veka v vzhodnih Alpah, ki je že skoraj stoletje ključni deziderat zgodnjesrednjeveške arheologije.

Raziskave grobišča so se začele leta 1953 in nadaljevala z daljšimi presledki do leta 2013. Grobišče obsega preko 2500 grobov. Gorenjski muzej hrani približno 3000 predmetov iz teh grobov, antropološko je bilo analiziranih več kot 1200 okostij. Kljub temu je bilo do danes analiziranih manj kot tridesetih grobov, večinoma izbor lepih zgodnjesrednjeveških predmetov, ki ne dajejo pravega vpogleda v najdišče (Kastelic, 1960; Valič, 1967; 1974, 1975; 1978; 1985; 1991, Sagadin, 1985; 1991; Bitenc in Knific, 2000).

Zato ZRC SAZU, Inštitut za arheologijo skupaj s partnerji od leta 2011 izvaja intenzivne analize arhiva dokumentacije, arheoloških najdb, antropoloških ostankov in arheoloških kontekstov. Več kot polstoletna zgodovina raziskav analizo otežuje, saj imamo opraviti z rezultati štirih različnih arheoloških metod dela in s štirimi tipi dokumentacije (Štular et al., 2013 in tam navedena literatura).

4 Uporaba digitalnih orodij za analizo arheološkega najdišča

Eden temeljnih podatkov vsake moderne arheološke raziskave je stratigrafski odnos (Harris, 1989; Harris, Brown in Brown III, 1993; Balme in Paterson, 2006). Grob, ki je vkopan v nek drug grob, je mlajši izmed teh dveh. Grob, preko katerega je bil zgrajen zid, je starejši od slednjega. Pri modernih arheoloških izkopavanjih, v Sloveniji nekako od 1990-tih dalje (npr. Novaković in Turk 1991; Grosman, 1991), je stratigrafski odnos najpomembnejši element opazovanja in dokumentiranja. Pred tem pa izkopavalci na ta podatek niso bili (tako) pozorni in so ga zato dokumentirali redkeje. Poleg velike množice podatkov in raznorodne dokumentacije v našem primeru torej nastopi še dodatna težava: iščemo podatke, ki so v dokumentacijo prišli posredno ali celo pomotoma (prim. Pleterski, 2008). V nadaljevanju bomo predstavili metodo analize stratigrafskih odnosov na arheološkem najdišču Župna cerkev v Kranju. Isto metodo je možno aplicirati na številne arhive starejših arheoloških izkopavanj.

Metoda temelji na treh digitalnih orodjih:

1. podatkovna zbirka (informatiziranih) arhivskih zapisov,
2. prostorska podatkovna zbirka in
3. program za analizo stratigrafskih odnosov.

4.1 Arheološki dnevnik

Temeljni element arheoloških izkopavanj je arheološki dnevnik. V našem primeru gre za več zvezkov rokopisa in drugih arhivskih zapisov. Arhiv smo inventarizirali in opisali v podatkovno zbirko. Vsi terenski dnevniki so bili transkribirani (Štular in Belak, 2012; Štular in Belak, 2012a; Belak, 2013; Štular in Belak, 2013; Sagadin, 2014; Belak, 2014). Oboje skupaj omogoča hitro iskanje podatkov v terenskih dnevnikih in, po potrebi, hiter dostop do originalnih dokumentov. Hkrati je to osnova za arhiv digitalnih podatkov.

4.2 Prostorska podatkovna zbirka

Prostorska dokumentacija je drugi steber arheološke dokumentacije. V našem primeru gre za meritve in arheološke risbe (načrti v merilu 1:100, 1:20 in 1:10), ki pa so umeščeni samo v vsakokraten *ad hoc* vzpostavljen relativni koordinatni sistem. V okolju geografskih informacijskih sistemov (GIS) smo vse podatke prenesli v moderni absolutni koordinatni sistem. Šele to omogoča medsebojno primerjavo podatkov iz različnih izkopavanj, na primer identificirati kateri skelet dokumentiran l. 1953 je ležal nad skeletom, dokumentiranim leta 2011.

4.3 Analiza stratigrafskih podatkov

Opisani podatkovni zbirki sta vir stratigrafskih podatkov. Na našem najdišču gre za razmeroma veliko število odnosov, približno 10.000, zelo pogosti pa so tudi nasprotujoči si ali napačni podatki. Ker je veliko število odnosov in razmeroma nizka kakovost podatkov v arheologiji pogosta, so v 1990-tih in zgodnjih 2000-tih razvili več programskih orodij za analizo (Hundack et al. s. a.; Herzog s. a.; Butina, Klasinc, Zorc, 2007). Vendar se je razvoj ustavil (zdi se, da je glavni razlog izjemno ozek krog razvijalcev in hkrati premajhen trg za komercializacijo) in večino teh orodij dandanes lahko uporabljamo le v virtualnih okoljih starejših operacijskih sistemov.

Literatura

Jane Balme in Alistair Paterson. 2006. *Archaeology in practice: a student guide to archaeological analyses*. Malden, Oxford, Carlton.

Mateja Belak (ur.). 2013. *Grobišče Župna cerkev v Kranju. Dnevniki izkopavanj 1969-1973*. Monographiae Instituti Archaeologici Sloveniae 5. Inštitut za arheologijo ZRC SAZU, Založba ZRC.

Mateja Belak (ur.). 2014. *Grobišče Župna cerkev v Kranju. Grobni zapisniki*. Monographiae Instituti Archaeologici Sloveniae 7. Inštitut za arheologijo ZRC SAZU, Založba ZRC.

Polona Bitenc in Timotej Knific (ur.). 2000. *Od Rimljanov do Slovanov. Predmeti*. Razstava, Narodni muzej Slovenije v Ljubljani, 2. junij - 15. november 2000.

Eva Butina, Rok Klasinc in Miha Zorc. 2007. Predstavitev prostorskega dokumentiranja arheoloških izkopavanj in programskega paketa Miniexplorer. *Arheo*, str. 93–115.

Darja Grosman. 1991. Kocka, kocka, kockica…: od arheološkega zapisa v zemlji do arheološkega zapisa na papirju. *Arheo* 12, str. 25–36.

Edward C. Harris. 1989. *Načela arheološke stratigrafije*. Slovensko arheološko društvo, Ljubljana.

Edward C. Harris, Marley R. Brown III, Gregory J. and Brown. 1993. *Practices of archaeological stratigraphy*. London, San Diego, New York, Boston, Sydney, Tokyo, Toronto.

Irmela Herzog. s. a., Stratify, 1.5, Manual, *http://www.stratify.org/Download/Stratify_Manual.pdf*

Christoph Hundack, Petra Mutzel, Igor Pouchkarev, Barbara Reitgruber, Barbara Schuhmacher in Stefan Thome. s. a. *ArchEd*. https://www.ac.tuwien.ac.at/files/archive/ArchEd/ArchEd_UsersGuide.pdf

Jože Kastelic. 1960. Staroslovanski Kranj 900 let Kranja. *Spominski zbornik*, Kranj, str. 41–50.

Predrag Novaković in Peter Turk. 1991. Kamen na kamen palača... (izkopavanje gradišča na Krasu). *Arheo* 12, str. 57–68.

Andrej Pleterski. 2008. *Zgodnjesrednjeveška naselbina na Blejski Pristavi: najdbe*. Ljubljana.

Milan Sagadin. 1985. Kranj. - Župna cerkev sv. Kancijana in tovarišev. *Varstvo spomenikov* 27, str. 283–284.

Milan Sagadin. 1991. Najstarejša cerkvena stavba v Kranju. *Pod zvonom Sv. Kancijana*, str. 31–44.

Milan Sagadin. 2014. *Grobišče Župna cerkev v Kranju, Dnevnik izkopavanj 1984*. Zbirka E-Monographiae Instituti Archaeologici Sloveniae 6. Inštitut za arheologijo ZRC SAZU, Založba ZRC.

Benjamin Štular (ur.) in Mateja Belak (ur.). 2012. *Grobišče Župna cerkev v Kranju, Dokumentacija o izkopavanjih v letu 1953*. Zbirka E-Monographiae Instituti Archaeologici Sloveniae 1. Inštitut za arheologijo ZRC SAZU, Založba ZRC.

Benjamin Štular (ur.) in Mateja Belak (ur.). 2012a. *Grobišče Župna cerkev v Kranju. Kartoteka najdb iz leta 1953*. Monographiae Instituti Archaeologici Sloveniae 2. Inštitut za arheologijo ZRC SAZU, Založba ZRC.

Benjamin Štular (ur.) in Mateja Belak (ur.). 2013. *Grobišče Župna cerkev v Kranju. Dokumentacija o izkopavanjih v letih 1964, 1965 in 1966*. Monographiae Instituti Archaeologici Sloveniae 4. Inštitut za arheologijo ZRC SAZU, Založba ZRC.

Benjamin Štular, Ana Ornik Turk in Andrej Pleterski. 2013. *Dotik dediščine: trirazsežni prikaz zgodnjesrednjeveškega naglavnega nakita iz najdišča župna cerkev v Kranju*. Ljubljana: Inštitut za arheologijo ZRC SAZU, Založba ZRC.

Andrej Valič. 1974. Arheološka dokumentacija pri izkopavanjih staroslovanskega grobišča v Kranju. *Varstvo spomenikov* 17-19/1, str. 47–50.

Andrej Valič. 1975. Oris 20-letnih raziskovanj grobišča v Kranju. *Kranjski zbornik*, str. 159–167.

Andrej Valič. 1978. La necropole slave a Kranj. *Inventaria Archaeologica* 21. Ljubljana.

Andrej Valič. 1967. Staroslovanski Kranj. Das Altslawische Kranj. *Arheološki vestnik* 18, Ljubljana, str. 417–426.

Andrej Valič. 1985. Osnovna izhodišča arheoloških proučevanj mesta Kranja (Carnium). *Kranjski zbornik* (1985), str. 88–94.

Andrej Valič. 1991. Poznoantični relikti v staroslovanskem okolju Gorenjske in Kranja. *Pod zvonom Sv. Kancijana*, str. 24–30.